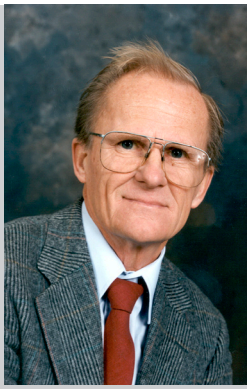
TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

Tampere International Center for Signal Processing
TICSP Series #38

Peter Grünwald, Petri Myllymäki, Ioan Tabus, Marcelo Weinberger & Bin Yu (eds.)

Festschrift in Honor of Jorma Rissanen on the Occasion of his 75th Birthday



Tampere 2008

Tampere International Center for Signal Processing.
TICSP series # 38

Peter Grünwald, Petri Myllymäki, Ioan Tabus, Marcelo Weinberger &
Bin Yu (eds.)

Festschrift in Honor of Jorma Rissanen on the Occasion of his 75th Birthday

Tampere International Center for Signal Processing
Tampere 2008

ISBN 978-952-15-1962-8
ISSN 1456-2774

Tampereen Yliopistopaino Oy, 2008

Contents

Preface

This *Festschrift* is presented in honor of Jorma Rissanen, on the occasion of his 75th birthday on October 20th, 2007. It contains 20 contributions by Jorma's colleagues and friends covering a wide range of topics in the areas of information theory, statistical modeling and inference, data compression and applications of modeling in science, engineering, and economics, reflecting the many areas where Jorma's work had a great impact.

A special session will be organized in the IEEE Information Theory Workshop to be held in Porto, Portugal during May 5-9, 2008. A number of the contributions will be presented in the special session of the workshop and the final printed copy of the edited version of the present volume will be available.

Jorma's exemplary career and his discoveries are a continuous source of inspiration for generations of researchers and students. His discoveries in science have had a substantial impact on the foundations of statistics and information theory. They prove fruitful in a wide range of applications, including science (psychology, molecular biology, astrophysics), engineering (control, computer science, signal processing), and economics.

Jorma is energetic as ever, constantly producing novel, galvanizing contributions in his positions as professor emeritus at Tampere University of Technology, visiting professor at Helsinki University of Technology and University of London, and HIIT Fellow at Helsinki Institute for Information Technology.

This volume is built out of contributions continuing in many ways the line of thoughts and principles promoted by Jorma, giving a glimpse of the diversity of areas relevant to his research. This volume will also reveal some insights of his remarkable personal life. All those who have had the privilege to meet and work with him have been amazed to see in him such an example of staying straight in life and being faithful to his pursuit of truth and values through science.

We wish to Jorma the fulfillment of all his wishes and strength to continue transforming his wishes into reality.

Happy Birthday!

Bin, Ioan, Marcelo, Peter, and Petri

Berkely, Tampere, Palo Alto, Amsterdam, Helsinki

October 20th, 2007

A Conversation with Jorma Rissanen

Pentti Huuhtanen, Erkki P. Liski and Simo Puntanen

Brief Biography

Jorma Johannes Rissanen was born in Pielisjärvi, Finland, on October 20, 1932. He received his Master's degree in electrical engineering in 1956, his Licentiate and Doctor of Technology degrees in control theory and mathematics in 1960 and 1965, respectively, all from the Helsinki University of Technology, Finland. He joined IBM's San Jose Research Laboratory, California, 1966, where he remained for more than three decades, except for the academic year 1973–74, when he held the chair of control theory in Linköping University, Sweden. He is a fellow of the Helsinki Institute for Information Technology, University of Helsinki. He is also a Professor Emeritus at the Tampere University of Technology. Since 1998 he has been appointed Visiting Professor in Computer Science at Royal Holloway, University of London. He is a Foreign Member of the Finnish Academy of Science and Letters and he has received an Honorary Doctorate degree from the Tampere University of Technology, Finland, in 1992.

Jorma Rissanen is the founder of the Minimum Description Length (MDL) principle, a new inductive principle in statistical modeling. He has carried out significant research in the fields of control, prediction and system theories, relation theory, numerical mathematics, information and coding theory, probability theory and statistics. He has published more than a hundred research papers and the books *Stochastic Complexity in Statistical Inquiry* (1989, World Scientific) and *Information and Complexity in Statistical Modeling* (2007, Springer), and he holds 15 US patents. He continues to be active in research.

Jorma Rissanen has earned many honors and awards, including the 2006 Kolmogorov Medal of the Computer Learning Research Centre at Royal Holloway, University of London; an IEEE Information Society Golden Jubilee Award for Technological Innovation for the invention of arithmetic coding in 1998; the IEEE 1993 Richard W. Hamming medal 'For fundamental contributions to information theory, statistical inference, control theory, and the theory of complexity'; an IBM Corporate Award in 1991 for the MDL/PMDL principles and stochastic complexity; an IBM Outstanding Innovation Award in 1988 for work in statistical inference, information theory, and the theory of complexity; the Best Paper Award from the IEEE Information Theory Group in 1986 (covered all papers published in information theory during the preceding two-year period); the Best Paper Award from the International Federation of Automatic Control in 1981; an IBM Outstanding Innovation Award in 1980 for the introduction of arithmetic codes. He is an IEEE Fellow.

In 1956 he married Riitta Åberg, and they have a son Juhani and a daughter Natasha, one grandson Juhani and one granddaughter Elissa.

The following conversation took place both by email in the end of July and meeting in Tampere in August 2007.

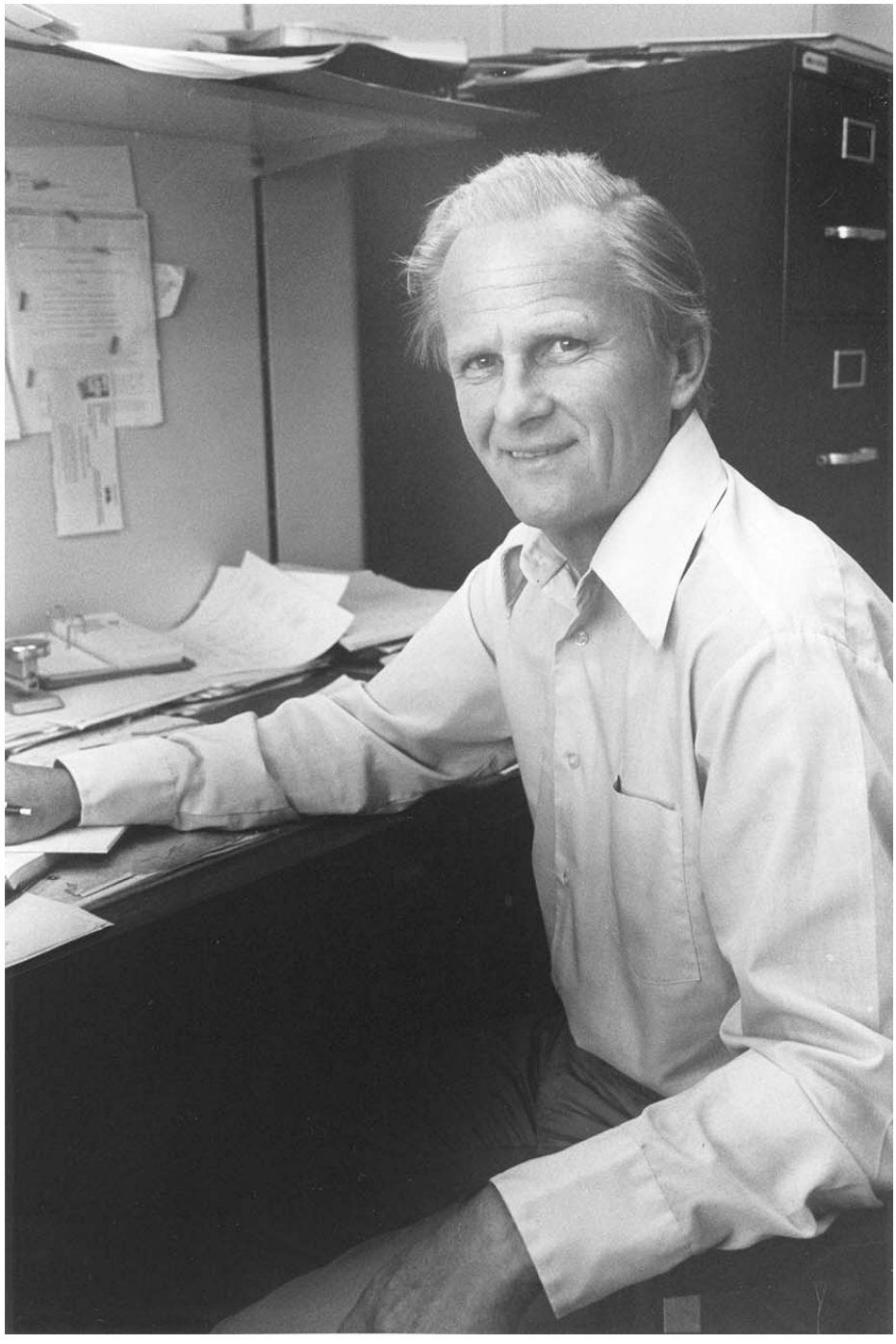


Figure 1: Jorma Rissanen in his San Jose office in 1985.

Early life, Kemi, wartime, 1932–55

Question: *Jorma, let us start at the very beginning. You were born on October 20, 1932 in Pielisjärvi, near then the Soviet border, but soon your family moved to Kemi, which is situated in the north-western part of Finland near to the Sweden–Finland border. Jorma, tell us something about your family and your childhood in Kemi.*

Jorma Rissanen: There isn't all that much to tell. We lived in rented apartments in various suburbs of the small town, and just about all I remember of the period before the war was that it was cold and there were huge piles of snow.

Did you have sisters or brothers?

I had four sisters and a brother.

Soon after you went to primary school in 1939 the Winter War broke out when the Soviet Union attacked Finland on November 30, 1939. The war lasted 100 days. What kind of time was that in your life?

It did not affect us too much. Since Kemi is hundreds of kilometers from the Soviet border we had no bombing, even though we had to go to bomb shelters a few times.

The period of peace after the Winter War was short. The Continuation War lasted from June 25, 1941 until September 19, 1944. Germany took part by providing critical material support and military cooperation to Finland and also some German troops operated in North Finland. When Finland in September 1944 made peace with the Soviet Union the so-called Lapland War broke out between Finland and Germany. Your home town Kemi was also in the middle of battles. Could you tell about your personal war experiences?

Just before the fighting reached the house where we lived we took off to the woods together with the family of the owner of the house, a horse and a cow. There we waited for a few days until fighting passed the area. When the sound of the cannon shots seemed to come from farther away I decided to take a look at our home on the main road to north. A miles long German column of vehicles of any sort had been stopped on the road with dead soldiers with their arms lying nearby. I, together with the neighborhood boys, had a keen interest in guns and ammunition, and I thought that this is an excellent opportunity to get Schmeisser machine pistols. I picked two and hid them under the house where we lived. Every room of the apartment had bullet holes so that none of us would have survived if we had stayed. On the yard behind an Y-shape tree a dead German soldier with his machine rifle was lying. I was tempted to take the rifle but it was too long and heavy and I left it. Unfortunately I had no chance to try the machine pistols, because the next few days a Finnish soldier came to check if there are any weapons around, and my father told him that I had a cache under the house. I remember those as exciting times!

How about your early schooling in wartime?

After the Germans were pushed to Norway there was almost nothing to eat in Kemi, and I was sent with many others to Sweden, where I stayed one year without attending any school and missed the third grade. I learned Swedish though, which I however kept forgetting faster than what I was taught an hour every day from the third grade on in the junior high school.

Figure 2: Conference in Ann Arbor, Michigan: Jorma Rissanen and Claude Shannon. Jorma Rissanen has received the Best Paper Award from IEEE Information Theory Group in 1986.

What can you recall about your life during the elementary, junior high and high school days? Did you start showing any added aptitude in one subject over others at some stage?

For a couple of years I had trouble with arithmetic and mathematics, but then all of a sudden, it seemed, I understood that arithmetic and mathematics is nothing but a game with certain rules, which you have to learn. After that it became my favorite field.

Did you happen to meet inspiring teachers in science and mathematics?

Our mathematics teacher happened to be excellent. He kept on reminding the students that there is nothing to learning mathematics and all of you are capable of getting at least the grade B. He left me with the long lasting impression that I should be able to learn anything if it is explained clearly.

Of science we were taught basic physics, the teacher of which was very good, too.

How about the matriculation examination in the early 1950s? What kind of thoughts does that bring to your mind?

The examination itself took one whole day. I remember the 3-hour mathematics test, which consisted of ten problems. One was an involved percentage type of problem, which was very hard

to decipher, and I missed it. However, since it was the only one I missed I still got the grade 'laudatur' and was allowed to attend the 2-week summer school for the Technical University of Helsinki. The other tests meant nothing to me and I must have passed them because I graduated.

Helsinki University of Technology, 1952–56

You enrolled in the undergraduate programme at Helsinki University of Technology (HUT) in 1952. How did you become interested in engineering studies?

At the time there were two favorite ways to proceed towards higher education with good career prospects, the medical and technical fields, and my choice was clear: the most prestigious technical university of Finland was in Helsinki.

After the years in Kemi, how did you like to live in Helsinki in your student years?

I actually lived in Helsinki itself only the first year, after which I moved to Otaniemi and took the bus to the old technical university building, where all the lectures were given. I had no social life other than soccer, which we played just about every night in the newly built indoors arena in Otaniemi.

Do you recall teachers or other people at HUT that you regard as influences on your career as it developed?

Hans Blomberg, the professor of theoretical electrical engineering was excellent. So was the mathematics professor Kalle Väisälä, from whom I took all the courses he gave.

You served in the army right after finishing the undergraduate degree in 1956. How did you cope with the mandatory military assignment?

My military career ended prematurely when I hurt my knee and was operated, which gave me a good chance to study for the Licentiate degree in the military hospital.

I recall that this knee problem has not been entirely solved and it has not allowed you to play soccer according to your talents? Is that right?

That is right. I didn't play soccer in the following ten years until a knee specialist in California told me that there is nothing much wrong with the knee and go ahead and start playing, which I did.

You started your Licentiate studies in 1957. What form were these studies? Was your aim already directed at a doctoral degree?

There were no courses on control theory, and I picked the topic for graduate studies because of Hans Blomberg's interest in it. My aim certainly was a doctoral degree.

You have told that one of your first places of employment was the Helsinki City Utility Company.

It was my first steady employment; after all, I had just married and had to make a living.

University of Tampere

Figure 3: Conference Dinner in Tampere Conference 1987: T. W. Anderson, Jorma Rissanen and Dorothy Anderson.

So you were working and continuing your studies at the same time?

Yes.

When did your interests in mathematics and science begin to emerge?

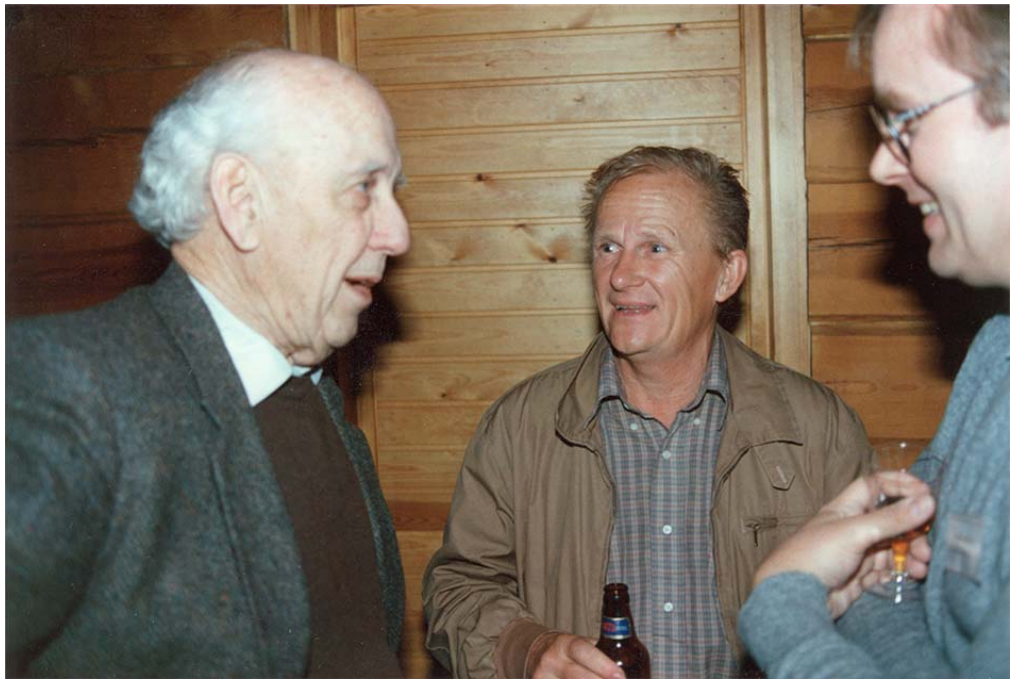
They grew gradually during the undergraduate studies.

What was your dissertation about?

It was about the horrible problem of adaptive control.

Could you please be a bit more specific? Your answers makes me curious...

Although I feel the less said about the thesis the better, let me just add that the most important and difficult part in problem selection is to pick a topic which both is tractable and reasonably significant. Since nobody had been able to do anything worthwhile about the adaptive control problem I should never have taken that as a PhD topic. Needless to say I wasted my time as well as that of the examiner, Olli Lokki, and produced absolutely nothing worthwhile.



Simo Puntanen

Figure 4: Sauna Party in Tampere Conference 1987: Ted Hannan, Jorma Rissanen, and Heikki Hella.

Who was your thesis advisor and what kind of process it was to work for a doctoral degree at that time at HUT?

My thesis advisor was Hans Blomberg, and since there were no formal lectures given on graduate level I simply started by reading literature given to me by him. I continued studying mathematics under the tutoring of Kalle Väisälä. The examination was done in my working out problems in books as home work – not avoiding the more difficult ones as instructed by Professor Väisälä.

Often, work done for doctoral thesis shapes one's future conception of the field. Is this the case for you?

No.

Are there other people at HUT that you regard as influences on your career as it developed?

No.

You received the Doctor of Technology Degree in control theory and mathematics from the Helsinki University of Technology in 1965. Were there many other doctoral students in your field of research at HUT?

Since I did most of my work away from the university while working for IBM, I didn't know of any other doctoral student in any field at HUT.

Did you have any contacts with the Department of Mathematics at the University of Helsinki? At that time there were many world famous mathematicians like e.g. Rolf Nevanlinna, Paul Kustaanheimo, P. J. Myrberg, Gustav Elfving and Olli Lehto.

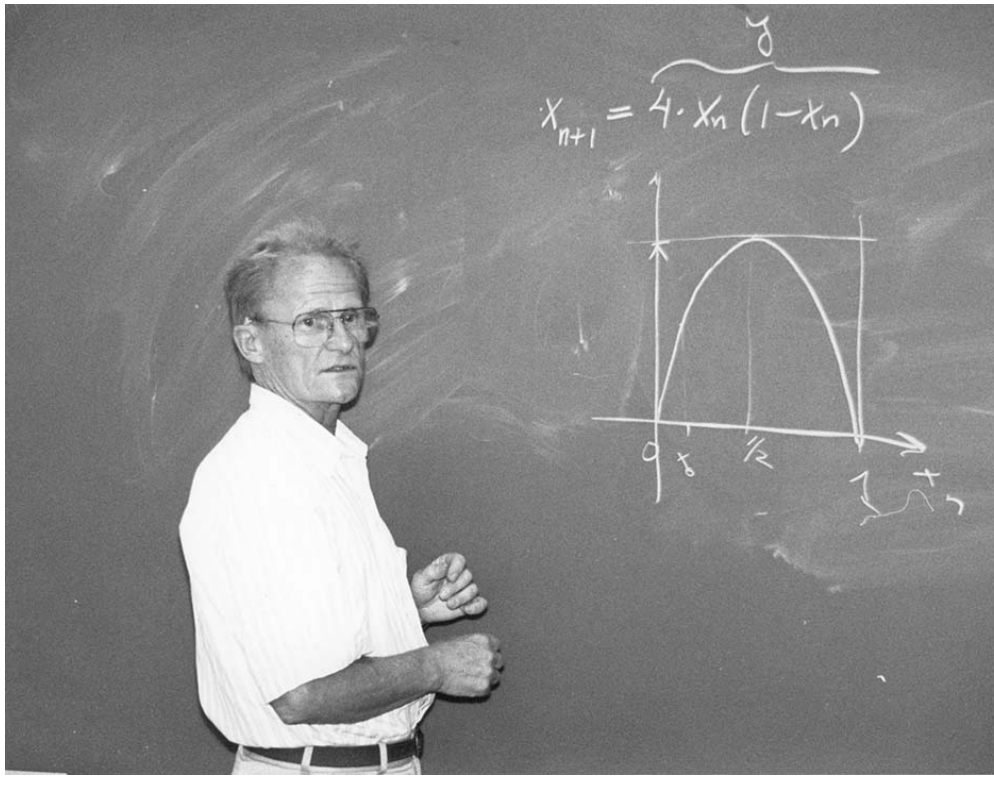
I had no contact with the Department of Mathematics at the University of Helsinki, but I did know the impressive Paul Kustaanheimo and of course I knew Olli Lokki who was at the HUT. In fact, my first course in statistics were lectures given at HUT by Olli Lokki, who was the pioneer of statistics teaching in HUT.

What were your plans for future after finishing your degree? How did you originally get interested in research career?

It seems now that I have always regarded research as the only career for me, which was only strengthened while working in IBM Research.

What about the teaching? Did you feel any call for teaching after finishing your degree?

I didn't have a chance until later when I found out that teaching is too difficult to be left to the professionals.



University of Tampere

Figure 5: Workshop in Tampere 1990. Jorma Rissanen talking on "Stochastic Complexity in Linear Models".

Sweden, 1958–64

You started your international career in Sweden in 1958. You worked at ASEA in Västerås? Could you tell something about that time?

I was told at the Helsinki City Utility Company that "by law we cannot fire you because of the military service, but don't come back". I never did. In fact I left the whole country and joined ASEA. I formulated and proved my very first theorem, one of the fundamental theorems in control (Rissanen 1961). It states that by a linear feedback you can move the poles wherever you wish but you cannot change their number.

You moved to the IBM Nordic Laboratory in Stockholm in 1960.

Yes, I did, and was sent for one year visit to Yorktown Heights and San Jose. These visits turned out to be crucial for my entire future.

Information technology and industry lived their early stages in these times. Do you find that this has some effect on your orientation in research?

The mission of the IBM Nordic Laboratories was industrial process control, but after a few years the company grew tired of waiting for economic results from that. The main problem was not control, which was easy if you knew what you were supposed to control. This turned my interests to modeling, and as it happens I never got out of it. The effect of information technology to my career was still some years away – and in fact came from an unexpected direction. But more of that a bit later.

Your first published works were on linear systems and prediction theory. Are we right?

Yes, I mentioned above my very first pole shifting theorem for linear feedback systems. In addition in the mid sixties I derived fast algorithms for factoring Hankel and Toeplitz matrices, which solve the so-called system identification problem as well as the Kalman type of prediction problem for ARMA processes (Rissanen 1973).

To the USA and to IBM, 1964–

You left Stockholm in 1964 and moved to the States. First to Electronic Associates in Princeton and Lockheed in Huntsville, but soon back to IBM's San Jose Research Laboratory, California. That was a time of many changes of residence. Tell us what made you to get off the ground.

Well, I needed a sponsor in the USA for emigration, and my boss in IBM USA during the visit had been promoted, which left me without a contact. I thought that who needs IBM and joined Electronic Associates. I realized however that maybe I had made a mistake and when IBM showed interest in getting us to San Jose we were just too happy to comply.

Did you join a research group in San Jose or did you work more or less independently? Tell us about the research culture and atmosphere at IBM.

I joined a research group but was given quite free hands to study and work on my own problems. This I took advantage of and became a professional student.

San Jose is geographically pretty close to some of America's greatest universities. Did you develop connections with people in Stanford and Berkeley?

Yes, indeed. I gave graduate lectures in Berkeley on prediction theory, and visited regularly professors Rudy Kalman and Tom Kailath in Stanford University. Later I actually was an adjunct professor in Stanford University.

What kind of working habits did you follow? How would you describe a normal day?

I started at about 8 AM and studied or worked on a problem leading to a paper until 5 PM. It took many years to get accustomed to such a regime without any obligations or directions – just study without focus and search for a meaningful problem. Every other day at noon time I played pickup soccer, which was a great distraction and broke the monotonous day.

In 1973–74 you held the chair of control theory in Linköping University, Sweden. It was not your cup of tea?

Immigrants often tend to feel nostalgia after a few years, and this happened to us as well. Also the lack of focus in my research had an effect, and when I saw the advertizement of a professor chair in Sweden I thought why not. It was a disastrous move. I found out that I don't like the field of control, I don't like to be a professor, and I don't like the climate nor the at that time socialistic Sweden. However, something happened, which maybe could not have happened otherwise. I was exposed to the exciting ideas of Chaitin, Kolmogorov, and Martin-Löf, which set my mind in fire. I found that this is what interests me, and finally I could focus on something that also could be of some interest to IBM. We returned to San Jose after just one year.

Did you have any personal contact with Kolmogorov?

I never had the honor of meeting him.

Did you meet Harald Cramér while you were in Sweden?

No, I never met him. But if you were to ask my statistical idols, I would definitely mention Harald Cramér and R. A. Fisher.

Arithmetic coding, stochastic complexity and MDL, 1975–

Many works you did at that time were somehow related to time series analysis. The Box–Jenkins approach to modeling ARIMA processes was described in a highly influential book by statisticians George Box and Gwilym Jenkins in 1970 (Box and Jenkins 1970). It seems that you already were familiar with these kinds of techniques at that time. Was it because of your system and prediction theory background?

Yes indeed. The modeling problems for AR and ARMA processes were bread and butter in the control field a decade before the Box–Jenkins book.

Later you had joint works with Ted Hannan who was one of the foremost experts in time series analysis. Would you like to say something about your collaborations with him?

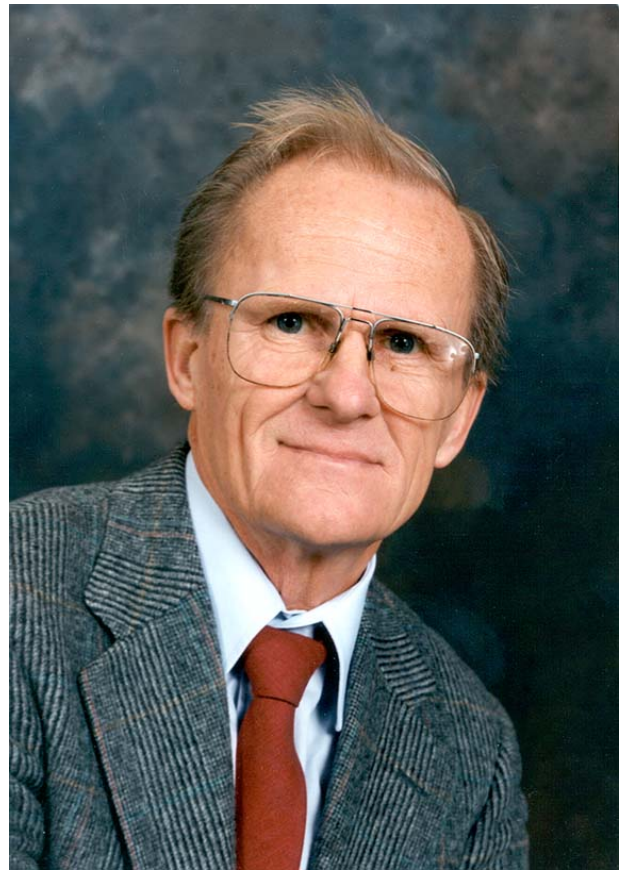


Figure 6: The IEEE announced that Jorma Rissanen is the recipient of the 1993 IEEE Richard W. Hamming Medal (IEEE Information Theory Society Newsletter 1993).

Ted loved to work on the difficult analysis problems arising in time series. His forte was analysis of existing problems rather than the creation of new ones. Among his many results was to derive a rather exact form for the penalty term in a criterion to find the number of parameters consistently. This also implies the consistency of the criterion BIC or the asymptotic approximation of the MDL criterion. To show something of my two-month visit to the Australian National University we had a joint *Biometrika* paper, written mostly by him, on estimation of ARMA order (Hannan and Rissanen 1982).

In 1975 you introduced a new coding technique for data compression, called Arithmetic Coding, *which is certainly one of your main contributions (Rissanen 1976). We may guess that coding theory was not one of your study subjects in the Helsinki University of Technology. Was it at IBM when you became interested in coding theory?*

I had never heard of the coding problem at the time I stayed at HUT. I got the idea of arithmetic coding from the brief paper by Kolmogorov, 'Three Measures of Information' (Kolmogorov 1965), during my stay as the professor of control theory in Sweden.

How do you see the status of Arithmetic Coding *nowadays?*

It is very simply the preferred way of doing coding for data compression. It has relegated the optimal, elegant, and then dominant Huffman codes to a graceful retirement – as someone put it.

You introduced in 1983 a universal modeling algorithm Context*, called now variable order Markov chains (Rissanen 1983). Tell us about the research that led to Context.*

I got the idea from the elegant data compression algorithm by Lempel and Ziv. Since I was advocating the view that all universal data compression algorithms must incorporate a model of the data, I first reinterpreted the main part in the LZ algorithm as a universal model in the huge class of ergodic processes. This then led to the universal modeling algorithm Context for the much smaller class of Markov chains. Because the class is smaller the model cost is smaller too, and you get a better compression if the data have any properties like the Markov chains.

Many people would like to know what led you into formulating the principle that is now known as the MDL principle?

Since with arithmetic coding you can encode any data modeled in any statistical manner in a completely mechanical and uniform way the key problem in data compression is to understand the statistical behavior of the data, which is modeling. I then turned the problem around and concluded that it is possible to measure the goodness of any model by its ability to compress the data. This relationship becomes evident by Kraft inequality, which establishes the logical equivalence between a distribution and the lengths of a (prefix) code tree.

Was the MDL principle formulated the first time in your Automatica *paper "Modeling by shortest data description" in 1978, or even before?*

Yes. The MDL principle as a concept was clear to me after the introduction of arithmetic coding in around 1975–1976, and in the *Automatica* paper I wanted to explain the principle to an audience not familiar with coding but one which I was familiar with, the control theory people.

You were inspired by the Kolmogorov's algorithmic theory of complexity. There are also the related works by Ray Solomonoff, G. J. Chaitin and Per Martin-Löf. Do you think that their works had some influence on your thinking?

In a fundamental way. You see, it is one thing to understand things intuitively but quite different to see formally the relationships, which is what I learned from the writings of these distinguished men. In fact, there are still certain baffling both conceptual and technical issues, which I hope to be able to sort out some day.

The idea of estimation-via-coding was presented in the computer science literature by Wallace and Boulton 1968. How is their approach related to MDL?

First, their principle is expressed as minimizing the mean code length, which is then estimated. This results in a two-part code, which Wallace and his students have applied to a number of practical cases. The MDL principle on the other hand has been developed into a theory of inference rather than just a criterion for model selection.

Figure 7: Peter Grünwald and Jorma Rissanen during the 2002 IEEE Information Theory Workshop in Bangalore, India.

Your paper 'Stochastic complexity and modeling' in Annals of Statistics *1986 extends certain fundamental results in information theory and statistics. Could you tell us in a few words about the significance of this contribution.*

At the time I wanted to prove that with a parametric family of distributions you cannot get a shorter code length for or, equivalently, a greater probability assigned to data sequences than a certain calculable bound – no matter how you construct the code. That is, such a bound, which I called stochastic complexity, is inherent for the family of distributions at hand. Later constructions were found with which the bound can be essentially achieved. This simply means that there are some absolute restrictions in modeling which you have to live with, and if your model gets away with such restrictions nobody can beat you. This is the aim with universal models.

During the last three decades you have developed such ideas as stochastic complexity, universal model, and universal minimal sufficient statistics, which provide an information theoretic foundation for statistics. How do you assess the present status of your theory now?

The universal minimal sufficient statistics in the algorithmic information theory, where a model of data is simply any finite set that includes the data, is due to Kolmogorov. When we deal with distributions as models, which must be estimated from data, there is a possibility to sharpen Kolmogorov's result at least in two respects: First, the code length of a model, represented by the maximum likelihood estimate of its parameters, corresponds to the Kolmogorov complexity of the model. But the model can actually be represented more completely by the sequence of the maximum likelihood estimates, obtained from all the prefixes of the data sequence. This has no

counter part in the algorithmic theory, and clearly there is more information about the model in the sequence of the maximum likelihood estimates than in just the last estimate. Hence, the idea of 'information' that an estimated model represents is now different and more complex.

The second aspect which is missing in the algorithmic models is the idea of models that can be optimally 'distinguished' from a given set of data. I have written about how this provides a sense of optimality in hypothesis testing and greatly reduces the number of hypotheses we need to test.

My most recent work is of a new generation of universal models, which, while related to the predictive ways of constructing universal models like Dawid's prequential 'plug-in' models or the equivalent predictive MDL models, are strictly better – provably so.

If you were asked to list your three most important contributions, what would you list?

The two most important contributions are undoubtedly Arithmetic Coding and the MDL principle. The third is tougher to pick. The theorem on the lower bound for the code length achievable for parametric families mentioned above together with the latest extensions of the MDL theory rank in my mind higher than Algorithm Context – perhaps because of the much greater difficulty in deriving them. In fact, I understand the proof in the *Annals of Statistics* paper only in my brighter moments.

Philosophy

In research, which is more important: conceptual foundations or technical perfection?

I think the conceptual foundations are more important although their clarifications require often difficult analysis, which, moreover, tends to modify the initial concepts, and hence we cannot really separate the two.

Do you see the computer as a tool in theoretical work?

Computations of anything I have proposed have influenced my theoretical work (shown that I had overlooked something).

What is your approach to a problem? Do you have certain techniques or would you say that is intuitive?

My intuition amounts to seeing, or better feeling, what's essential in the problem. Then by analysis comes a sharper isolation of that essential, which amounts to better understanding of the problem. The rest is detail – albeit sometimes crucial detail. An example is coding, in which the essential is just sorting and counting. However, the crucial thing is to understand exactly what it is that should be counted. In fact, this is also behind the entropy and the algorithmic information, so that in essence Kolmogorov's three measures of information could be reduced to one. Clearly, to find the essence can be difficult, although in case of arithmetic coding it was not difficult, because of Kolmogorov's paper, where the essence was spelled out. Nevertheless, after the invention of arithmetic codes my boss cum secretary exclaimed that he could not have invented that in a million years, or, more realistically, a million men could not find it in a year. In truth, the story of arithmetic coding is more involved than what I made it out here to be.

Figure 8: Jorma Rissanen and Stefano Zambelli at the home of K. Vela Velupillai in Galway, Ireland, March 2005.

What is your view on breadth versus depth in research? You have solved problems over a broad range and reached many deep findings. Some people tend to think that working deeply on a small topic is in contrast to having a wider interest.

I see no contradiction in the two views. I sometimes tell young people that you need mathematics and deep analytic tools if you want to accomplish anything of lasting value; you cannot do that simply by being clever in the small.

In this context I might mention a nice poem by Piet Hein, which I saw in the office door of Terry Speed in Berkeley: Its name is Wide Road and it goes as follows:

> To make a name for learning
> when other roads are barred,
> take something very easy
> and make it very hard.

You have developed foundations of statistical modeling and carried out significant research in many fields like e.g. prediction and system theories, information and coding theory, computer learning, probability and statistics, and you have worked with computer scientists, engineers, information theorists, mathematicians, and statisticians. Could you give us your projections for statistical modeling in this new century?

In my view statistics needs a solid foundation rather than just a collection of isolated techniques however clever.

It is not enough to claim to have found a method which works well or better than other methods on some data. It's even not enough to prove that the technique works on data generated by an imagined 'true' distribution. We need to understand why a technique works as it does, and why it is better than the competing approaches. This is what's missing in current statistics,

where all sorts of criteria for the model selection problem have been proposed. In addition, a sound statistical theory should be able to treat the estimation of both the parameter values and their number within a common theory. We ought to be able to formalize ideas like 'information', 'complexity', and 'noise', which I'm afraid can be done only with information theoretic means. As a specific example, there are good techniques for denoising. However, unless the idea of 'noise' is defined it becomes the part in the data that is removed. I hope that statistics will progress in a manner which makes sense and in which the fundamental concepts are defined and the limitations clearly understood.

In many applications we are going to be faced with a lot of data generated by computational statistics or by a measuring device and then the theory has to keep up with it. Are we in a situation where there is more information than theory? Do you think there will be a revolution in statistical theory?

I already explained above that the traditional statistics is unable to formalize the central concepts that are needed to capture properties in complex data, so that indeed there is more in the data than what the traditional statistical theory can explain. In my opinion, there will be drastic changes in statistics although one may suspect that the changes will be gradual, perhaps disguised as modifications of old approaches so that no foundation need be changed. Currently, foundations, such as they are, are ignored.

Some writers seem to mix up MDL with Bayesian procedures. What is your view of the Bayesian philosophical framework?

The MDL theory is based upon the MDL principle, which opens up an entirely different approach to statistics, free from the untenable assumption of a 'true' data generating distribution, while the Bayesian philosophy has no principle other than an unrestricted use of probabilities. Instead the central concept is the posterior distribution, whose interpretation is just as fuzzy as that of the prior. Moreover, since the prior affects the posterior its selection is crucial. In the MDL theory, where the use of priors is optional, their selection must be restricted so that they are encodable. This permits optimization of their selection, which cannot be done within the Bayesian philosophy, because nothing prevents you from putting it to unity on the data. The irresistible desire to peek into the data has created concepts like 'empirical' priors which clearly contradicts the very foundation of Bayesianism.

The usual confusion between the two approaches is understandable if one equates the MDL principle with the early criterion for model selection, also derived by Bayesian arguments and called BIC. Even though they are identical the MDL derivation attaches an asymptotic optimality to the criterion, while no such status can be given to BIC in terms of the Bayesian concepts. As a result the MDL criterion has been refined and developed further, while the BIC is a dead end.

Finally, the concepts like universal models, noise, statistical information, and complexity, which are central in the MDL theory, have no meaning in the Bayesian philosophy. It is true that Bayes' formula creates a universal model, the so-called Bayesian mixture, which is good, but not because its goodness could be assessed by Bayesian means. Rather, it is good because it reaches the lower bound referred to above. Moreover, there are other universal models, which have not been found in the Bayesian philosophy and which have properties preferable to the Bayesian mixture. In summary, the MDL principle has created an entire theory with new concepts, which goes beyond any Bayesian technique and Bayesian philosophy whatever that is.

What definition of probability do you use?

Strictly speaking I use the probability that satisfies Kolmogorov's axioms. Sometimes, when it is preferable to talk about code length I mean by probability the number two raised to the negative power of the code length. After all, prefix codes are very concrete and there is nothing fuzzy about them. An example is the assignment of probability to a closed curve on the plane. It is easy to encode such curves by the chain link method, which then assigns a probability to the curve. Compare this with the horrendous task of defining a prior distribution for the set of all continuous closed curves on the plane.

Teemu Roos

Figure 9: Jorma Rissanen with Wojciech Szpankowski and Jacob Ziv at the ITA Workshop, San Diego, California, 2007.

Affiliations to Finland and current research

You were invited to "The 2nd International Tampere Conference in Statistics" in 1987. In consequence of this meeting we had the privilege to learn to know you personally. Since then you have been a regular visitor to Tampere. In 1987 Conference we had T. W. Anderson, George E. P. Box, C. R. Rao and Ted Hannan as keynote speakers. You surely had met all of them before?

I do not recall having met Professor Box before that meeting. Actually, as you may know, I met Tarmo Pukkila, who was in charge for the Tampere Conference, first time in 1985 in Las Vegas in the ASA Conference – and there Tarmo invited me to Tampere in 1987.

And now you even have a flat in Hervanta, Tampere. How do you share your time between California, Finland, and the numerous conference and lecturing trips?

I spend about two months each year in Finland on three different time periods. The conference and lecture trips do not take that much of my time.

You have been invited to numerous conferences, visited a number of universities, and have delivered many prestigious lectures, e.g. the Kolmogorov Lecture 2006. Could you mention some highlights?

In addition to the Kolmogorov lecture I remember with pleasure a talk I gave in Norbert Wiener's 100 year memorial meeting in 1994, and the visit to Ann Arbor 1986, where Shannon himself handed me the best paper award. I also remember lectures in Beijing in 2005 as a guest of Microsoft, which provided a car with the driver for a week, and a memorial meeting of Z. C. Wei in Academia Sinica in Taipei 2005.

From May 1995 until your mandatory retirement you held a part time professorship at the Tampere University of Technology. Tell us now how you came to TUT.

TUT made me an honorary doctor in 1992, which followed with a part time professorship. When I had to retire from that TUT made other special arrangements, which allowed me to visit Tampere three times each year.

Now every year you are teaching a course on statistical modeling at TUT, you are a fellow of the Helsinki Institute for Information Technology, and you have joint research projects in both places. Could you tell about these activities?

I enjoy them greatly. Now I have access to peers and graduate students, which I never had in IBM Research.

What do you want to say about your current research interests?

I'm involved in applications of the MDL principle to practical problems. I'm also involved in theoretical work, some of which has been inspired by the applications.

In addition to papers by only yourself you have plenty of papers with your collaborators. Can you shortly comment on the role of collaboration in your research?

Although I have mostly worked alone collaboration has been quite important, in particular on arithmetic coding, whose practical implementation would have been outside of my skills.

What about your PhD students? You surely must have a bunch of them?

I have had access to PhD students mostly only in Finland, and I have found them both useful and invigorating. They tend to come up with unexpected questions, which expose embarrassing shortcomings in my original suggestions. Also, of course, their superb programming skills make the applications possible, which, in turn, create further problems.

Personal

You officially had to take the mandatory retirement from your part time professorship almost ten years ago. Do you consider yourself retired?

I had no mandatory retirement from IBM Research. Rather, having already been involved in the institutions in Finland and London, I decided to retire from IBM five years ago and spend more time in Finland. I certainly do not consider myself retired. I often think that I should have retired from IBM earlier.

What things do you like to do when you are not doing research? Tell us a little bit about your life outside of research and your hobbies or other activities.

I must say that nowadays, when I don't play soccer any more, I have very little activities other than research. I walk my dog and work on the difficult to maintain yard. We live on a wild mountain side.

You have held onto your passion for soccer. Tell us what kind of resonance soccer and bandy play have had for you.

When I grew up in Kemi this game 'bandy', which is like soccer on ice with skates and a curved club to control a small ball, was my passion in winter and soccer in summer. I started to play recreational soccer in California in 1965 until my retirement in 2002. I would still play if there was an 'over 70 team' (no younger than 70 years old allowed in the team). Unfortunately, there does not seem to be enough players of that age.

As far as we remember you enjoy reading mysteries. Who are your favorite writers?

The British writers Ken Follet, Frederic Forsyth, Jack Higgins, and the Americans Clive Cussler, and Dan Brown.

You did not learn English at school? What was your way of learning your perfect English?

First, my command of English is far from perfect, but having lived in California for 40 years has helped.

What does the future hold for you?

I don't know.

Yes indeed, we don't know what the future holds. We wish you well! Thanks very much Jorma for sharing your thoughts about your career and about science with us.

I thank you for bothering to find out my thinking and making me feel as if I had done something worthwhile.

Ioan Tăbuș

Figure 10: Jorma Rissanen looking at the sketch of a plot on sale in Kuru, Finland, April 2007.

References

G. E. P. Box, G. M. Jenkins (1970). *Time Series Analysis: Forecasting and Control*. Holden-Day, New York. [Third edition by G. E. P. Box, G. M. Jenkins and G. C. Reinsel, Prentice Hall, Englewood Cliffs, NJ, 1994.]

E. J. Hannan, J. Rissanen (1982). Recursive estimation of mixed autoregressive-moving average order. *Biometrika*, 69, 81–94.

A. N. Kolmogorov (1965). Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1, 4–7.

J. Rissanen (1961). Control system synthesis by analogue computer based on the "generalized linear feedback" concept. In *Analogue computation applied to the study of chemical processes, international seminar. Le calcul analogique appliqué à l'étude des processus chimiques, séminaire international. Actes/Proceedings.* (Robert Vichnevetsky, ed.), Gordon & Breach, New York.

J. Rissanen (1973). Algorithms for triangular decomposition of block Hankel and Toeplitz matrices with application to factoring positive matrix polynomials. *Mathematics of Computation*, 27, 147–154.

J. Rissanen (1976). Generalized Kraft inequality and arithmetic coding. *IBM Journal of Research and Development*, 20, 198–203.

J. Rissanen (1978). Modeling by shortest data description. *Automatica*, 14, 465–471.

J. Rissanen (1983). A universal data compression system. *IEEE Transactions on Information Theory*, 29, 656–664.

J. Rissanen (1986). Stochastic complexity and modeling. *Annals of Statistics*, 14, 1080–1100.

J. Rissanen (1989). *Stochastic Complexity in Statistical Inquiry.* World Scientific, Teaneck, NJ.

J. Rissanen (2007). *Information and Complexity in Statistical Modeling.* Springer, New York.

C. S. Wallace, D. M. Boulton (1968). An information measure for classification. *Computing Journal*, 11, 185–195.

Acknowledgements

Special thanks go to Jarmo Niemelä for his assistance with LaTeX and for his help with the final editing of this interview, and in particular, for fixing up the list of Jorma Rissanen's research publications. Thanks go also to Peter Grünwald, Virve Larmila, Teemu Roos, Terry Speed, Ioan Tăbuş and K. Vela Velupillai for their help.

Research publications of Jorma Rissanen

Monographs and edited books

[A1] Jorma Rissanen (1989). *Stochastic complexity in statistical inquiry.* World Scientific. (Series in Computer Science.) ISBN: 978-997150859-3.

[A2] George Cybenko, Dianne P. O'Leary, Jorma Rissanen (eds.) (1999). *The mathematics of information coding, extraction, and distribution.* Proceedings of a workshop that was an integral part of the 1996–97 IMA program on Mathematics in high-performance computing. Univ. of Minnesota, Minneapolis, MN, USA, November 11–15, 1996. Springer. (The IMA Volumes in Mathematics and its Applications, vol. 107.) ISBN: 978-0-387-98665-4.

[A3] Jorma Rissanen (2007). *Information and complexity in statistical modeling.* Springer. (Series: Information Science and Statistics.) ISBN: 978-0-387-36610-4.

Articles in refereed journals and collections

[B1] J. Rissanen (1961). Control system synthesis by analogue computer based on the "generalized linear feedback" concept. In *Analogue computation applied to the study of chemical processes, international seminar. Le calcul analogique appliqué à l'étude des processus chimiques, séminaire international. Actes/Proceedings.* (Robert Vichnevetsky, ed.), Gordon & Breach, New York.

[B2] J. Rissanen (1963). On the theory of self-adjusting models. *Automatica*, 1 (4), 297–309.

[B3] J. Rissanen (1966). Performance deterioration of optimum systems. *IEEE Transactions on Automatic Control*, 11 (3), 530–532.

[B4] N. McClamroch, J. Rissanen (1967). A result on the performance deterioration of optimum systems. *IEEE Transactions on Automatic Control*, 12 (2), 209–210.

[B5] J. Rissanen (1967). On duality without convexity. *Journal of Mathematical Analysis and Applications*, 18 (2), 269–275.

[B6] A. Chang, J. Rissanen (1968). Regulation of incompletely identified linear systems. *SIAM Journal on Control and Optimization*, 6 (3), 327–348.

[B7] J. Rissanen, L. Barbosa (1969). A factorization problem and the problem of predicting nonstationary vector-valued stochastic processes. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 12, 255–266.

[B8] J. Rissanen, L. Barbosa (1969). Properties of infinite covariance matrices and stability of optimum predictors. *Information Sciences*, 1 (3), 221–236.

[B9] J. Rissanen (1970). On factoring positive operators. *Journal of Mathematical Analysis and Applications*, 32 (3), 505–511.

[B10] J. Rissanen (1970). Reproof of a theorem on duality without convexity. *Journal of Mathematical Analysis and Applications*, 29 (2), 429–431.

[B11] Thomas Kailath, J. Rissanen (1971). The stochastic realization problem. In *Proceedings of the 1971 IEEE Conference on Decision and Control including the 10th Symposium on Adaptive Processes, Miami Beach, FL, 15–17 December 1971*, vol. 10, p. 546.

[B12] J. Rissanen (1971). Maximum power feedback laws. *International Journal of Control, I. Ser.*, 14, 233–240.

[B13] J. Rissanen (1971). On optimum root-finding algorithms. *Journal of Mathematical Analysis and Applications*, 36 (1), 220–225.

[B14] J. Rissanen (1971). Recursive identification of linear systems. *SIAM Journal on Control and Optimization*, 9, 420–430.

[B15] J. Rissanen (1972). Recursive evaluation of Padé approximants for matrix sequences. *IBM Journal of Research and Development*, 16, 401–406.

[B16] J. Rissanen, T. Kailath (1972). Partial realization of random systems. *Automatica*, 8 (4), 389–396.

[B17] J. Rissanen, T. Kailath (1972). Partial realization of random systems. In *Proceedings of the IFAC Fifth World Congress (Paris, 1972), Part 4: Education, feedback, regulators, linear and nonlinear systems; Identification, differential games, discrete and stochastic systems, Paper No. 35.6*, International Federation of Automatic Control, Düsseldorf, p. 6.

[B18] J. Rissanen (1973). Algorithms for triangular decomposition of block Hankel and Toeplitz matrices with application to factoring positive matrix polynomials. *Mathematics of Computation*, 27, 147–154.

[B19] J. Rissanen (1973). Bounds for weight balanced trees. *IBM Journal of Research and Development*, 17, 101–105.

[B20] J. Rissanen (1973). A fast algorithm for optimum linear predictors. *IEEE Transactions on Automatic Control*, 18 (5), 555.

[B21] P. E. Caines, J. Rissanen (1974). Maximum likelihood estimation of parameters in multivariate Gaussian stochastic processes (corresp.). *IEEE Transactions on Information Theory*, 20 (1), 102–104.

[B22] J. Rissanen (1974). Basis of invariants and canonical forms for linear dynamic systems. *Automatica*, 10, 175–182.

[B23] Jorma Rissanen (1974). Solution of linear equations with Hankel and Toeplitz matrices. *Numerische Mathematik*, 22, 361–366.

[B24] J. Rissanen, Bostwick F. Wyman (1975). Duals of input/output maps. In *Category theory applied to computation and control (Proc. First Internat. Sympos., San Francisco, Calif., February 25–26, 1974), Lecture Notes in Computer Science*, vol. 25, Springer, Berlin, pp. 204–208.

[B25] Jorma Rissanen (1975). Canonical Markovian representations and linear prediction. In *Proceedings of the IFAC 6th World Congress (Boston/Cambridge, Mass., 1975), Part 1, Paper No. 29.3*, International Federation of Automatic Control, Düsseldorf, p. 9.

[B26] J. Rissanen (1976). Minimax entropy estimation of models for vector processes. In *System identification: Advances and Case Studies* (D. G. Lainiotis, R. K. Mehra, eds.), *Mathematics in Science and Engineering*, vol. 126, Academic Press, New York, pp. 97–117.

[B27] J. Rissanen, L. Ljung (1976). Estimation of optimum structures and parameters for linear systems. In *Mathematical systems theory: Proceedings of the international symposium held in Udine, Italy, June 16–27, 1975, Lecture Notes in Economics and Mathematical Systems*, vol. 131, Springer, Berlin, pp. 92–110.

[B28] J. J. Rissanen (1976). Generalized Kraft inequality and arithmetic coding. *IBM Journal of Research and Development*, 20 (3), 198–203.

[B29] Jorma Rissanen (1976). System estimation by entropy criterion. In *Proceedings of the Ninth Hawaii International Conference on System Sciences (Univ. Hawaii, Honolulu, Hawaii, 1976)*, Western Periodicals, North Hollywood, Calif., pp. 199–200.

[B30] Jorma Rissanen (1977). Independent components of relations. *ACM Transactions on Database Systems*, 2 (4), 317–325.

[B31] Lennart Ljung, Jorma Rissanen (1978). On canonical forms, parameter identifiability and the concept of complexity. In *Proceedings of the 4th IFAC Symposium on Identification and system parameter estimation, Tbilisi, USSR, 1976, Part 3* (N. S. Rajbman, ed.), North-Holland, Amsterdam, pp. 1415–1426.

[B32] J. Rissanen (1978). Modeling by shortest data description. *Automatica*, 14, 465–471.

[B33] J. Rissanen (1978). Theory of relations for databases – a tutorial survey. In *Mathematical foundations of computer science 1978: Proceedings, 7th Symposium, Zakopane, Poland, 1978* (Józef Winkowski, ed.), *Lecture Notes in Computer Science*, vol. 64, Springer-Verlag, Berlin, pp. 537–551.

[B34] Jorma Rissanen (1978). Minimax codes for finite alphabets (corresp.). *IEEE Transactions on Information Theory*, 24 (3), 389–392.

[B35] J. Rissanen (1979). Shortest data description and consistency of order estimates in ARMA-processes. In *International Symposium on Systems Optimization and Analysis (Rocquencourt, 1978)*, *Lecture Notes in Control and Information Sciences*, vol. 14, Springer, Berlin, pp. 92–98.

[B36] J. Rissanen, P. E. Caines (1979). The strong consistency of maximum likelihood estimators for ARMA processes. *The Annals of Statistics*, 7 (2), 297–315.

[B37] J. Rissanen, G. G. Langdon, Jr. (1979). Arithmetic coding. *IBM Journal of Research and Development*, 23 (2), 149–162.

[B38] Jorma Rissanen (1979). Arithmetic codings as number representations. *Acta Polytechnica Scandinavica, Mathematics and Computer Science Series*, 31, 44–51.

[B39] J. Rissanen (1980). Consistent order estimates of autoregressive processes by shortest description of data. In *Analysis and optimisation of stochastic systems. Based on the Proceedings of the International Conference held at the University of Oxford from 6–8 September, 1978, organised by The Institute of Mathematics and its Applications* (O. Jacobs, M. Davis, M. Dempster, C. Harris, P. Parks, eds.), Academic Press, New York.

[B40] G. Langdon, J. Rissanen (1981). Compression of black-white images with arithmetic coding. *IEEE Transactions on Communications*, 29 (6), 858–867.

[B41] Jorma Rissanen (1981). Order estimation in Box-Jenkins model for time series. *Methods of Operations Research*, 44, 143–150.

[B42] Jorma Rissanen, Glen G. Langdon, Jr. (1981). Universal modeling and coding. *IEEE Transactions on Information Theory*, 27 (1), 12–23.

[B43] E. J. Hannan, J. Rissanen (1982). Recursive estimation of mixed autoregressive-moving average order. *Biometrika*, 69 (1), 81–94.

[B44] Glen G. Langdon, Jr., Jorma Rissanen (1982). A simple general binary source code. *IEEE Transactions on Information Theory*, 28 (5), 800–803.

[B45] J. Rissanen (1982). Estimation of structure by minimum description length. *Circuits, Systems, and Signal Processing*, 1 (3–4), 395–406.

[B46] Jorma Rissanen (1982). On equivalences of database schemes. In *Proceedings of the ACM symposium on principles of database systems, March 29–31, 1982, Los Angeles, California*, Association for Computing Machinery, pp. 23–26.

[B47] Jorma Rissanen (1982). Tight lower bounds for optimum code length. *IEEE Transactions on Information Theory*, 28 (2), 348–349.

[B48] E. J. Hannan, J. Rissanen (1983). Errata: "Recursive estimation of mixed autoregressive-moving average order" [Biometrika 69 (1), 81–94, 1982; MR0655673 (84e:62136)]. *Biometrika*, 70 (1), 303.

[B49] Glen G. Langdon, Jr., Jorma Rissanen (1983). Correction to "A simple general binary source code" (Sep 1982, 800–803). *IEEE Transactions on Information Theory*, 29 (5), 778–779.

[B50] Glen G. Langdon, Jr., Jorma J. Rissanen (1983). A double-adaptive file compression algorithm. *IEEE Transactions on Communications*, 31 (11), 1253–1255.

[B51] J. Rissanen (1983). Probability estimation for symbols – observed or not. In *International Symposium on information theory held at Quebec, Canada on September 26–30, 1983. Abstracts of papers*, Institute of Electrical and Electronics Engineers Inc., New York.

[B52] Jorma Rissanen (1983). Information in prediction and estimation. In *Proceedings of the 22nd IEEE Conference on Decision and Control, San Antonio, TX, 14–16 December 1983*, vol. 22, pp. 308–310.

[B53] Jorma Rissanen (1983). A universal data compression system. *IEEE Transactions on Information Theory*, 29 (5), 656–664.

[B54] Jorma Rissanen (1983). A universal prior for integers and estimation by minimum description length. *The Annals of Statistics*, 11 (2), 416–431.

[B55] K. Mohiuddin, J. J. Rissanen, R. Arps (1984). Lossless binary image compression based on pattern matching. In *Proceedings of the International Conference on Computers, Systems, and Signal Processing, December 1984, Bangalore, India*, pp. 447–451.

[B56] Jorma Rissanen (1984). Universal coding, information, prediction, and estimation. *IEEE Transactions on Information Theory*, 30 (4), 629–636.

[B57] J. Rissanen (1985). Minimum description length principle. In *Encyclopedia of Statistical Sciences* (S. Kotz, N. L. Johnson, eds.), vol. 5, John Wiley and Sons, pp. 523–527.

[B58] J. Rissanen, V. Wertz (1985). Structure estimation by accumulated prediction error criterion. In *Proceedings of the 7th IFAC/IFORS Symposium on Identification and System Parameter Estimation held in 3–7 July 1985 in York, England* (H. A. Barker, Young P. C., eds.), Pergamon Press.

[B59] Stephen Todd, Glen G. Langdon, Jr., Jorma Rissanen (1985). Parameter reduction and context selection for compression of gray-scale images. *IBM Journal of Research and Development*, 29 (2), 188–193.

[B60] L. Gerencsér, J. Rissanen (1986). A prediction bound for Gaussian ARMA processes. In *Proceedings of the 25th IEEE Conference on Decision and Control, Athens, Greece*, vol. 3, pp. 1487–1490.

[B61] Jorma Rissanen (1986). Complexity of strings in the class of Markov sources. *IEEE Transactions on Information Theory*, 32 (4), 526–532.

[B62] Jorma Rissanen (1986). Order estimation by accumulated prediction errors. *Journal of Applied Probability*, Special vol. 23A: Essays in time series and allied processes, 55–61.

[B63] Jorma Rissanen (1986). Predictive and nonpredictive minimum description length principles. In *Time series and linear systems* (S. Bittanti, ed.), *Lecture Notes in Control and Information Sciences*, vol. 86, Springer, Berlin, pp. viii, 115–140.

[B64] Jorma Rissanen (1986). A predictive least squares principle. *IMA Journal of Mathematical Control and Information*, 3 (2–3), 211–222. Special issue: Parametrization problems.

[B65] Jorma Rissanen (1986). Stochastic complexity and modeling. *The Annals of Statistics*, 14 (3), 1080–1100.

[B66] Jorma Rissanen (1986). Stochastic complexity and statistical inference. In *Analysis and optimization of systems (Antibes, France, 1986)*, *Lecture Notes in Control and Information Sciences*, vol. 83, Springer, Berlin, pp. 393–407.

[B67] J. Rissanen (1987). Complexity and information in contingency tables. In *Proceedings of The Second International Tampere Conference in Statistics, June 1–4, 1987, Tampere, Finland* (Tarmo Pukkila, Simo Puntanen, eds.), Report A 184, University of Tampere, Department of Mathematical Sciences, Tampere, Finland. Invited paper.

[B68] Jorma Rissanen (1987). Comment on "Rational transfer function approximation" by E. J. Hannan. *Statistical Science*, 2 (2), 156–157.

[B69] Jorma Rissanen (1987). Stochastic complexity and the MDL principle. *Econometric Reviews*, 6 (1), 85–102.

[B70] Jorma Rissanen (1987). Stochastic complexity (with discussion). *Journal of the Royal Statistical Society. Series B. Methodological*, 49 (3), 223–239, 253–265.

[B71] Jorma Rissanen, Mati Wax (1987). Measures of mutual and causal dependence between two time series. *IEEE Transactions on Information Theory*, 33 (4), 598–601.

[B72] E. J. Hannan, J. Rissanen (1988). The width of a spectral window. *Journal of Applied Probability*, Special vol. 25A: A celebration of applied probability, 301–307.

[B73] Jorma Rissanen (1988). Comment on "ARMA memory index modeling of economic time series" by H. J. Bierens. *Econometric Theory*, 4 (1), 61–64.

[B74] Jorma Rissanen (1988). Stochastic complexity and the maximum entropy principle. In *Maximum-entropy and Bayesian methods in science and engineering, vol. 1 (Laramie, WY, 1985, and Seattle, WA, 1986/1987)*, Fundamental Theories of Physics, Kluwer Acad. Publ., Dordrecht, pp. 161–171.

[B75] J. Rissanen, K. M. Mohiuddin (1989). A multiplication-free multialphabet arithmetic code. *IEEE Transactions on Communications*, 37 (2), 93–98.

[B76] J. Rissanen (1990). Complexity of models. In *Complexity, Entropy and the Physics of Information: Proceedings of the Santa Fe Institute Workshop, May 29 to June 10, 1989* (Wojciech H. Zurek, ed.), *Santa Fe Institute studies in the sciences of complexity*, vol. 8, Westview Press, pp. 117–126.

[B77] J. Rissanen (1992). Discussion of "Prequential analysis, stochastic complexity and Bayesian inference" by A. P. Dawid. In *Bayesian Statistics 4: Proceedings of the Fourth Valencia International Meeting* (J. M. Bernardo, J. O. Berger, A. P. Dawid, A. F. M. Smith, eds.), Oxford University Press, Oxford, pp. 121–122.

[B78] Jorma Rissanen, Terry P. Speed, Bin Yu (1992). Density estimation by stochastic complexity. *IEEE Transactions on Information Theory*, 38 (2), 315–323.

[B79] László Gerencsér, Jorma Rissanen (1993). Asymptotics of predictive stochastic complexity. In *New directions in time series analysis, Part II. Proceedings of the 1990 IMA Workshop* (D. Brillinger, P. Caines, J. Geweke, E. Parzen, M. Rosenblatt, M. S. Taqqu, eds.), *IMA Volumes in Mathematics and its Applications*, vol. 46, Springer, New York, pp. 93–112.

[B80] J. Rissanen (1993). The minimal description length principle. In *Encyclopedia of Microcomputers* (Allen Kent, James G. Williams, eds.), vol. 11, Marcel Dekker, New York, NY, pp. 151–158.

[B81] M. J. Weinberger, M. Feder, J. Rissanen (1993). Sequential model estimation for universal coding and the predictive stochastic complexity of finite-state sources. In *Proceedings of IEEE International Symposium on Information Theory, San Antonio, TX, January 17–22, 1993*, Institute of Electrical and Electronics Engineers, inc, Piscataway, NJ, p. 52.

[B82] L. Gerencsér, J. H. van Schuppen, J. Rissanen, Z. Vágó (1994). Stochastic complexity, selftuning and optimality. In *Proceedings of the 33rd IEEE Conference on Decision and Control, Lake Buena Vista, FL, 14–16 December 1994*, vol. 1, pp. 652–654.

[B83] J. Rissanen (1994). Noise separation and MDL modeling of chaotic processes. In *From Statistical Physics to Statistical Inference and Back. Proceedings of NATO Summer School* (P. Grassberger, J.-P. Nadal, eds.), NATO ASI Series, Kluwer Academic Publishers, pp. 317–330.

[B84] Jorma Rissanen, Eric Sven Ristad (1994). Language acquisition in the MDL framework. In *Language computation. DIMACS workshop on human language, March 20–22, 1992, at Princeton Univ., Princeton, NJ, USA* (Eric Sven Ristad, ed.), *DIMACS Ser. Discrete Math. Theor. Comput. Sci.*, vol. 17, American Mathematical Society, Providence, RI, pp. 149–166.

[B85] Jorma Rissanen, Eric Sven Ristad (1994). Unsupervised classification with stochastic complexity. In *Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: an Informational Approach (Knoxville, TN, 1992)* (Hamparsum Bozdogan, ed.), vol. 2, Kluwer Academic Publ., Dordrecht, pp. 14–15, 171–182.

[B86] Søren Forchhammer, Jorma Rissanen (1995). Coding with partially hidden Markov models. In *Proceedings of the IEEE Data Compression Conference, DCC 1995, Snowbird, Utah, March 28–30, 1995* (James A. Storer, Martin Cohn, eds.), IEEE Computer Society Press, pp. 92–101.

[B87] L. Gerencsér, J. H. van Schuppen, J. Rissanen, Z. Vágó (1995). Parametric uncertainty and control performance in stochastic adaptive control. In *Proceedings of the 5th IFAC Symposium on Adaptive Systems in Control and Signal Processing (ACASP'95), Budapest, Hungary, 1995*, pp. 79–82.

[B88] Manish Mehta, Jorma Rissanen, Rakesh Agrawal (1995). MDL-based decision tree pruning. In *Proceedings of the First International Conference on Knowledge Discovery in Databases and Data Mining (KDD-95), Montreal, Canada, August 20–21, 1995*, pp. 216–221.

[B89] J. Rissanen (1995). Stochastic complexity and its applications. In *Workshop on Model Uncertainty and Model Robustness, June 30th to July 2nd 1995, Bath, England*. Electronic publication. Available from `http://www-old.stat.duke.edu/conferences/bath/rissanen.ps`.

[B90] Jorma Rissanen (1995). Stochastic complexity in learning. In *Computational Learning Theory, Second European Conference, EuroCOLT '95, Barcelona, Spain, March 1995, Proceedings* (Paul M. B. Vitányi, ed.), *Lecture Notes in Artificial Intelligence*, vol. 904, Springer, pp. 196–210. Invited lecture.

[B91] Marcelo J. Weinberger, Jorma Rissanen, Meir Feder (1995). A universal finite memory source. *IEEE Transactions on Information Theory*, 41 (3), 643–652.

[B92] Søren Forchhammer, Jorma Rissanen (1996). Partially hidden Markov models. *IEEE Transactions on Information Theory*, 42 (4), 1253–1256.

[B93] Manish Mehta, Rakesh Agrawal, Jorma Rissanen (1996). Sliq: a fast scalable classifier for data mining. In *Advances in Database Technology – EDBT'96, 5th International Conference on Extending Database Technology, Avignon, France, March 25–29, 1996, Proceedings* (Peter M. G. Apers, Mokrane Bouzeghoub, Georges Gardarin, eds.), *Lecture Notes in Computer Science*, vol. 1057, Springer, pp. 18–32.

[B94] J. Rissanen (1996). Information theory and neural nets. In *Mathematical Perspectives in Neural Networks* (P. Smolensky, M. C. Mozer, D. E. Rumelhart, eds.), Lawrence Erlbaum Associates, Mahwah, NJ, chapter 16, pp. 567–602.

[B95] J. Rissanen (1996). Stochastic complexity – an introduction. In *Computational Learning and Probabilistic Reasoning*, Wiley, Chichester, England, chapter 2, pp. 33–41.

[B96] J. Rissanen, Bin Yu (1996). Learning by MDL. In *Learning and Geometry: Computational Approaches* (David W. Kueker, Carl H. Smith, eds.), *Progress in Computer Science and Applied Logic*, vol. 14, Birkhäuser, Boston, MA, pp. 3–19.

[B97] Jorma J. Rissanen (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, 42 (1), 40–47.

[B98] Marcelo J. Weinberger, Jorma Rissanen, Ronald Arps (1996). Applications of universal context modeling to lossless compression of gray-scale images. *IEEE Transactions on Image Processing*, 5 (4), 575–586.

[B99] J. Rissanen (1997). Shannon-Wiener information and stochastic complexity. In *Proceedings of the Norbert Wiener Centenary Congress, 1994 (East Lansing, MI, November 27–December 3, 1994)* (V. Mandrekar, P. R. Masani, eds.), *Proceedings of Symposia in Applied Mathematics*, vol. 52, American Mathematical Society, Providence, RI, pp. 331–341.

[B100] J. Rissanen (1997). Stochastic complexity in learning. *Journal of Computer and System Sciences*, 55 (1, part 2), 89–95. Second Annual European Conference on Computational Learning Theory (EuroCOLT '95), Barcelona, 1995.

[B101] Ioan Tăbuş, Jorma Rissanen, Jaakko Astola (1997). Adaptive L-predictors based on finite state machine context selection. In *Proceedings of ICIP'97 International Conference on Image Processing, Santa Barbara, California, October 1997*, pp. 401–404.

[B102] Andrew Barron, Jorma Rissanen, Bin Yu (1998). The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, 44 (6), 2743–2760. Special issue: Information Theory: 50 Years of Discovery.

[B103] J. Rissanen, G. Shedler (1998). Failure-time prediction. *Journal of Statistical Planning and Inference*, 66 (2), 193–210.

[B104] I. Tăbuş, C. Popeea, J. Rissanen, Astola J. (1998). Alphabet extensions for Markov sources. In *Proceedings of ITW'98 Information Theory Workshop, San Diego, California, February 8–11, 1998*, p. 78.

[B105] J. Rissanen (1999). MDL denoising with wavelets. In *IEEE Information Theory Worskhop on Detection, Estimation, Classification and Imaging (DECI), February 24–26, 1999, Santa Fe, NM*. Electronic publication. Available from `http://www.ifp.uiuc.edu/itw-deci/psfiles/rissanen.ps`.

[B106] Jorma Rissanen (1999). Discussion of paper "Minimum message length and Kolmogorov complexity" by C. S. Wallace and D. L. Dowe. *The Computer Journal*, 42 (4), 327–329.

[B107] Jorma Rissanen (1999). Fast universal coding with context models. *IEEE Transactions on Information Theory*, 45 (4), 1065–1071.

[B108] Jorma Rissanen (1999). Hypothesis selection and testing by the MDL principle. *The Computer Journal*, 42 (4), 260–269. Invited paper for the special issue devoted to Kolmogorov complexity.

[B109] Jorma Rissanen (1999). Rejoinder. *The Computer Journal*, 42 (4), 343–344.

[B110] C. D. Giurcaneanu, I. Tabus, J. Rissanen (2000). MDL based digital signal segmentation. In *Proceedings of EUSIPCO-2000, X European Signal Processing Conference, Tampere, Finland, September 4–8*, vol. 1, pp. 339–342.

[B111] J. Rissanen (2000). A generalized minmax bound for universal coding. In *Proceedings of IEEE International Symposium on Information Theory (ISIT2000), Sorrento, Italy, June 25–30, 2000*, IEEE Press, p. 324.

[B112] J. Rissanen, B. Yu (2000). Coding and compression: a happy union of theory and practice. *Journal of the American Statistical Association*, 95, 986–988. Invited Year 2000 Commemorative Vignette on Engineering and Physical Sciences.

[B113] Jorma Rissanen (2000). MDL denoising. *IEEE Transactions on Information Theory*, 46 (7), 2537–2543.

[B114] Ioan Tabus, Gergely Korodi, Jorma Rissanen (2000). Text compression based on variable-to-fixed codes for Markov sources. In *2000 Data Compression Conference (DCC 2000), 28–30 March 2000, Snowbird, UT, USA*, pp. 133–142.

[B115] Rissanen Jorma (2001). Information, complexity and the MDL principle. In *Cycles, Growth and Structural Change: Theories and empirical evidence* (Lionello F. Punzo, ed.), Routledge Siena Studies in Political Economy, Routledge, New York, NY, pp. 339–350.

[B116] J. Rissanen (2001). Complexity and information in data. In *System Identification (SYSID 2000): A Proceedings volume from the 12th IFAC Symposium, Santa Barbara, California, USA, 21–23 June 2000* (R. Smith, ed.), vol. 1, Elsevier Science Ltd, pp. 1–6. Plenary paper.

[B117] Jorma Rissanen (2001). Simplicity and statistical inference. In *Simplicity, Inference and Modelling: Keeping It Sophisticatedly Simple* (Arnold Zellner, Hugo A. Keuzenkamp, Michael McAleer, eds.), Cambridge University Press, Cambridge, chapter 9, pp. 156–164.

[B118] Jorma Rissanen (2001). Strong optimality of the normalized ML models as universal codes and information in data. *IEEE Transactions on Information Theory*, 47 (5), 1712–1717.

[B119] J. Rissanen, B. Yu (2002). Coding and compression: a happy union of theory and practice. In *Statistics in the 21st Century* (Adrian E. Raftery, Martin A. Tanner, Martin T. Wells, eds.), *Monographs on Statistics and Applied Probability*, vol. 93, CRC Press, Boca Raton, FL, and American Statistical Association, Alexandria, VA, pp. 229–236.

[B120] I. Tabus, J. Rissanen, J. Astola (2002). A classifier based on normalized maximum likelihood model for classes of boolean regression models. In *Proceedings of EUSIPCO 2002, XI European Signal Processing Conference, September 3–6, 2002, Toulouse, France*, vol. 1, pp. 119–122.

[B121] I. Tabus, J. Rissanen, J. Astola (2002). Normalized maximum likelihood models for Boolean regression with application to prediction and classification in genomics. In *Computational And Statistical Approaches To Genomics* (Wei Zhang, Ilya Shmulevich, eds.), Kluwer Academic Publishers, pp. 173–196, 1st ed. Second Edition: Springer, 2006, pp. 235–258.

[B122] Ioan Tabus, Jorma Rissanen (2002). Asymptotics of greedy algorithms for variable-to-fixed length coding of Markov sources. *IEEE Transactions on Information Theory*, 48 (7), 2022–2035.

[B123] Petri Kontkanen, Wray Buntine, Petri Myllymäki, Jorma Rissanen, Henry Tirri (2003). Efficient computation of stochastic complexity. In *AI and statistics: Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics, January 3–6, 2003, Key West, Florida* (Christopher M. Bishop, Brendan J. Frey, eds.), Society for Artificial Intelligence and Statistics, pp. 181–188.

[B124] J. Rissanen (2003). Complexity and information in data. In *Entropy* (Andreas Greven, Gerhard Keller, Gerald Warnecke, eds.), Princeton Series in Applied Mathematics, Princeton University Press, Princeton, NJ, pp. 299–312.

[B125] Jorma Rissanen (2003). Complexity of simple nonlogarithmic loss functions. *IEEE Transactions on Information Theory*, 49 (2), 476–484.

[B126] B. Ryabko, J. Rissanen (2003). Fast adaptive arithmetic code for large alphabet sources with asymmetrical distributions. *IEEE Communications Letters*, 7 (1), 33–35.

[B127] Ioan Tabus, Gergely Korodi, Jorma Rissanen (2003). DNA sequence compression using the normalized maximum likelihood model for discrete regression. In *2003 Data Compression Conference (DCC 2003), 25–27 March 2003, Snowbird, UT, USA*, IEEE Computer Society, pp. 253–262.

[B128] Ioan Tabus, Jorma Rissanen, Jaakko Astola (2003). Classification and feature gene selection using the normalized maximum likelihood model for discrete regression. *Signal Processing*, 83 (4), 713–727. Special issue on genomic signal processing.

[B129] Janne Ojanen, Timo Miettinen, Jukka Heikkonen, Jorma Rissanen (2004). Robust denoising of electrophoresis and mass spectrometry signals with minimum description length principle. *Federation of European Biochemical Societies Letters*, 570 (1–3), 107–113. Invited paper.

[B130] J. Rissanen (2004). Complexity and information in modeling. In *Computability, Complexity and Constructivity in Economic Analysis* (K. Vela Velupillai, ed.), Blackwell Publishing, Oxford, chapter 4.

[B131] J. Rissanen, I. Tabus (2004). Modelling with distortion. In *CISS 2004, 38th Annual Conference on Information Sciences and Systems, Princeton University, March 17–19*, pp. 560–563.

[B132] Adriana Vasilache, Ioan Tăbuş, Jorma Rissanen (2004). Algorithms for constructing min-max partitions of the parameter space for MDL inference. In *Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International Workshops SSPR 2004 and SPR 2004, Lisbon, Portugal, August 2004, Proceedings* (Ana Fred, Terry Caelli, Robert P. W. Duin, Aurélio Campilho, Dick de Ridder, eds.), *Lecture Notes in Computer Science*, vol. 3138, Springer, pp. 930–938.

[B133] B. Barliga, I. Tabus, J. Rissanen, J. Astola (2005). Image denoising based on Kolmogorov structure function for a class of hierarchical image models. In *Proceedings of SPIE Optics and Photonics 2005: Algorithms, Architectures, and Devices and Mathematical Methods, Mathematical Methods in Pattern and Image Analysis, San Diego, California, USA, 31 July – 4 August 2005* (J. T. Astola, I. Tabus, J. Barrera, eds.), pp. 1–10.

[B134] Petri Kontkanen, Petri Myllymäki, Wray Buntine, Jorma Rissanen, Henry Tirri (2005). An MDL framework for data clustering. In *Advances in Minimum Description Length: Theory and Applications* (Peter D. Grünwald, In Jae Myung, Mark A. Pitt, eds.), Neural Information Processing, MIT Press, Cambridge, Massachusetts, chapter 13, pp. 323–353.

[B135] Gergely Korodi, Jorma Rissanen, Ioan Tăbuş (2005). Lossless data compression using optimal tree machines. In *2005 Data Compression Conference (DCC 2005), 29–31 March 2005, Snowbird, UT, USA*, IEEE Computer Society, pp. 348–357.

[B136] J. Rissanen, I. Tabus (2005). Kolmogorov's structure function in MDL theory and lossy data compression. In *Advances in Minimum Description Length: Theory and Applications* (Peter D. Grünwald, In Jae Myung, Mark A. Pitt, eds.), Neural Information Processing, MIT Press, Cambridge, Massachusetts, chapter 10, pp. 245–262.

[B137] Jorma Rissanen (2005). An inroduction to the MDL principle. In *Proceedings of the COBERA Workshop on Computable Economics, March 20–24, 2005, National University of Ireland, Galway, Ireland.*

[B138] Ciprian Doru Giurcăneanu, Jorma Rissanen (2006). Estimation of AR and ARMA models by stochastic complexity. In *Time Series and Related Topics. In Memory of Ching-Zong Wei* (Tze Leung Lai Hwai-Chung Ho, Ching-Kang Ing, ed.), *Lecture notes–monograph series*, vol. 52, Institute of Mathematical Statistics, Beachwood, Ohio, USA, pp. 48–59.

[B139] V. Kumar, J. Heikkonen, J. Rissanen, K. Kaski (2006). MDL denoising with histogram models. *IEEE Transactions on Signal Processing*, 54 (8), 2922–2928.

[B140] Jorma Rissanen (2006). The structure function and distinguishable models of data. *The Computer Journal*, 49 (6), 657–664.

[B141] I. Tabus, A. Hategan, C. Mircean, J. Rissanen, I. Shmulevich, W. Zhang, J. Astola (2006). Nonlinear modeling of protein expressions in protein arrays. *IEEE Transactions on Signal Processing*, 54 (6), 2394–2407.

[B142] Ioan Tabus, Jorma Rissanen, Jaakko Astola (2006). Nonlinear signal modeling and structure selection with applications to genomics. In *Advances in nonlinear signal and image processing* (Stephen Marshall, Giovanni L. Sicuranza, eds.), *EURASIP Book Series on Signal Processing and Communications*, vol. 6, Hindawi Publishing Corporation, New York, NY, chapter 4, pp. 79–102.

[B143] Jaakko Astola, Gergely Korodi, Jorma Rissanen, Ioan Tabus (2007). DNA sequence compression – based on the normalized maximum likelihood model. *IEEE Signal Processing Magazine*, 24 (1), 47–53.

[B144] Teemu Roos, Jorma Rissanen (2007). Conditional NML universal models. In *Information Theory and Applications Workshop (ITA-07), January 29 – February 2, 2007, University of California, San Diego.* Electronic publication. Available from `http://eprints.pascal-network.org/archive/00002972/01/CNMLITA.pdf`.

Patents

[C1] Glenn George Langdon Jr., Jorma J. Rissanen (1978), Method and means for arithmetic string coding. US Patent No. 4122440.

[C2] Glen G. Langdon Jr., Jorma J. Rissanen (1981), Method and means for arithmetic coding utilizing a reduced number of operations. US Patent No. 4286256.

[C3] Glen G. Langdon Jr., Jorma J. Rissanen (1984), High-speed arithmetic compression coding using concurrent value updating. US Patent No. 4467317.

[C4] Glen G. Langdon Jr., Jorma J. Rissanen (1984), Method and means for carry-over control in the high order to low order pairwise combining of digits of a decodable set of relatively shifted finite number strings. US Patent No. 4463342.

[C5] Glen G. Langdon Jr., Jorma J. Rissanen (1985), Adaptive source modeling for data file compression within bounded memory. US Patent No. 4494108.

[C6] Kottappuram M. A. Mohiuddin, Jorma J. Rissanen (1987), Multiplication-free multi-alphabet arithmetic code. US Patent No. 4652856.

[C7] Jorma J. Rissanen, Mati Wax (1988), Algorithm for constructing tree structured classifiers. US Patent No. 4719571.

[C8] Glen G. Langdon Jr., Joan L. Mitchell, William B. Pennebaker, Jorma J. Rissanen (1990), Arithmetic coding encoder and decoder system. US Patent No. 4905297.

[C9] Joan L. Mitchell, William B. Pennebaker, Jorma J. Rissanen (1991), Dynamic model selection during data compression. US Patent No. 5045852.

[C10] Paul D. Healey, Jorma J. Rissanen (1994), Adaptive computation of symbol probabilities in n-ary strings. US Patent No. 5357250.

[C11] Gilbert Furlan, Jorma J. Rissanen, Dafna Sheinvald (1996), Distributed coding and prediction by use of contexts. US Patent No. 5550540.

[C12] Gilbert Furlan, Jorma Johannes Rissanen (1997), Quantization method for image data compression employing context modeling algorithm. US Patent No. 5640159.

[C13] Gilbert Furlan, Jorma Johannes Rissanen, Dafna Sheinvald (1997), Distributed coding and prediction by use of contexts. US Patent No. 5652581.

[C14] Rakesh Agrawal, Manish Mehta, Jorma Johannes Rissanen (1998), Data mining method and system for generating a decision tree classifier for data records based on a minimum description length (MDL) and presorting of records. US Patent No. 5787274.

[C15] Jorma Rissanen (2006), Lossless data compression system. US Patent No. 7028042.

The 49 coauthors of Jorma Rissanen

Ioan Tăbuş	17	Z. Vágó	2
Jaakko Astola	9	J. H. van Schuppen	2
Glen G. Langdon Jr.	7	B. Barliga	1
Bin Yu	5	Andrew Barron	1
László Gerencsér	4	F. Wyman Bostwick	1
Gergely Korodi	4	A. Chang	1
Edward J. Hannan	3	A. Hategan	1
Thomas Kailath	3	K. Kaski	1
Marcelo J. Weinberger	3	V. Kumar	1
Rakesh Agrawal	2	N. Mc Clamroch	1
Ronald Arps	2	Timo Miettinen	1
L. Barbosa	2	C. Mircean	1
Wray Buntine	2	Janne Ojanen	1
Peter E. Caines	2	C. Popeea	1
Meir Feder	2	Teemu Roos	1
Søren Forchhammer	2	B. Ryabko	1
Ciprian Doru Giurcăneanu	2	G. Shedler	1
Jukka Heikkonen	2	I. Shmulevich	1
Petri Kontkanen	2	Terry Speed	1
Lennart Ljung	2	Stephen Todd	1
Manish Mehta	2	Adriana Vasilache	1
K. M. Mohiuddin	2	Mati Wax	1
Petri Myllymäki	2	V. Wertz	1
Eric Sven Ristad	2	W. Zhang	1
Henry Tirri	2		

THE MDL PRINCIPLE, PENALIZED LIKELIHOODS, AND STATISTICAL RISK

ANDREW R. BARRON, CONG HUANG, JONATHAN Q. LI, AND XI LUO

ABSTRACT. We determine, for both countable and uncountable collections of functions, information-theoretic conditions on a penalty pen(f) such that the optimizer $\hat{f}$ of the penalized log likelihood criterion $\log 1/\text{likelihood}(f) + \text{pen}(f)$ has statistical risk not more than the index of resolvability corresponding to the accuracy of the optimizer of the expected value of the criterion. If $\mathcal{F}$ is the linear span of a dictionary of functions, traditional description-length penalties are based on the number of non-zero terms of candidate fits (the ℓ_0 norm of the coefficients) as we review. We specialize our general conclusions to show the ℓ_1 norm of the coefficients times a suitable multiplier λ is also an information-theoretically valid penalty.

1. INTRODUCTION

From work in the information theory and statistics communities, there are close connections between high-quality data compression and accurate statistical estimation. The original Shannon (1948) code construction and the condition of Kraft characterizing valid codelengths show the correspondence between probability distributions $p(data)$ for data and optimal variable-length binary codes of length essentially $\log_2 1/p(data)$ bits (see, e.g., Cover and Thomas 2007). The development of universal data compression and, in particular, the minimum description-length (MDL) principle has built this correspondence further to deal with the case of distributions $p_f(data)$ that depend on an unknown function f believed to belong to a family $\mathcal{F}$ which may be given parametrically (see, Barron, Rissanen and Yu 1998 or Grünwald 2007 and work cited therein). The function f may provide a density or log-density function (for instance we may have $p_f(x) = p_0(x)e^{f(x)}/c_f$ where p_0 is a reference distribution and c_f is a normalizing constant), or, in the case that the data consists of pairs of inputs X and outputs Y, the function $f(x)$ may refer to a regression function, classification function, Poisson intensity function, etc. that captures an essential aspect of the conditional distribution of Y given X. Starting from a discussion of coding redundancy, we analyze statistical risk of estimation, capturing its relationship to the accuracy of approximation and the level of complexity of functions f in $\mathcal{F}$, to contribute to a general theory of penalized likelihood.

Ideal procedures adapt to the complexity revealed by the data. We discuss results for mixture-based and prediction-based procedures and present new results for procedures that optimize penalized likelihood. Penalties $pen(f)$ are typically related to parameter dimension or to function irregularity. We develop means to determine when such penalties capture information-theoretic complexity to provide for quality compression and accurate function estimation.

An index of resolvability, the optimum sum of relative entropy approximation error and penalty relative to the sample size, is used to capture the performance of these procedures. It upper bounds the

Andrew Barron, Cong Huang, and Xi Rossi Luo are with the Department of Statistics, Yale University, P.O. Box 208290, New Haven, CT 06520-8290; Andrew.Barron@yale.edu, Cong.Huang@yale.edu and Xi.Luo@yale.edu; Jonathan Qiang Li is with Radar Networks, Inc., 410 Townsend St., San Francisco, CA 94107: qiang.li@aya.yale.edu.

statistical risk as does a related expression involving an expected redundancy of data compression. These resolvability and redundancy bounds on risk have been developed for penalized likelihood restricted to a countable set of functions which discretizes $\mathcal{F}$, with complexity penalty $pen(f) = L(f)$ equal to an information-theoretic codelength for f (Barron and Cover 1991, Barron 1990, Li 1999, Kolaczyk and Nowak 2004, 2005, and Grünwald 2007). The estimator is interpretable as a maximizing posterior probability with $L(f)$ equal to the log reciprocal prior probability of f. Some results for classes formed from continuous finite-dimensional families with penalty proportional to the dimension have been developed (Yang and Bafrron 1998, Barron, Birgé and Massart 1999), giving resolvability bounds on risk of such penalized likelihood estimators. Moreover, resolvability bounds on risk of Bayes predictive density estimators with general priors have been developed as will be discussed below. The present paper gives a simple and natural method to extend the previous information-theoretic bounds for penalized likelihood to deal with more general penalties not restricted to countable $\mathcal{F}$ or to penalties based on dimension.

Early advocates of penalized likelihood estimation with penalty on the roughness of the density include Good and Gaskins (1971,1980), de Montricher, Tapia and Thompson (1975), and Silverman (1982). Reproducing kernel Hilbert space penalties are championed in Wahba (1990). Statistical rate results for quadratic penalties in Hilbert space settings corresponding to weighted ℓ_2 norms on coefficients in function expansions (including Sobolev-type penalties equal to squared L_2 norms of derivatives) are developed in Cox and O'Sullivan (1990) based on functional analysis tools. Later developments in this direction are in Cucker and Smale (2001). Empirical process techniques built around metric entropy calculations yield rate results for penalties designed for a wide variety of function classes in Shen (1998). Related theory for constrained maximum likelihood in nonparametric settings is in Nemirovski, Polyak and Tysbakov (1985) and for minimum contrast estimators and sieves in Birgé and Massart (1993,1998).

The use of ℓ_1 penalization of log-likelihoods is a currently popular approach, see Park and Hastie (2007). The penalty is applied to coefficients in linear models for f, coinciding with a generalized linear model $p_f(\underline{u})$ for the data, where the terms of the linear model are members of a dictionary of candidates. For special cases, see Koh, Kim and Boyd (2007), Banerjee, Ghaoui and d'Aspermont (2007), Friedman, Hastie and Tibshirani (2007b), or Zhang, Wahba et al (2005). That work has focussed on algorithmic development, related to work for penalized least squares in Tibshirani's 1996 Lasso, Chen and Donoho's 1994,1999 basis pursuit, the LARS algorithm (Efron et al 2004), coordinate algorithms (Friedman et al 2007a) and relaxed greedy algorithms (Jones 1992, Barron 1993, Lee, Bartlett and Williamson 1996, Barron and Cheang 2001, Zhang 2003, and Barron, Cohen, et al 2008). A new algorithmic result is established at the end of this paper.

Recently there is activity to analyze risk of ℓ_1 penalized procedures. Much of it, requiring restrictions on the correlation of dictionary members, focusses on whether the procedure performs as well as the best subset selection rule, as in the work on ℓ_1 penalized least squares regression in Bunea, Tsybakov and Wegkamp (2006,2007a) and Zhang (2007), on ℓ_1 penalized empirical L_2 criteria for density estimation in Bunea, Tsybakov and Wegkamp (2007b), and ℓ_1 penalized logistic regression in Meier, van de Geer and Bühlmann (2008), and the general method of van de Geer (2008). For general dictionaries without correlation conditions, it is natural to ask whether an ℓ_1 penalized criterion performs as well as the best tradeoff between approximation error and ℓ_1 norm of coefficients. This is examined for ℓ_1 penalized

least squares by Huang, Cheang and Barron (2008) and for ℓ_1 penalized likelihood in the present paper. Risk bounds for penalized likelihood should capture the corresponding tradeoff of Kullback-Leibler approximation error and the penalty, as is available for Bayes predictive estimators. This motivates our analysis of the risk of penalized likelihood estimators and demonstration that the ℓ_1 penalty satisfies the information-theoretic requirements for the results we seek.

Extending information-theoretic risk results to penalized likelihood with an uncountable family $\mathcal{F}$, the main tool developed in Section 3 is that of a variable-complexity cover. Such covers allow for variable penalty levels. The distortion used in measuring closeness to the cover is based on discrepancies between log-likelihood and its theoretical analog rather than based on the metrics of traditional metric entropy. In brief, a valid penalty $pen(f)$ is one for which for each f in $\mathcal{F}$ there is a representor in the cover for which $pen(f)$ is not less than its complexity plus distortion.

The theory is simplified compared to alternatives that would glue together bounds for subclasses with their separate metric entropy (fixed complexity) covering properties. Indeed, it is not necessary to organize $\mathcal{F}$ to come from a list of function subclasses. Nevertheless, to relate to past work, various subclasses $\mathcal{F}_s$ may arise, corresponding to functions of various regularity s, quantified by number of derivatives or by weighted norms of coefficients in function expansions.

Often $\mathcal{F}$ is arranged as a union of families $\mathcal{F}_m$ of functions of similar characteristics, e.g., parametric families $\mathcal{F}_m = \{f_{\theta,m} : \theta \in R^{d_m}\}$ of given parameter dimension d_m. For instance, consider linear combinations of a dictionary $\mathcal{H}$ of functions. Such $f_\theta(x) = \sum_{h\in\mathcal{H}} \theta_h h(x)$ are specified by the coefficients $\theta = (\theta_h : h \in \mathcal{H})$. The set of linear combinations $\mathcal{F}$ is the union of models $\mathcal{F}_m$ for subsets m of $\mathcal{H}$ in which the $f_{\theta,m}(x) = \sum_{h\in m} \theta_h h(x)$. These families have dimension $d_m = \text{card}(m)$ when the functions in m are linearly independent.

The data are assumed to come from a sample space over which distributions indexed by f are provided. For our most general statements, other than a measure space, no particular structure need be assumed for this space. It is traditional to think of data in the form of a finite length string $\underline{U} = \underline{U}_n = (U_1, U_2, \ldots, U_n)$, consisting of a sequence of outcomes $X_1, X_2, \ldots, X_n$ or outcome pairs $(X_i, Y_i)_{i=1}^n$. We write $\underline{\mathcal{U}}$ for the sample space and $P_{\underline{U}|f}$ (or sometimes more briefly P_f if clear from the context) for the distributions on $\underline{U}$. Likewise $E_{\underline{U}|f}$ or sometimes more briefly E_f denotes the expected value. When being explicit about sample size, we index by n, as in $P_{\underline{U}_n|f}$ or $P_f^{(n)}$.

For lossless data compression, the space $\underline{\mathcal{U}}$ is countable, such as a discretization of an underlying continuous space, $p_f(\underline{u})$ is the probability mass function, and $q(\underline{u})$, satisfying Kraft's inequality $\sum_{\underline{u}\in\underline{\mathcal{U}}} q(\underline{u}) \leq 1$, is a coding distribution with codelengths $\log_2 1/q(\underline{u})$ in bits. Then the pointwise coding redundancy is $\log 1/q(\underline{u}) - \log 1/p_f(\underline{u})$, the difference between the actual codelength and the codelength we would have had if f were given. Following past MDL work, we allow continuous sample spaces and density functions relative to a given reference measure, yet, we refer to the log density ratio as a redundancy. See Barron (1985) for a limiting code redundancy interpretation of the absolutely continuous case involving fine discretizations.

Thus our setting is that the distributions $P_{\underline{U}|f}$ have density functions $p(\underline{u}|f) = p_f(\underline{u})$ relative to a fixed reference measure on $\underline{\mathcal{U}}$. The likelihood function $likelihood(f)$ is $p_f(\underline{U})$ at specified data $\underline{U}$. When the

sample space is a sequence space the reference measure is assumed to be a product of measures on the individual spaces. For the special case of i.i.d. modeling, there is a space $\mathcal{U}$ for the individual outcomes with distributions $P_f^{(1)} = P_f$ and then $\underline{\mathcal{U}}$ is taken to be the product space $\mathcal{U}^n$ and $P_{\underline{U}_n|f} = P_f^n$ is taken to be the product measure with joint density $p_f(\underline{u}_n) = \prod_{i=1}^n p_f(u_i)$.

The object of universal data compression and universal modeling in general is the choice of a single distribution $q(\underline{u})$, $\underline{u} \in \underline{\mathcal{U}}$, such that the redundancy $\log 1/q(\underline{u}) - \log 1/p_f(\underline{u})$ is kept not larger than need be (measured either pointwise or in expectation over $\underline{u}$ and either on the average or in worst case over f) for functions in each class of interest.

As discussed in Rissanen (1989), Barron, Rissanen and Yu (1998) and Grünwald (2007), minimum description-length methods choose q in one of several interrelated ways: by Bayes *mixtures*, by *predictive models*, by *two-stage* codes, or by *normalized maximum-likelihood* codes. We discuss some aspects of these with an eye toward redundancy and resolvability bounds on risk.

Our treatment of penalized likelihood gives general information-theoretic penalty formulation in sections 2 and 3, with risk bounds given for squared Hellinger and related distances, and then application to ℓ_1 penalties in sections 4 and 5. To put these results into an information-theoretic context, we first review below redundancy and resolvability bounds for mixture models and their implications for the risk of predictive estimators. These risk bounds are for the stronger Kullback-Leiber loss. This material shows that tools are already in place for dealing with uncountable families by mixture models, and their associated predictive interpretations. Then penalized likelihood is studied because of its familiarity and comparative ease of computation.

1.1. **Mixture models.** These models for $\underline{U}$ use a prior distribution w on $\mathcal{F}$ leading to a mixture density $q(\underline{u}) = q_w(\underline{u}) = \int p_f(\underline{u}) w(df)$. For instance with $\mathcal{F}$ a union of families $\mathcal{F}_m$, the prior may be built from a probability $w(m)$ and a distribution on $\mathcal{F}_m$ for each m. If $\mathcal{F}_m$ is given parametrically the prior may originate on the parameters yielding $q(\underline{u}|m) = q_{w_m}(\underline{u}) = \int p_{f_{\theta,m}}(\underline{u}) w(d\theta|m)$, and an overall mixture $q(\underline{u}) = \sum_m w(m) q(\underline{u}|m)$. A mixture distribution has average case optimal redundancy, averaging over $\underline{u}$ according to $p_f(\underline{u})$ and averaging over functions according to the prior. These mixture densities are the same objects used in Bayesian model selection and Bayesian prediction. However, a difference is that with MDL we use data compression thinking to guide the choice of the prior weights to achieve operationally desirable properties.

We discuss tools for redundancy and resolvability bounds for mixtures and the bounds they yield on risk. First we recall results for parametric families in which the aim is to uniformly control the redundancy.

The expected redundancy of the mixture q_{w_m} takes the form $E_f[\log p(\underline{U}|f)/q_{w_m}(\underline{U})]$ which we recognize as the Kullback-Leibler divergence between the mixture and the target. In a well-studied problem, initiated in the characterization of communication channel capacity and extended to minimax redundancy of universal data compression (Gallager 1968,1974, Davisson 1973, Davisson and Leon-Garcia 1980, Haussler 1997, Clarke and Barron 1990,1994 and Xie and Barron 1997) the minimax procedure yielding the smallest worst case expected redundancy in each $\mathcal{F}_m$ corresponds to a choice of prior w_m

yielding the largest minimum average redundancy, interpretable as a maximum Shannon mutual information $I(f;\underline{U})$, and suitable approximate forms for the optimal w_m, called the least favorable prior or capacity achieving prior, are available in an asymptotic setting. Indeed, for smooth parametric families with a Fisher information $I(\theta|m)$, an asymptotically optimal prior is proportional to $|I(\theta|m)|^{1/2}$ with a sequence of boundary modifications, and the resulting redundancy behaves asymptotically like $\frac{d_m}{2}\log n$ plus a specified constant determined by the logarithm of the integral of this root Fisher information. There are finite sample bounds of the same form but with slightly larger constants, from examination of the resolvability of mixtures we come to shortly.

Building on the work of Shtarkov (1987), the theory of pointwise minimax redundancy identifies what is the smallest constant penalty that can be added to $\log 1/p(\underline{U}|\hat{f}_m)$, where $\hat{f}_m = f_{\hat{\theta},m}$ is the maximizer of the likelihood, such that the result retains a data-compression interpretation. This problem has been studied in an asymptotic setting in Rissanen (1996), Barron, Rissanen and Yu (1998), Takeuchi et al (1997a,1997b,1998,2007), and Xie and Barron (2000). One of the conclusions, in the cases studied there, is that the same value $\frac{d_m}{2}\log\frac{n}{2\pi} + \log\int |I(\theta|m)|^{1/2}d\theta$ characterizes this smallest constant penalty asymptotically. That theory provides data compression justification for a penalty with main term proportional to the dimension d_m. Certain mixture procedures have asymptotically minimax pointwise redundancy and are shown to be close to the exact optimal normalized maximum likelihood. These mixtures use the same Fisher information based prior with boundary modification, with an additional modification required for non-exponential family cases, that puts some small mass on an enlargement of the family. That there are solutions of mixture form is of interest for our subsequent discussion of predictive distributions.

Choosing weights $w(m)$ to assign to the families can also be addressed from an information-theoretic standpoint, thinking of $\log 1/w(m)$ as a codelength. Indeed, since the MDL parameter cost, approximately $\frac{d_m}{2}\log n$, is determined by the dimension d_m, it is customary to set $\log 1/w(m)$ using the log-cardinality of models of the same dimension (one can not do much better than that for most such models). For example, for models which correspond to subsets m of size d chosen out of p candidate terms in a dictionary, the $\log 1/w(m)$ can be set to be $\log\binom{p}{d}$, plus a comparatively small additional description length for the dimension d. Often p is large compared to the sample size n, while the critical dimensions d which lead to the best resolvability are small compared to n, so this $\log 1/w(m)$ of order $d_m \log p/d_m$ substantially adds to $\frac{d_m}{2}\log n$ in the total description length. Use of $\frac{d_m}{2}\log n$ alone is not in accord with the total minimum description length principle in such cases in which the contribution from $\log 1/w(m)$ is comparable or larger.

1.2. **Index of resolvability of mixtures.** We now come to a bound on expected redundancy of mixtures developed in Barron (1998), which is shown to bound an associated statistical risk. Recall the Kullback divergence $D(P_{\underline{U}}||Q_{\underline{U}}) = E\log p(\underline{U})/q(\underline{U})$ is the total expected redundancy if data $\underline{U}$ are described using $q(\underline{u})$ but the governing measure has a density $p(\underline{u})$. Suppose this density has the form $p_{f^*}(\underline{u})$, sometimes abbreviated $p_*(\underline{u})$. A tool in the examination of the redundancy is $D_n(f^*,f) = D(P_{\underline{U}|f^*}||P_{\underline{U}|f})$ which measures how well f approximates a hypothetical f^*. In the i.i.d. modeling case this divergence takes the form $D_n(f^*,f) = nD(f^*,f)$ where $D(f^*,f)$ is the divergence between the single observation distributions $D(P_{f^*}||P_f)$. It is an important characteristic of mixtures that the divergence of a mixture

from a product measure is considerably smaller than the order n divergence between pairs of distributions in the family.

Indeed, the resolvability bound on expected redundancy of mixtures is given as follows. Let the distribution $Q_{\underline{U}}$ be a general mixture with density $q(\underline{u}) = \int p(\underline{u}|f)W(df)$ formed from a prior W. Let B be any measurable subset of functions in $\mathcal{F}$. Then, as in Barron (1998), by restriction of the integral to B followed by Jensen's inequality, the redundancy of the mixture Q is bounded by the sum of the maximum divergence of distributions in B from the target f^* and the log reciprocal prior probability of B, and thus, minimizing over any collection of such subsets B,

$$D(P_{\underline{U}|f^*}||Q_{\underline{U}}) \;\leq\; \min_B \left\{ \max_{f\in B} D_n(f^*, f) + \log 1/W(B) \right\}.$$

In i.i.d. modeling, we divide by n to obtain the following redundancy rate bound. This shows the redundancy of mixture codes controlled by an *index of resolvability*, expressing the tradeoff between the accuracy of approximating sets and their log prior probability relative to the sample size,

$$(1/n)D(P_{\underline{U}_n|f^*}||Q_{\underline{U}_n}) \;\leq\; \min_B \left\{ \max_{f\in B} D(f^*, f) + \frac{1}{n}\log 1/W(B) \right\}.$$

When the $B = \{f\}$ are singleton sets, the right side is the same as the index of resolvability given in Barron and Cover (1991), used there for two-stage codes, as will be discussed further. The optimal sets for the resolvability bound for mixture codes take the form of Kullback balls $B_{r,f^*} = \{f : D(f^*, f) \leq r^2\}$, yielding

$$(1/n)D(P_{\underline{U}_n|f^*}||Q_{\underline{U}_n}) \;\leq\; \min_{r\geq 0} \left\{ r^2 + \frac{1}{n}\log 1/W(B_{r,f^*}) \right\}.$$

As illustrated in Barron (1998) with suitable choices of prior, it provides the usual $(d_m/2)(\log n)/n$ behavior of redundancy rate in finite-dimensional families, and rates of the form $(1/n)^\rho$ for positive $\rho < 1$ for various infinite-dimensional families of functions. Similar characterization arises from a stronger Bayes resolvability bound $D(P_{\underline{U}|f^*}||Q_{\underline{U}}) \leq -\log \int e^{-D_n(f^*,f)}W(df)$ as developed in Barron (1988,1998), Haussler and Barron (1993), and Zhang (2006).

1.3. **Implications for predictive risk.** For predictive models the data are presumed to arise in a sequence $\underline{U}_N = (U_n)_{n=1}^N$ and the joint distribution $q(\underline{U}_N)$ (for universal modeling or coding) is formed by gluing together predictive distributions $q(u_n|\underline{u}_{n-1})$, that is, by multiplying together these conditional densities for $n = 1, 2, \ldots, N$. In the i.i.d. modeling case, given f, the density for U_n given the past is $p(u_n|f)$. Predictive distributions are often created in the form $p(u_n|\hat{f}_{n-1})$ by plugging in an estimate $\hat{f}_{n-1}$ based on the past $\underline{u}_{n-1} = (u_i)_{i=1}^{n-1}$. Nevertheless, predictive distribution need not be restricted to be of such a plug-in form. Indeed, averaging with respect to a prior w, a one-step-ahead predictive redundancy is optimized by a Bayes predictive density $q(u_n|\underline{u}_{n-1})$. The one-step-ahead predictive redundancy is $E_f D(P_{U_n|f}||Q_{U_n|\underline{U}_{n-1}})$, which we recognize to be the Kullback risk of the predictive density, based on a sample of size $n-1$, as an estimate of the target density $p(u_n|f)$. Here and in what follows, it is to be understood that if the variables are not i.i.d. given f, the target becomes the conditional density $p(u_n|\underline{u}_{n-1}, f)$. The model built by multiplying the Bayes predictive densities together is the mixture $q_w(\underline{u})$. Correspondingly, by the chain rule, the total codelength and its redundancy yield the same values, respectively, as the mixture codelength and redundancy discussed in (1) above. Indeed, the total

redundancy of the predictive model is

$$D(P_{\underline{U}_N|f}||Q_{\underline{U}_N}) \;=\; \sum_{n=1}^{N} E_f D(P_{U_n|f}||Q_{U_n|\underline{U}_{n-1}}),$$

which is the cumulative Kullback risk. Dividing by N we see in particular that the Cesàro average of the risks of the predictive distributions is bounded by the index of resolvability discussed above.

This chain rule property has been put to use for related conclusions. For example, it is the basis of the analysis of negligibility of superefficiency in Barron and Hengartner (1998). That work shows for d-dimensional families that $\frac{d}{2n}$ is the asymptotically efficient level of individual Kullback risk based on samples of size n. Indeed, summing across sample sizes $n = 1, 2, \ldots, N$, it corresponds to a total Kullback risk (total redundancy) of $\frac{d}{2} \log N$, which cannot be improved upon asymptotically (except in a negligible set of parameters) according to Rissanen's (1984) award winning result. The predictive interpretation also plays a critical role for non-finite dimensional families $\mathcal{F}_s$ in identifying the efficient rates of estimation (also in Barron and Hengartner 1998) and in establishing the minimax rates of estimation (in Yang and Barron 1999 and Haussler and Opper 1997). For these cases typical individual risk rates are of the form some constant times $(1/n)^\rho$ for some positive rate $\rho \leq 1$. At the heart of that analysis, one observes that taking the Cesàro average Kullback risk across sample sizes up to N recovers the same form $(1/N)^\rho$ (albeit with a different constant multiplier). The idea is that minimax rates for total expected redundancy is somewhat easier to directly analyze than individual Kullback risk, though they are related by the chain rule given above.

1.4. **Two-stage codes.** We turn our attention to models based on $two\text{-}stage$ codes, also called two-part codes. We recall some previous results here, and give in the next sections some simple generalizations to penalized likelihoods. Two-stage codes were used in the original formulation of the MDL principle by Rissanen (1978,1983) and in the analysis of Barron and Cover (1991). One works with a countable set $\tilde{\mathcal{F}}$ of possible functions, perhaps obtained by discretization of the underlying family $\mathcal{F}$. A key ingredient in building the total two-stage description length are assignments of complexities $L_n(f)$, for $f \in \mathcal{F}$, satisfying the Kraft inequality $\sum_{f\in\mathcal{F}} 2^{-L_n(f)} \leq 1$, given the size n of the sample.

These complexities typically have the form of a codelength for the model class m (of the form $L(m) = \log 1/w(m)$ as discussed above), plus a codelength $L(f|m)$ or $L(\theta|m)$ for the parameters that determine the functions in $\mathcal{F}_m$, which may be discretized to a grid of precision δ for each coordinate, each of which is described using about $\log 1/\delta$ bits. Under, respectively, first or second order smoothness conditions on how the likelihood depends on the parameters, the codelength for the parameters comes out best if the precision δ is of order $\frac{1}{n}$ or $\frac{1}{\sqrt{n}}$, leading to $L(f|m)$ of approximately $d_m \log n$ or $\frac{d_m}{2} \log n$, respectively, for functions in smooth families $\mathcal{F}_m$.

We are not forced to always have such growing parameter complexities. Indeed, as suggested by Cover and developed in Barron (1985) and Barron and Cover (1991), one may consider a more general notion of parameter complexity inspired by Kolmogorov. That work shows when any computable parameter value govern the data, ultimately a shorter total codelength obtains with it than for all other competitors and the true parameter value is discovered with probability one. Nevertheless, for any coding scheme, in second order smooth families with parameters in R^{d_m}, except for a null set of Lebesgue

measure 0 as shown by Rissanen (1984,1986), the redundancy will not be of smaller order than $\frac{d_m}{2}\log n$. The implication for parameter coding is that for most parameters the representor in the code will need to have complexity of order not smaller than $\frac{d_m}{2}\log n$.

For each function f and data $\underline{U}$, one has a two-stage codelength $L_n(f) + \log 1/p_f(\underline{U})$ corresponding to the bits of description of f followed by the bits of the Shannon code for $\underline{U}$ given f. Then the minimum total two-stage codelength takes the form

$$\min_{f\in\tilde{\mathcal{F}}} \left\{ \log \frac{1}{p_f(\underline{U})} + L_n(f) \right\}.$$

The minimizer $\hat{f}$ (breaking ties by choosing one of minimal $L_n(f)$) is called the minimum complexity estimator in the density estimation setting of Barron and Cover (1991) and it is called the complexity regularization estimator for regression and classification problems in Barron (1990).

Typical behavior of the minimal two stage codelength is revealed by investigating what happens when the data $\underline{U}_n$ are distributed according to $p_{f^*}(\underline{u}_n)$ for various possible f^*. As we have noted, eventually exact discovery is possible when f^* is in $\tilde{F}$, but its complexity, as will be ultimately revealed by the data, may be too great for full specification of f^* to be the suitable description with moderate sample sizes. It is helpful to have the notion of a surrogate function f_n^* in the list $\tilde{F}$, appropriate to the current sample size n, in place of f^* which is not necessarily in the countable $\tilde{F}$. The appropriateness of such an f_n^* is judged by whether it captures expected compression and estimation properties of the target.

The redundancy rate of the two-stage description (defined as $\frac{1}{n}$ times the expected difference between the total codelength and the target $\log 1/p_{f^*}(\underline{U}_n)$) is shown in Barron and Cover (1991) to be not more than the index of resolvability defined by

$$R_n(f^*) \;=\; \min_{f\in\tilde{\mathcal{F}}} \left\{ \frac{1}{n} D(P_{\underline{U}_n|f^*} || P_{\underline{U}_n|f}) + \frac{1}{n} L_n(f) \right\}.$$

For i.i.d. modeling it takes the form

$$R_n(f^*) \;=\; \min_{f\in\tilde{\mathcal{F}}} \left\{ D(f^*, f) + L_n(f)/n \right\},$$

capturing the ideal tradeoff in error of approximation of f^* and the complexity relative to the sample size. The function f_n^* which achieves this minimum is the population counterpart to the sample-based $\hat{f}$. It best resolves the target for the given sample size. Since $\hat{f}$ is the sample-based minimizer, one has an inequality between the pointwise redundancy and a pointwise version of the resolvability

$$\log \frac{p_{f^*}(\underline{U})}{p_{\hat{f}}(\underline{U})} + L_n(\hat{f}) \;\;\leq\;\; \log \frac{p_{f^*}(\underline{U})}{p_{f_n^*}(\underline{U})} + L_n(f_n^*).$$

The resolvability bound on the expected redundancy is recognized as the result of taking the expectation of this pointwise inequality.

This $R_n(f^*)$ also bounds the statistical risk of $\hat{f}$, as we recall and develop further in Section 2, with a simplified proof and with extension in Section 3 to uncountable $\mathcal{F}$. The heart of our statistical analysis will be the demonstration that the loss function we examine is smaller in expectation and stochastically not much more than the pointwise redundancy.

Returning to the form of the estimator, we note that when $\tilde{\mathcal{F}}$ is a union of sets $\tilde{\mathcal{F}}_m$, the complexities may take the form $L(m, f) = L(m) + L(f|m)$ for the description of m followed by the description of f given m. Thus, in fact, though it is customary to refer to two-stage coding, there are actually three-stages with minimum total

$$\min_m \min_{f \in \tilde{\mathcal{F}}_m} \{L(m) + L(f|m) + \log 1/p_f(\underline{U})\}.$$

The associated minimizer $\hat{m}$ (again breaking ties by choosing the simplest such m) provides a model selection in accordance with the MDL principle. Likewise the resolvability takes the form

$$R_n(f^*) = \min_m \min_{f \in \tilde{\mathcal{F}}_m} \{D(f^*, f) + L(m, f)/n\}.$$

The ideal model selection, best resolving the target, is the choice m_n^* achieving the minimum, and the performance of the sample based MDL selection $\hat{m}$ is captured by the resolvability provided by m_n^*.

Two-stage codes in parametric families are closely related to average-case optimal mixture-codes. Indeed, in second order smooth families of dimension d, Laplace approximation, as in Barron (1985), shows that log mixture likelihood is approximately the maximum log-likelihood minus the log ratio between the square root of the determinant of empirical total Fisher information and the prior density, plus $\frac{d}{2} \log 2\pi$. Two-stage codes can achieve the same form (although with a slightly suboptimal constant) provided one uses more elaborate parameter quantizations based on local diagonalization of the Fisher information with a rectangular grid in the locally transformed parameter space, as explained in Barron (1985), rather than merely using a rectangular grid in the original parameters. To avoid such complications and to have exact average-case optimality, when it is computationally feasible, it is preferable to use mixture models in such smooth families rather than two-stage codes.

Nevertheless, in many estimation settings, it is common to proceed by a penalized likelihood (or penalized squared error) criterion, and it is the intent of the present paper to address associated information-theoretic and statistical properties of such procedures.

To recap, we have seen in the minimum description-length principle that there are close connections between compression and statistical estimation.

The connections of information theory and statistics have additional foundations. While it is well-known that information-theoretic quantities determine fundamental limits of what is possible in communications, it is also true that corresponding information-theoretic quantities determine fundamental limits of what is possible in statistical estimation, as we recall in the next subsection.

1.5. **Information-theoretic determination of minimax rates.** In Haussler and Opper (1997) and Yang and Barron (1999) the problem of minimax rates of function estimation are shown to have an information-theoretic characterization. Suppose we have a loss function $\ell(f^*, f)$ which is a squared metric locally equivalent to Kullback divergence $D(f^*, f)$ (i.e., they agree to within a constant factor in a suitable subset of the function space), and assume that the data $\underline{U}_n = (U_1, \ldots, U_n)$ are i.i.d. from p_{f^*} with an f^* in a given function class $\mathcal{F}_s$. Here we use the subscript s to remind ourselves that we are referring to function subclasses that permit control on the quantities of interest that characterize minimax rates (that is, finite metric entropy or finite capacity). In the language of information theory the family of distributions $(P_{\underline{U}_n|f}, f \in \mathcal{F}_s)$ is a channel with inputs f and outputs $\underline{U}_n$.

Three quantities are shown to be important in the study of the statistical procedures: these are the Kolmogorov *metric entropy*, the Shannon *channel capacity*, and the *minimax risk* of Wald's statistical decision theory. The *metric entropy* $H_\epsilon = H_\epsilon(\mathcal{F}_s)$ is defined by

$$H_\epsilon(\mathcal{F}_s) \;=\; \inf_{\tilde{\mathcal{F}}} \left\{ \log \text{card}(\tilde{\mathcal{F}}) \,:\, \inf_{\tilde{f}\in\tilde{\mathcal{F}}} \ell(f,\tilde{f}) \le \epsilon^2, \quad \forall f \in \mathcal{F}_s \right\},$$

for which a critical $\epsilon_n = \epsilon_n(\mathcal{F}_s)$ is one for which ϵ_n^2 is of the same order as H_{ϵ_n}/n. Shannon's *channel capacity* $C_n = C_n(\mathcal{F}_s)$ is

$$C_n \;=\; \max_W I(f;\underline{U}_n)/n$$

where the maximum is over distributions W restricted to $\mathcal{F}_s$ and $I(f;\underline{U})$ is Shannon's mutual information for the channel, equal also to the Bayes average (w.r.t. W) of the redundancy of the mixture code. Finally, the *minimax risk* $r_n = r_n(\mathcal{F}_s)$ is

$$r_n \;=\; \inf_{\hat{f}} \sup_{f\in\mathcal{F}_s} E_f\, \ell(f,\hat{f}),$$

where the infimum is over all estimators based on the sample of size n. Suppose also that we are not in a finite-dimensional setting (where the metric entropy is order of a multiple of $\log 1/\epsilon$), but rather we are in an infinite-dimensional setting, where the metric entropy is of order at least $(1/\epsilon)^\gamma$ for some positive γ. Then from Haussler and Opper (1997) and Yang and Barron (1999), as also presented by Lafferty (2007), we have the equivalence of these quantities.

Theorem 1.1. *The minimax estimation rate equals the channel capacity rate equals the metric entropy rate at the critical precision. That is,*

$$r_n \sim C_n \sim \frac{H_{\epsilon_n}}{n} \sim \epsilon_n^2,$$

where $\sim$ means that the two sides agree to within constant factors.

In modern statistical practice it is rarely the case that one designs an estimator solely around one function class of bounded metric entropy. Indeed, even if one knew, in advance of seeing the data, how one wants to characterize regularity of the function (e.g. through a certain norm on coefficients), one usually does not have advance knowledge of an appropriate size of that norm, though such knowledge would be required for such metric entropy control. Instead, one creates an estimate that adapts, that is, it simultaneously gives the right levels of risk for various function subclasses. Penalized likelihood estimators provide a means by which to achieve such aims.

We say that the countable set $\tilde{\mathcal{F}}$ together with its variable complexities provides an *adaptive cover* of each of several function subclasses $\mathcal{F}_s$, if for each of the subclasses there is a subset of functions in $\tilde{\mathcal{F}}$ that have complexity bounded by a multiple of H_{ϵ_n} and that cover the subclass to within precision ϵ_n. Then application of the index of resolvability shows that the minimum complexity estimator is simultaneously minimax rate optimal for each such subclass. In this setting the role of Theorem 1.1 is to give the lower bounds showing that the achieved rates are indeed best possible.

As discussed in Yang and Barron (1998,1999), Barron, Birgé and Massart (1999), and Barron, Cohen, Dahmen and DeVore (2008), such adaptation (sometimes to within log-factors of the right rates) is shown

to come for free for a variety of function classes when the models consist of subsets of basis functions from suitable dictionaries and the penalties are given by the dimension (times a log-factor).

The resolvability bounds go beyond such asymptotic rate considerations to give finite sample performance characterization specific to properties of the target f^*, not required to be tied to the worst case for functions in various classes.

2. Risk and Resolvability for Countable $\tilde{\mathcal{F}}$

Here we recall risk bounds for penalized likelihood with a countable $\tilde{\mathcal{F}}$. Henceforth we use base e exponentials and logarithms to simplify the mathematics (the units for coding interpretations become nats rather than bits).

In setting our loss function, we will have need of another measure of divergence. Analogous to the Kullback-Leibler divergence we have already discussed, for pairs of probability distributions P and $\tilde{P}$ on a measurable space, we consider the Bhattacharyya, Hellinger, Chernoff, Rényi divergence (Bhattacharyya 1943, Cramér 1946, Chernoff 1952, Rényi 1960) given by $d(P, \tilde{P}) = 2 \log 1/ \int (p(u)\tilde{p}(u))^{1/2}$ where p and $\tilde{p}$, respectively, are the densities of P and $\tilde{P}$ with respect to a reference measure that dominates the distributions and with respect to which the integrals are taken. Writing $D(P||\tilde{P}) = -2E \log(\tilde{p}(U)/p(U))^{1/2}$ and employing Jensen's inequality shows that $D(P||\tilde{P}) \geq d(P, \tilde{P})$.

On a sequence space $\mathcal{U}^n$, if P^n and $\tilde{P}^n$ are n-fold products of the measures P and $\tilde{P}$, then $d(P^n, \tilde{P}^n) = nd(P, \tilde{P})$ and $D(P^n, \tilde{P}^n) = nD(P, \tilde{P})$. Analogous to notation used above, we use $d_n(f^*, f)$ to denote the divergence between the joint distributions $P_{\underline{U}|f^*}$ and $P_{\underline{U}|f}$, and likewise $d(f^*, f)$ to be the divergence between the distributions $P_{U_1|f^*}$ and $P_{U_1|f}$.

We take this divergence to be our loss function in examination of the accuracy of penalized likelihood estimators. One reason is its close connection to familiar distances such as the L_1 distance between the densities and the Hellinger distance (it upper bounds the square of the L_1 distance and the square of the Hellinger distance with which it is equivalent as explained below). Another is that $d(P, \tilde{P})$, like the squared Hellinger distance, is locally equivalent to one-half the Kullback-Leibler divergence when $\log p(u)/\tilde{p}(u)$ is upper-bounded by a constant. Thirdly, it evaluates to familiar quantities in special cases, e.g., for two normals of mean μ and $\tilde{\mu}$ and variance 1, this $d(P, \tilde{P})$ is $\frac{1}{4}(\mu - \tilde{\mu})^2$. Most important though for our present purposes is the cleanness with which it allows us to examine the risk, without putting any conditions on the density functions $p_f(\underline{u})$.

The integral used in the divergence is called the Hellinger affinity $A(P, \tilde{P}) = \int p^{1/2}\tilde{p}^{1/2}$. It is related to the squared Hellinger distance $H^2(P, \tilde{P}) = \int (p(u)^{1/2} - \tilde{p}(u)^{1/2})^2$ by $A = 1 - \frac{1}{2}H^2$ and hence the divergence $d(P, \tilde{P}) = -2 \log A = -2 \log(1 - \frac{1}{2}H^2)$ is not less than $H^2(P, \tilde{P})$. In thinking about the affinity note that it is less than or equal to 1 with equality only when $P = \tilde{P}$. We let $A_n(f^*, f)$ denote the Hellinger affinity between the joint distributions $P_{\underline{U}|f^*}$ and $P_{\underline{U}|f}$. Its role in part of our analysis will be as a normalizer, equaling the expectation of $[p_f(\underline{U})/p_{f^*}(\underline{U})]^{1/2}$ for each fixed f.

The following result from Jonathan Li's 1999 Yale thesis is a simplification of a conclusion from Barron and Cover (1991). It is also presented in Kolaczyk and Nowak (2004) and in Grünwald (2007). We repeat it here because it is a stepping stone for the extensions we give in this paper.

Theorem 2.1. *Resolvability bound on risk (Li 1999). For a countable $\tilde{\mathcal{F}}$, and $\mathcal{L}_{\rm n}(f) = 2L_n(f)$ satisfying $\sum e^{-L_n(f)} \leq 1$, let $\hat{f}$ be the estimator achieving*

$$\min_{f\in\tilde{\mathcal{F}}} \left\{ \log \frac{1}{p_f(\underline{U}_n)} \; + \; \mathcal{L}_{\rm n}(f) \right\}.$$

Then, for any target function f^* and for all sample sizes, the expected divergence of $\hat{f}$ from f^* is bounded by the index of resolvability

$$Ed_n(f^*, \hat{f}) \; \leq \; \min_{f\in\tilde{\mathcal{F}}} \{ \, D_n(f^*, f) + \mathcal{L}_n(f) \, \}.$$

In particular with i.i.d. modeling, the risk satisfies

$$Ed(f^*, \hat{f}) \; \leq \; \min_{f\in\tilde{\mathcal{F}}} \left\{ D(f^*, f) + \frac{\mathcal{L}_n(f)}{n} \right\}.$$

Proof of Theorem 2.1: We have

$$2\log \frac{1}{A_n(f^*, \hat{f})} \; = \; 2\log \left[\frac{(p_{\hat{f}}(\underline{U})/p_{f^*}(\underline{U}))^{1/2} e^{-L(\hat{f})}}{A_n(f^*, \hat{f})} \right] \; + \; \log \frac{p_{f^*}(\underline{U})}{p_{\hat{f}}(\underline{U})} \; + \; \mathcal{L}_n(\hat{f}).$$

Inside the first part on the right side the ratio is evaluated at $\hat{f}$. We replace it by the sum of such ratios over all $f \in \tilde{\mathcal{F}}$ obtaining the bound

$$\leq 2 \; \log \sum_{f\in\tilde{\mathcal{F}}} \left[\frac{(p_f(\underline{U})/p_{f^*}(\underline{U}))^{1/2} e^{-L(f)}}{A_n(f^*, f)} \right] \; + \; \log \frac{p_{f^*}(\underline{U})}{p_{\hat{f}}(\underline{U})} \; + \; \mathcal{L}_n(\hat{f}).$$

Now we take the expected value for $\underline{U}$ distributed according to $P_{\underline{U}|f^*}$. For the expectation of the first part, by Jensen, obtaining a further upper bound, we may bring the expectation inside the log and then bring it also inside the sum. There we note for each fixed f that $E(p_f(\underline{U})/p_{f^*}(\underline{U}))^{1/2} = A_n(P_{f^*}, P_f)$, so there is a cancelation of the ratio. Then all that is left inside the log is $\sum e^{-L(f)}$ which by assumption is not more than 1. Thus the expected value of the first part is bounded by 0. What then remains is the expectation of the pointwise redundancy, which being less than the value at f_n^*, is bounded by the index of resolvability, which completes the proof for the general case. Dividing through by n gives the conclusion for the i.i.d. case.

If $\log p_{f*}(u)/p_f(u) \leq B$ for all u in $\mathcal{U}$, then by Yang and Barron (1998), Lemma 4, we have

$$d(f^*, f) \; \leq \; D(f^*||f) \; \leq \; C_B \, d(f^*, f),$$

for a constant C_B given there that is less than $2 + B$. Consequently, we have the following.

Corollary 2.2. *If, in the i.i.d. case, the log density ratios are bounded by a constant B, that is, if $|\log p_{f*}(u)/p_f(u)| \leq B$ for all $f \in \tilde{\mathcal{F}}$, then there is a constant $C_B \leq 2 + B$ such that the Kullback risk satisfies*

$$ED(f^*, \hat{f}) \leq C_B \min_{f\in\tilde{\mathcal{F}}} \left\{ D(f^*, f) + \frac{\mathcal{L}_n(f)}{n} \right\}.$$

Remarks.

We comment that the presence of the factor 2 in the penalty $\mathcal{L}(f) = 2L(f)$ is a byproduct of using the Chernoff-Rényi divergence with parameter $1/2$. As in the original Barron and Cover (1991) bound, one may replace the 2 with any multiplier strictly bigger than 1, though the best bound there occurs with the factor 2. See Zhang (2006, Thm. 4.1) or Grünwald (2007, Ch. 15) for analogous risk bounds for Chernoff-Rényi divergences with parameter λ between 0 and 1.

Producing an exact minimizer of the complexity penalized estimator can be computationally difficult, but an approximate minimizer is still amenable to analysis by the above method. For instance in Li (1999) and Li and Barron (2000) a version of a greedy algorithm is given for estimating densities by sums of m components from a given dictionary of possible component densities (e.g. Gaussian mixtures). Analysis there shows that with m steps the complexity penalized likelihood is within order $1/m$ of the optimum.

Refinements of the risk bound in Li's thesis deal with the case that the distribution of the data is not near any of the P_f. In this case he extends the result to bound the distance of the estimate from a reversed information projection of the distribution onto a convex hull of the P_f.

Some implications of resolvability bounds on risk are discussed in Barron and Cover (1991). Corresponding results for complexity penalized least squares and other bounded loss functions were developed in Barron (1990). Applications to neural nets were developed in Barron (1991,1994), providing risk bounds for estimation of linear combinations of a dictionary, by penalized least squares with a penalty that incorporates aspects of the ℓ_0 and ℓ_1 norms of the coefficients, but restricted to a countable set (that restriction is lifted by Huang, Cheang and Barron (2008) and the developments we give in the next section). Analogous resolvability bounds for regression and log-density estimation by neural nets in a weakly dependent setting were given in Modha and Masry (1996a,b). For mixture density estimation (including Gaussian mixtures), direct implications of Theorem 2.1 using resolvability calculations are given in Li (1999) and Li and Barron (2000) and, building in part on those developments, Rakhlin, Panchenko, and Murherjee (2005) give related risk results using bounds for Rademacher averages of convex hulls.

Kolaczyk and Nowak (2004,2005), and Willett and Nowak (2005) give implications of Li's theorem for multiscale wavelet image estimation and Poisson intensity function estimation. In some of their investigations the data are functions (e.g. of continuous time or location) but the theory nevertheless applies as they make clear in their settings. Indeed, as we have indicated, the structure of the data $\underline{U}$ (other than that there be a dominating measure for the candidate distributions) is not essential for the validity of the general bounds.

The proof of Theorem 2.1 given here is essentially the same as in Li's Thesis. One slight difference is that along the way we have pointed out that the expected redundancy of the two-stage code is also a bound on the risk. This is also noted by Grünwald (2007) and, as he emphasizes, it even more closely relates the risk and coding notions. The resolvability form is more useful in obtaining bounds that exhibit the tradeoff between approximation accuracy and dimension or complexity.

To be specific, the proof of Theorem 2.1 compares the loss $d_n(f^*, \hat{f})$ with the pointwise redundancy $r_n = \log p_{f^*}(\underline{U})/p_{\hat{f}}(\underline{U}) + \mathcal{L}_n(\hat{f})$ and shows that the difference is a random variable of mean bounded by 0. In a similar manner one can obtain a measure of concentration of this difference.

Theorem 2.3. *Tightness of the relationship between loss and redundancy: The difference between the loss $d_n(f^*, \hat{f})$ and the pointwise redundancy r_n is stochastically less than an exponential random variable of mean 2.*

Proof of Theorem 2.3: As shown in the proof of Theorem 2.1 the difference in question is bounded by

$$2 \log \sum_{f \in \tilde{\mathcal{F}}} \left[\frac{(p_f(\underline{U})/p_{f^*}(\underline{U}))^{1/2} e^{-L(f)}}{A_n(f^*, f)} \right].$$

The probability that this exceeds any positive τ is bounded first by dividing through by 2, then exponentiating and using Markov's inequality, yielding $e^{-\tau/2}$ times an expectation shown in the proof of Theorem 2.1 to be not more than 1. This completes the proof of Theorem 2.3.

Further remarks:

In the i.i.d. case we measure the loss by the individual divergence obtained by dividing through by n. Consequently, in this case the difference between the loss $d(f^*, \hat{f})$ and pointwise redundancy rate is stochastically less than an exponential of mean $2/n$. It is exponentially unlikely (probability not more than $e^{-n\tau/2}$) to be greater than any positive τ.

The original bound of Barron and Cover (1991) also proceeded by a tail probability calculation, though it was noticeably more elaborate than given here. An advantage of that original proof is its change of measure from the one at f^* to the one at f_n^*, showing that questions about the behavior when f^* is true can indeed be resolved by the behavior one would have if one thought of the distribution as being governed by the f_n^* which best resolves f^* at the given sample size.

Remember that in this section we assumed that the space $\tilde{\mathcal{F}}$ of candidate fits is countable. From both statistics and engineering standpoints, it is awkward to have to force a user of this theory to construct a discretization of his space of functions in order to use this penalized likelihood result. We overcome this difficulty in the next section.

3. Risk and Resolvability for uncountable $\mathcal{F}$

We come to the main new contributions of the paper. We consider estimators $\hat{f}$ that maximize $p_f(\underline{U})e^{-pen(f)}$ or, equivalently, that achieve the following minimum:

$$\min_{f \in \mathcal{F}} \left\{ \log \frac{1}{p_f(\underline{U})} + pen(f) \right\}.$$

Since the log ratio separates, for any target p_*, this sample minimization is equivalent to the following,

$$\min_{f \in \mathcal{F}} \left\{ \log \frac{p_*(\underline{U})}{p_f(\underline{U})} + pen(f) \right\}.$$

We want to know for proposed penalties $pen(f)$, $f \in \mathcal{F}$, when it will be the case that $\hat{f}$ has risk controlled by the population-based counterpart:

$$\min_{f \in \mathcal{F}} \left\{ E \log \frac{p_*(\underline{U})}{p_f(\underline{U})} + pen(f) \right\},$$

where the expectation is with respect to $p_*(\underline{U})$. One may specialize to $p_* = p_{f^*}$ in the family. In general, it need not be a member of the family $\{p_f : f \in \mathcal{F}\}$, though when such a bound holds, it is only useful when the target is approximated by such densities.

There are two related aspects to the question of whether such a bound holds. One concerns whether the optimal sample quantities suitably mirror the population quantities even for such possibly larger $\mathcal{F}$, and the other is to capture what is essential for the penalty.

A quantity that may be considered in examining this matter is the discrepancy between sample and population values, defined by,

$$\log \frac{p_*(\underline{U})}{p_f(\underline{U})} - E \log \frac{p_*(\underline{U})}{p_f(\underline{U})}.$$

Perhaps it is ideally centered, yielding mean 0 when defined in this way, with subtraction of the Kullback divergence. However, control of this discrepancy, at least by the techniques of which we are aware, would require control of higher order moments, particularly the variance, which, in order to produce bounds on Kullback risk (using, e.g., Bernstein-type bounds), would require conditions relating the variance of the log density ratios to the expected log ratio. Furthermore Bernstein-type bounds would entail a finite moment generating function of the log-likelihood ratio for generating function parameters in an open neighborhood of 0. Though such development is possible, e.g., if the log densities ratios are bounded, it is not as clean an approach as what follows.

Instead, we use the following discrepancy which is of similar spirit to the above and easier to control in the desired manner,

$$\log \frac{p_*(\underline{U})}{p_f(\underline{U})} - 2 \log \frac{1}{E(p_f(\underline{U})/p_*(\underline{U}))^{1/2}}.$$

This discrepancy does not subtract off as large a value, so it is not mean centered, but that is not necessarily an obstacle if we are willing to use the Hellinger risk, as the control needed of the discrepancy is one-sided in character. No moment conditions will be needed in this analysis other than working with the expected square-roots that give the Hellinger affinities, which are automatically bounded by 1. Note that this expected square root is a value of the moment generating function of the log-likelihood ratio $\log p_f(\underline{U})/p_*(\underline{U})$ and that its logarithm is a value of its cumulant generating function, but only evaluated at the specific positive value $1/2$.

In Theorem 2.1, the penalty $\mathcal{L}(f) = 2L(f)$ is used to show that if it is added to the discrepancy, then uniformly for f in the countable $\tilde{\mathcal{F}}$ (i.e. even with a data-based $\hat{f}$ in place of a fixed f) we have that the expectation of the penalized discrepancy is positive.

This leads us to consider, in the uncountable case, penalties which exhibit a similar discrepancy control. We say that a collection $\mathcal{F}$ with a penalty $pen(f)$ for $f \in \mathcal{F}$ has a *variable–complexity variable–discrepancy cover* suitable for p_* if there exists a countable $\tilde{\mathcal{F}}$ and $\mathcal{L}(\tilde{f}) = 2L(\tilde{f})$ satisfying

$\sum_{\tilde{f}} e^{-L(\tilde{f})} \leq 1$, such that the following condition $(*)$ holds for all $\underline{U}$:

$$\inf_{\tilde{f}\in\tilde{\mathcal{F}}} \left\{ \log \frac{p_*(\underline{U})}{p_{\tilde{f}}(\underline{U})} - 2\log \frac{1}{E(p_{\tilde{f}}(\underline{U})/p_*(\underline{U}))^{1/2}} \; + \; \mathcal{L}(\tilde{f}) \right\}$$

$$\leq \inf_{f\in\mathcal{F}} \left\{ \log \frac{p_*(\underline{U})}{p_f(\underline{U})} - 2\log \frac{1}{E(p_f(\underline{U})/p_*(\underline{U}))^{1/2}} \; + \; pen(f) \right\}. \qquad (*)$$

This condition captures the aim that the penalty in the uncountable setting mirrors an information-theoretically valid penalty in the countable case. We drop reference to dependence of the penalty on the sample size, but since the bounds we develop hold for any size data, there is no harm in allowing any of the quantities involved to change with n. In brief, the above condition will give what we want because the minimum over the countable $\tilde{f}$ is shown to have non-negative expectation and so the minimum over all f in $\mathcal{F}$ will also.

Equivalent to condition $(*)$ is that there be a $\tilde{\mathcal{F}}$ and $L(\tilde{f})$ with $\sum e^{-L(\tilde{f})} \leq 1$ such that for every f in $\mathcal{F}$ the penalty satisfies

$$pen(f) \; \geq \; \min_{\tilde{f}\in\tilde{\mathcal{F}}} \left\{ \log \frac{p_f(\underline{U})}{p_{\tilde{f}}(\underline{U})} - 2\log \frac{E(p_f(\underline{U})/p_*(\underline{U}))^{1/2}}{E(p_{\tilde{f}}(\underline{U})/p_*(\underline{U}))^{1/2}} \; + \; 2L(\tilde{f}) \right\}.$$

That is, the penalty exceeds the minimum complexity plus discrepancy difference. The log ratios separate so the minimizing $\tilde{f}$ does not depend on f. Nevertheless, the following characterization $(**)$ is convenient. For each f in $\mathcal{F}$ there is an associated representor $\tilde{f}$ in $\tilde{\mathcal{F}}$ for which

$$pen(f) \; \geq \; \{ \log \frac{p_f(\underline{U})}{p_{\tilde{f}}(\underline{U})} - 2\log \frac{E(p_f(\underline{U})/p_*(\underline{U}))^{1/2}}{E(p_{\tilde{f}}(\underline{U})/p_*(\underline{U}))^{1/2}} \; + \; 2L(\tilde{f}) \}. \qquad (**)$$

The idea is that if $\tilde{f}$ is close to f then the discrepancy difference is small. Then we use the complexity of such $\tilde{f}$ along with the discrepancy difference to assess whether a penalty $pen(f)$ is suitable. The countable set $\tilde{\mathcal{F}}$ of possible representors is taken to be non-stochastic. Nevertheless, the minimizer in $\tilde{\mathcal{F}}$ will depend on the data and accordingly we allow the representor $\tilde{f}$ of f to also have such dependence. With this freedom, in cases of interest, the variable complexity cover condition indeed holds for all $\underline{U}$, though it would suffice for our purposes that (*) hold in expectation.

One strategy to verify the condition would be to create a metric-based cover of $\mathcal{F}$ with a metric chosen such that for each f and its representor $\tilde{f}$ one has $|\log p_f(\underline{U})/p_{\tilde{f}}(\underline{U})|$ plus the difference in the divergences arranged if possible to be less than a distance between f and $\tilde{f}$. Some examples where this can be done are in Barron and Cover (1991). Such covers give a metric entropy flavor, though the $L(\tilde{f})$ provides variable complexity rather than the fixed log-cardinality of metric entropy. The present theory and applications show such covering by metric balls is not an essential ingredient.

Condition $(**)$ specifies that there be a cover with variable distortion plus complexity rather than a fixed distance and fixed cardinality. This is analogous to the distortion plus rate tradeoff in Shannon's rate-distortion theory. In our treatment, the distortion is the discrepancy difference (which does not need to be a metric), the codebook is the cover $\tilde{\mathcal{F}}$, the codelengths are the complexities $L(\tilde{f})$. Valid penalties $pen(f)$ exceed the minimal sum of distortion plus complexity.

An alternative formulation of penalized likelihood conditions is in Shen (1998). He argues that the estimator is likely to have a penalty value in the set $\{f : pen(f) \leq C\, pen(f^*)\}$, with C near 1. Under his conditions, this set is compact with metric entropy properties permiting appeal to uniform large deviation bounds of log likelihood ratios from Wong and Shen (1995) (which do require relationship between the variance and mean of the log-likelihood ratios) and to results on constrained maximum likelihood from Nemirovskii, Polyak and Tsybakov (1985). One could also appeal then to results in Birgé and Massart (1993,1998) on minimum contrast estimation. There are applications to function classes, including some that go beyond the traditional Sobolev type. However, for that machinery, possibly large constants arise giving a asymptotic rate flavor to the conclusions. Unlike Shen's method, we don't assume that the target f^* must have a finite penalty. What matters is that there be functions f close to f^* that do. Moreover, in seeking risk bounds of the form $\inf\{D(f^*, f) + pen(f)/n\}$ with constants equal to 1, we are striving to make the results of practical interest in non-asymptotic settings.

The ideas we develop here have parallels with other empirical loss, such as the average squared error in regression, explored in a concurrent paper for which some of us are coauthors (Huang, Cheang and Barron 2008), building on work with Cheang originating with his 1998 Yale thesis. That work does center by subtracting the expected loss to define the discrepancies and forces uniform boundedness of the fits, so that variances of the squared errors are proportional to the mean squared errors. The idea bridging from the countable to the uncountable classes by the assumption that the penalty exceeds a complexity penalized discrepancy difference originates with Cong Huang in this regression work. Its use here with densities is simpler, because we use a milder loss function that allow arbitrary densities.

Our main theorem, generalizing Theorem 2.1 to uncountable $\mathcal{F}$, is the following.

Theorem 3.1. *Consider $\mathcal{F}$ and $pen_n(f)$ satisfying the discrepancy plus complexity requirement $(*)$ and the estimator $\hat{f}$ achieving the optimum penalized likelihood*

$$\min_{f\in\mathcal{F}} \left\{ \log \frac{1}{p_f(\underline{U})} + pen_n(f) \right\}.$$

If the data $\underline{U}$ are distributed according to $P_{\underline{U}|f^*}$, then

$$Ed_n(f^*, \hat{f}) \;\leq\; \min_{f\in\mathcal{F}} \left\{ E \log \frac{p_{f^*}(\underline{U})}{p_f(\underline{U})} \;+\; pen_n(f) \right\}.$$

In particular, for i.i.d. modeling,

$$Ed(f^*, \hat{f}) \;\leq\; \min_{f\in\mathcal{F}} \left\{ D(f^*, f) + pen_n(f)/n \right\}.$$

Proof of Theorem 3.1. From the characterization $(**)$, at $f = \hat{f}$ in $\mathcal{F}$ there is an associated $\tilde{f}$ in $\tilde{F}$ with

$$2\log \frac{1}{A_n(P_{f^*}, P_{\hat{f}})} \;\leq\; 2\log \left[\frac{(p_{\tilde{f}}(\underline{U})/p_{f^*}(\underline{U}))^{1/2} e^{-L(\tilde{f})}}{A_n(P_{f^*}, P_{\tilde{f}})} \right] \;+\; \left[\log \frac{p_{f^*}(\underline{U})}{p_{\hat{f}}(\underline{U})} + pen(\hat{f}) \right].$$

The first part of the right side has expectation not more than 0 by the same analysis as in Theorem 2.1 (replacing the ratio inside the log, which is there evaluated at a random $\tilde{f}$, by its sum over all of $\tilde{\mathcal{F}}$ and bringing the expectation inside the log by Jensen's inequality). The expectation of the second part is an expected minimum which is bounded by the minimum expectation. This completes the proof.

In like manner we have the following.

Corollary 3.2. *For $\mathcal{F}$ and $pen_n(f)$ satisfying the discrepancy-complexity requirement, the difference between the loss $d_n(f^*, \hat{f})$ and the pointwise redundancy $r_n = \log p_{f^*}(\underline{U})/p_{\hat{f}}(\underline{U}) + pen_n(\hat{f})$ is stochastically less than an exponential random variable of mean 2.*

Proof of Corollary 3.2. An interpretation of this assertion is that at a particular $f = \hat{f}$ the penalized discrepancy $\log p_{f^*}(\underline{U})/p_f(\underline{U}) - 2\log 1/A_n(f^*, f) + pen_n(f)$ is stochastically greater than $-Z$ where Z is an exponential random variable of mean 2. The requirement on the penalty enforces that uniformly in $\mathcal{F}$ this penalized discrepancy exceeds a minimum complexity penalized discrepancy from the countable class case, which as in the proof of Theorem 2.2 is already seen to be stochastically greater than such a random variable. This completes the proof.

Remark: We complete this section with a further comment on the tool for verification of the requirement on the penalty. Consider the case that f models the log density function of independent random variables $X_1, \ldots, X_n$, in the sense that for some reference density $p_0(x)$ we have

$$p_f(x) = \frac{p_0(x)\, e^{f(x)}}{c_f}$$

where c_f is the normalizing constant. Examining the difference in discrepancies at f and a representing $\tilde{f}$ we see that both $p_0(x)$ and c_f cancel out. What remains for our penalty requirement is that for each f in $\mathcal{F}$ there is a $\tilde{f}$ in a countable $\tilde{\mathcal{F}}$ with complexities $L(\tilde{f})$ for which

$$pen(f) \;\geq\; 2L(\tilde{f}) + \sum_{i=1}^{n} (f(X_i) - \tilde{f}(X_i)) + 2n \log E \exp\left\{\tfrac{1}{2}(\tilde{f}(X) - f(X))\right\}$$

where the expectation is with respect to a distribution for X constructed to have density which is the normalized pointwise affinity $p_a(x) = [p_{f^*}(x) p_f(x)]^{1/2}/A(f^*, f)$.

In the final section, with an ℓ_1 penalty on coefficients, we illustrate how to demonstrate the existence of such representors $\tilde{f}$ of functions f in the linear span of a dictionary of candidate basis functions.

4. Information-theoretic validity of the ℓ_1 penalty

Let $\mathcal{F}$ be the linear span of a dictionary $\mathcal{H}$ of functions. Thus any f in $\mathcal{F}$ is of the form $f(x) = f_\theta(x) = \sum_h \theta_h h(x)$ where the coefficients are denoted $\theta = (\theta_h : h \in \mathcal{H})$. We assume that the functions in the dictionary are bounded. We want to show that a weighted ℓ_1 norm of the coefficients $||\theta||_1 = \sum_h |\theta_h| a_h$ can be used to formulate a valid penalty. Here we use weights $a_h = \|h\|_\infty$. For f in $\mathcal{F}$ we denote $V_f = \min\{\|\theta\|_1 : f_\theta = f\}$. With the definition of V_f extended to a closure of $\mathcal{F}$, this V_f is called the variation of f with respect to $\mathcal{H}$ (this terminology is suggested by the notion of total variation which corresponds to the case that $\mathcal{H}$ consists of indicators of half-spaces). We show that certain multiples of V_f are valid penalties.

The dictionary $\mathcal{H}$ is a finite set of p candidate terms, typically much larger than the sample size. (One can also work with an infinite $\mathcal{H}$ together with an empirical cover as explored in Section 5.) As we shall see, the codelengths of our representors will arise via a variable number of terms times the log cardinality

of the dictionary. Accordingly, for sensible risk bounds, it is only the logarithm of p, and not p itself, that we need to be small compared to the sample size n.

A valid penalty will be seen to be a multiple of V_f, by arranging the number of terms in the representor to be proportional to V_f and by showing that a representor with that many terms suitably controls the discrepancy difference. We proceed now to give the specifics.

The countable set $\tilde{\mathcal{F}}$ of representors is taken to be the set of all functions of the form $\tilde{f}(x) = V\frac{1}{K}\sum_{k=1}^{K} h_k(x)/a_{h_k}$ for terms h_k in $\mathcal{H} \cup -\mathcal{H} \cup \{0\}$, where the number of terms K is in $\{1, 2, \ldots\}$ and the nonnegative multipliers V will be determined from K in a manner we will specify later. We let p be the cardinality of $\mathcal{H} \cup -\mathcal{H} \cup \{0\}$, allowing for h or $-h$ or 0 to be a term in $\tilde{f}$ for each h in $\mathcal{H}$.

The main part of the codelength $L(\tilde{f})$ is $K \log p$ nats to describe the choices of $h_1, \ldots, h_K$. The other part is for the description of K and it is negligible in comparison, but to include it simply, we may use a possibly crude codelength for the integer K such as $K \log 2$ (or more standard codelengths for integers may be used, e.g, of size slightly larger than $\log K$). Adding these contributions of $K \log 2$ for the description of K and of $K \log p$ for the description of $\tilde{f}$ given K, we have

$$L(\tilde{f}) = K \log(2p).$$

Some shortening of this codelength is possible, taking advantage of the fact that the order of the terms $h_1, \ldots, h_K$ does not matter and that repeats are allowed, as will be briefly addressed in Section 5. For simplicity we take advantage of the present form linear in K in the current section.

To establish the existence of a representor $\tilde{f}$ of f with the properties we want, we consider a distribution on choices of $h_1, h_2, \ldots, h_K$ in which each is selected independently, where h_k is h with probability $|\theta_h|a_h/V$ (with a sign flip if θ_h is negative). Here $K = K_f = \lceil V_f/\delta \rceil$ is set to equal V_f/δ rounded up to the nearest integer, where $V_f = \sum_h |\theta_h| a_h$, where a small value for δ will be specified later. Moreover, we set $V = K\delta$, which is V_f rounded up to the nearest point in a grid of spacings δ. When V_f is strictly less than V there is leftover an event of probability $1 - V_f/V$ in which h_k is set to 0.

As f varies, so does the complexity of its representors. Yet for any one f, with $K = K_f$, each of the possibilities for the terms h_k produces a possible representor $\tilde{f}$ with the same complexity $K_f \log 2p$.

Now the critical property of our random choice of $\tilde{f}(x)$ representing $f(x)$ is that, for each x, it is a sample average of i.i.d. choices $Vh_k(x)/a_{h_k}$. Each of these terms has expectation $f(x)$ and variance $V\sum_h |\theta_h| h^2(x)/a_h - f^2(x)$ not more than V^2.

As the sample average of K such independent terms, $\tilde{f}(x)$ has expectation $f(x)$ and variance $(1/K)$ times the variance given for a single draw. We will also need expectations of exponentials of $\tilde{f}(x)$ which is made possible by the representation of such an exponential of sums as the product of the exponentials of the independent summands.

The existence argument proceeds as follows. The quantity we need to bound to set a valid penalty is the minimum over $\tilde{\mathcal{F}}$ of the complexity-penalized discrepancy difference:

$$2L(\tilde{f}) + \sum_{i=1}^{n} (f(X_i) - \tilde{f}(X_i)) + 2n \log \int p(x) \exp\{\tfrac{1}{2}(\tilde{f}(x) - f(x))\}$$

where $p(x) = p_a(x)$ is a probability density function as specified in the preceding section. The minimizing $\tilde{f}$ gives a value that is not more than the expectation over random $\tilde{f}$ obtained by the sample average of randomly selected h_k. We condition on the data $X_1, \ldots X_n$. The terms $f(X_i) - \tilde{f}(X_i)$ have expectation 0 so it remains to bound the expectation of the log term. The expected log is less than or equal to the log of the expectation and we bring that expectation inside the integral. So, indeed, at each x we are to examine the expectation of the exponential of $\frac{1}{2}[\tilde{f}(x) - f(x)]$. By the independence and identical distribution of the K summands that comprise the exponent, the expectation is equal to the Kth power of the expectation of $\exp\{\frac{1}{2K}[Vh(x)/a_h - f(x)]\}$ for a randomly drawn h.

We now take advantage of a classical bound of Hoeffding, easily verified by using the series expansion of the exponential. If T is a random variable with range bounded by B, then $E\exp\{\frac{1}{K}(T-\mu)\} \leq \exp\{\frac{B^2}{8K^2}\}$.

For any given x, let $R(x) = \max_h h(x)/a_h - \min_h h(x)/a_h$ be the range of $h(x)/a_h$ as h varies, which is uniformly bounded by 2. At x fixed, $T = \frac{1}{2}Vh(x)/a_h$ is a random variable, induced by the random h, having range $\frac{V}{2}R(x)$ not more than V. Then at the given x, using the Hoeffding inequality gives that the expectation of $\exp\{\frac{1}{2}(\tilde{f}(x) - f(x))\}$ is bounded by $\exp\{\frac{(V^2}{8K}\}$.

The expectation of the log of the integral of this exponential is bounded by $\frac{V^2}{8K}$ or equivalently $\frac{1}{8}V\,\delta$, when multiplied by $2n$ yields a discrepancy difference bound of

$$\tfrac{1}{4}n\,V\,\delta,$$

where V is not more than $V_f + \delta$.

Now twice the complexity plus the discrepancy bound has size $2K\log(2p) + \frac{1}{4}nV_f\delta + \frac{1}{4}n\delta^2$, which, with our choice of $K = \lceil V_f/\delta \rceil$ not more than $V_f/\delta + 1$, shows that a penalty of the form

$$pen_n(f) \;\geq\; \lambda V_f \;+\; C$$

is valid as long as λ is at least $\frac{2}{\delta}\log(2p) + \frac{1}{4}n\delta$ and $C = 2\log(2p) + \frac{1}{4}n\delta^2$. We set $\delta = (\frac{8\log 2p}{n})^{1/2}$ as it optimizes the bound on λ producing a critical value λ_n^* equal to $(2n\log 2p)^{1/2}$ and a value of $C = 4\log(2p)$. We note that the presence of the constant term $C = 4\log(2p)$ in the penalty does not affect the optimization that produces the penalized likelihood estimator, that is, the estimator is the same as if we used a pure ℓ_1 penalty equal to λV_f. Nevertheless, for application of our theory giving risk bounds, the C found here is part of our bound.

We summarize the conclusion with the following Theorem. The setting is as above with the density model $p_f(x)$ with exponent $f(x)$. The estimate is chosen with f in the linear span of the dictionary $\mathcal{H}$. The data are i.i.d. according to $p_{f^*}(x)$.

Theorem 4.1. *The ℓ_1 penalized likelihood estimator $\hat{f} = f_{\hat{\theta}}$ achieving*

$$\min_{\theta} \left\{ \log \frac{1}{p_{f_\theta}(\underline{X}_n)} + \lambda_n ||\theta||_1 \right\},$$

or, equivalently,

$$\min_{f} \left\{ \log \frac{1}{p_f(\underline{X}_n)} + \lambda_n\, V_f \right\},$$

has risk $Ed(f^*, \hat{f})$ *bounded for every sample size by*

$$R_n(f^*) \;\le\; \inf_{f \in \mathcal{F}} \left\{ D(f^*, f) + \frac{\lambda_n V_f}{n} \right\} \;+\; \frac{4 \log 2p}{n}$$

provided $\frac{\lambda_n}{n} \ge \left[\frac{2 \log(2p)}{n} \right]^{1/2}$.

In particular, if f^* has finite variation V_{f^*} then for all n,

$$Ed(f^*, \hat{f}) \;\le\; R_n(f^*) \le \frac{\lambda_n V_{f^*}}{n} + \frac{4 \log 2p}{n}.$$

Note that the last term $\frac{4 \log 2p}{n}$, is typically negligible compared the main term, which is near

$$\left[\frac{2 \log 2p}{n} \right]^{1/2} V_{f^*} \, .$$

Not only does this result exhibit $\left[\frac{\log p}{n} \right]^{1/2}$ as the rate of convergence, but also it gives clean finite sample bounds.

Even if V_{f^*} is finite, the best resolvability can occur with simpler functions. In fact, until n is large compared to $V_{f^*}^2 \log p$, the index of resolvability will favor approximating functions f_n^* with smaller variation.

5. Refined resolvability for ℓ_1 penalized log likelihood

Three directions of refinement of this risk conclusion for ℓ_1 penalized log likelihood are presented briefly here, using the techniques introduced above. These parallel corresponding refinements for ℓ_1 penalized least squares in Huang et al (2008). This section may be skipped by those who only want the overview and who want to move to the computation results of Section 6. The present material is for readers who want to see some of the nuances of statistical rates of density estimation using ℓ_1 controls.

One refinement is that a valid codelength bound for $\tilde{f}$ can take the form $K \log(4e \max\{p/K, 1\})$ which is smaller when $K > 2e$. This leads to an improvement in the risk conclusion in which λ_n^* is as above but with $4e \max\{p/\sqrt{n}, 1\}$ in place of $2p$ inside the log factor so that the log factor may be replaced by a constant when p is small, not more than a multiple of $\sqrt{n}$. The idea of this improvement originates in the setting of Bunea et al (2007a). This improved codelength and risk conclusion follows directly from the above argument using Huang et al (2008), Lemmas 8.5 and 8.6 so we omit the detail. This refinement does not improve the order of the bound when the dictionary size p is a larger order power of n.

Secondly, we consider infinite dictionaries with a finite metric dimension property, and show that a suitable cover of the dictionary has size about $n^{d/2}$ where d is the metric dimension of the dictionary. Then analogous conclusions obtain with $\log p$ replaced by $(d/2) \log n$, so that if f^* has finite variation with respect to the dictionary then the risk is of order bounded by $\left[\frac{d \log n}{n} \right]^{1/2}$. Thus the performance of the ℓ_1 penalized log-likelihood estimator is in agreement with what was obtained previously for other estimators in Barron (1991,1994), Modha and Masry (1996a,b), Lee, Bartlett, and Williamson (1996), Juditsky and Nemirovski (2000), Barron, Cohen, Dahmann and Devore (2008); where a noteworthy

feature is that unlike standard derivative-based regularity conditions which lead to rates that degrade with dimension, the variation condition with respect to a finite-dimensional dictionary has rate of statistical risk at least as good as the power $1/2$.

To explain, suppose a library $\mathcal{H}$ has the properties that $\|h\|_\infty \leq b$ and there are positive constants c and d such that for each positive $\varepsilon \leq b$ there is a finite L_∞ cover $\tilde{\mathcal{H}}$ of size $M_\varepsilon \leq (c/\varepsilon)^d$. Here the cover property is that for each h in $\mathcal{H}$ there is a representor $\tilde{h}$ in $\tilde{\mathcal{H}}$ with $\|h - \tilde{h}\|_\infty \leq \varepsilon$. Then d is called the metric dimension of $\mathcal{H}$. (As shown in Barron (1991,1994) this property holds for the dictionary of Lipshitz sigmoidal functions in d variables used in single hidden layer neural networks and related classes of sinusoidal functions; see Barron, Birgé and Massart (1999) for other examples.) Again with $\mathcal{F}$ the linear span of $\mathcal{H}$, functions take the form $f(x) = \sum_{h\in\mathcal{H}} \theta_h h(x)$ in $\mathcal{F}$ and we consider optimization of the ℓ_1 penalized log likelihood.

To adapt the above proof, we use the unweighted ℓ_1 norm $\|\theta\|_1 = \sum_{h\in\mathcal{H}} |\theta_h|$, multiplying by b^2 in the bound on the $v(x)$ to account a_h equal to 1 rather than $\|h\|_\infty$, and let V_f is the infimum of such $\|\theta\|_1$ among representations satisfying $f_\theta = f$. To obtain a representor $\tilde{f}$, we again draw $h_1, \ldots, h_K$ independently, with distribution that yields h with probability $|\theta_h|/V$, with $K = K_f$ and V as before. The new step is to replace each such h_j with its representor $\tilde{h}_j$ in $\tilde{\mathcal{H}}$, which changes the value of each $\tilde{f}(x)$ by at most $V\varepsilon$. Thus the discrepancy studied above

$$\sum_{i=1}^n (f(X_i) - \tilde{f}(X_i)) + 2n \log \int p(x) \exp\{\tfrac{1}{2}(\tilde{f}(x) - f(x))\}$$

is increased by at most $2nV\varepsilon$, while the complexity is the same as before with the cardinality p replaced by M_ε. This yields a complexity penalized discrepancy bound of $2K_f \log(2M_\varepsilon) + \frac{1}{4}nb^2(V_f + \delta)\delta + 2n(V_f + \delta)\varepsilon$, where the three terms correspond to the three parts of the above analysis: namely, the complexity, the discrepancy, and the contribution of the cover of the dictionary.

Consequently, we have validity of the penalty $pen_n(f) = \lambda_n V_f + C$, with λ_n at least $\lambda_n^* = \frac{2}{\delta}\log(2M_\varepsilon) + \frac{1}{4}nb^2\delta + 2n\varepsilon$ and $C = 2\log(2M_\varepsilon) + \frac{1}{4}nb^2\delta^2 + 2n\delta\varepsilon$. Setting $\delta = \frac{1}{b}\left[\frac{8}{n}\log(2M_\varepsilon)\right]^{1/2}$ produces the best such λ_n^* equal to $b\left[2n\log(2M_\varepsilon)\right]^{1/2} + 2n\varepsilon$. With M_ε replaced by the bound $(c/\varepsilon)^d$, to balance the two terms in λ_n^* we set $\varepsilon = b\sqrt{d/n}$, valid for $d \leq n$. Then M_ε is within a constant factor of $(n/d)^{d/2}$ and we have the desired risk conclusion in a slightly improved form. Indeed, for any sequence of dictionaries and sample sizes with d/n small, $\frac{\lambda_n^*}{n}$ is near $b\left[\frac{d}{n}\log\frac{n}{d}\right]^{1/2}$ and $\frac{C}{n}$ is near $2\left[\frac{d}{n}\log\frac{n}{d}\right]$. To summarize, with λ_n not less than this λ_n^* for dictionaries of finite metric dimension, we have the resolvability bound on risk of the ℓ_1 penalized likelihood estimator:

$$Ed(f^*, \hat{f}) \;\leq\; R_n(f^*) \;\leq\; \inf_{f\in\mathcal{F}} \left\{ D(f^*, f) + \frac{\lambda_n V_f}{n} \right\} \;+\; \frac{C}{n}.$$

A feature of this analysis of resolvability of densities is that the constructed variable-complexity cover $\tilde{\mathcal{F}}$ is not data-dependent. This necessitated our appeal to L_∞ covering properties of the dictionary in constructing the set of representors $\tilde{\mathcal{F}}$. Results for least squares in Lee, Bartlett, and Williamson (1996) and for penalized least squares in Huang et al (2008) allow for data-dependent covers (depending on observed and hypothetical input data), and accordingly allow for empirical L_1 or L_2 covering properties of the dictionary, thus allowing traditional step sigmoids in the neural net case. It is not clear whether

there is a method to allow data-dependent covers in risk analysis for density estimation by penalized likelihood.

Thirdly, an improved method of approximation with probabilistic proof originates in the L_2 case in Makovoz (1996), with a stratified sampling interpretation in Huang et al (2008). It yields an improvement in which V^2/K is replaced by $\varepsilon_0^2 V^2/(K-K_0)$ where ε_0 is the distance attained by the best covering of the dictionary of size $K_0 < K$. We find it allows a somewhat smaller λ_n and improved risk bounds for ℓ_1 penalized log-likelihood estimators of order $\left[\frac{d}{n}\log\frac{n}{d}\right]^{\frac{1}{2}+\frac{1}{2d+2}}$, which remains near the rate $1/2$ when the dimension d is large. This conclusion is in agreement with what is achieved by other estimators in Yang and Barron (1999) and close to the lower bound on optimal rates given there. Similar implications for classification problems using convex hulls of a dictionary are in Koltchinskii and Panchenko (2005). The refined conclusion for ℓ_1 penalized least squares is given in Huang et al (2008) using empirical L_2 covering properties based on Makovoz's result.

Adapting the stratified sampling argument to ℓ_1 penalized log likelihood and the use of L_∞ covering properties proceeds as follows. Partition $\mathcal{H}$ into K_0 disjoint cells c. Let $v(c) \geq \sum_{h\in c}|\theta_h|$ and $f_c(x) = \frac{1}{v(c)}\sum_{h\in c}\theta_h h(x)$ which decomposes $f(x) = \sum_{h\in\mathcal{H}}\theta_h h(x)$ as $f(x) = \sum_c v(c) f_c(x)$. Consider positive integers $K(c)$. A convenient choice is $v(c) = \eta K(c)$ with $K(c) = \lceil \sum_{h\in c}|\theta_h|/\eta \rceil$. For each cell c draw $h_{c,k}$, for $k = 1, 2, \ldots, K(c)$, independently with outcome h with probability $\theta_h/v(c)$ for h in c (and outcome 0 with any leftover probability due to $v(c)$ possibly larger than $\sum_{h\in c}|\theta_c|$). Form the within cell sample averages $f_{c,K}(x) = \frac{1}{K(c)}\sum_{k=1}^{K(c)} h_{c,k}(x)$ and the random representor $\tilde{f}(x) = \sum_c v(c) f_{c,K}(x)$, which is seen to be an equally weighted average when $v(c)$ is proportional to $K(c)$. Now with $a_h = 1$ we proceed as in the analysis in the previous section, with the following exception.

For each x, the expectation of $\exp\{\frac{1}{2}[\tilde{f}(x) - f(x)]\}$ with respect to the distribution of the random terms is again straightforward by the independence of the $h_{c,k}$, but now they are not all identically distributed. This expectation becomes the product across the cells c of the $K(c)$ power of the expectation of $\exp\{\frac{1}{2}\frac{v(c)}{K(c)}[h_{c,1}(x) - f_c(x)]\}$. By the Hoeffding bound each of these expectations is not more than $\exp\{\frac{1}{32}(v(c)/K(c))^2 R_c(x)\}$, where $R_c(x) = \max_{h\in c} h(x) - \min_{h\in c} h(x)$ is the range of $h(x)$ for $h \in c$ for each x. With $mid_c(x) = [\min_{h\in c}(x) + \max_{h\in c}(x)]/2$ equal to the midrange function we recognize that it is the choice of function representing cell c optimizing $\max_{h\in c}|h(x) - mid_c(x)|$, equal to the half-range $R_c(x)/2$. Then we bound $R_c(x)$ by $\|R_c\|_\infty = 2\max_{h\in c}\|h - mid_c\|_\infty$, which is not more than $2\varepsilon_0$ if the partition is arranged to correspond to the best L_∞ cover of $\mathcal{H}$ of size K_0. Accordingly, the expectation of $2n\log\int p_a(x)\exp\{\frac{1}{2}[\tilde{f}(x) - f(x)]\}$ is not more than $\frac{1}{4}n\sum_c \frac{v(c)^2}{K(c)}\varepsilon_0^2$. Choosing $v(c)/K(c) = \eta$ to equal δ/ε_0 and $V = \sum_c v(c)$, this is $\frac{1}{4}nV\delta\varepsilon_0$, improving on the previous bound by the presence of the factor ε_0.

The other difference with the previous analysis is that with $K(c)$ equal to $\sum_{h\in c}|\theta_h|/\eta$ rounded up to an integer, the sum of these counts over the K_0 cells is a total count of $K = K_f$ between V_f/η and $V_f/\eta + K_0$. Likewise, $V = K\eta$ is between V_f and $V_f + K_0\eta$, with $\eta = \delta/\epsilon_0$.

So the complexity penalized discrepancy bound is now $2K_f\log(2M_\varepsilon) + \frac{1}{4}nV\delta\varepsilon_0 + 2nV\varepsilon$. Using the indicated bounds on K and V, and setting $\delta = \left[\frac{8}{n}\log(2M_\varepsilon)\right]^{1/2}$, it is not more than $\lambda_n^* V_f + C$, with $\lambda_n^* = \varepsilon_0\left[2n\log(2M_\varepsilon)\right]^{1/2} + 2n\varepsilon$ the same as before but with the smaller ε_0 in place of

b, which is the source of the improved rate. One sees that a good choice for the relationship between the precisions is $\varepsilon = \varepsilon_0/\sqrt{n}$, with which $C = K_0\left[4\log(2M_\varepsilon) + 2n\delta\varepsilon/\varepsilon_0\right]$ becomes $C = K_0\left[4\log(2M_\varepsilon) + (8\log(2M_\varepsilon))^{1/2}\right]$, the same order as before but with the multiplication by $K_0 \geq 1$. Again the resolvability is $\inf_{f\in\mathcal{F}}\left\{D(f^*,f) + \frac{\lambda_n^* V_f}{n} + \frac{C}{n}\right\}$, with the improved λ_n^* and the inflated C. In particular, in the finite metric dimension case with K_0 of order $(1/\varepsilon_0)^d$, setting ε_0 of order $\left[\frac{d}{n}\log\frac{n}{d}\right]^{\frac{1}{2d+2}}$ one finds that both λ_n^*/n and C/n are of order $\left[\frac{d}{n}\log\frac{n}{d}\right]^{\frac{1}{2}+\frac{1}{2d+2}}$, providing the claimed improvement in rate.

This completes our story of the risk of penalized log likelihood. Common penalties for functions in uncountable sets $\mathcal{F}$ may be used, such as the ℓ_1 norm of the coefficients of f, which may, at first glance, not look like a complexity penalty. Nevertheless, variable cover arguments show that the ℓ_1 penalty does have the property we require. For suitable multipliers λ, the ℓ_1 penalized discrepancy exceeds the complexity penalized discrepancy, and hence inherits its clean risk properties.

6. A Note on Computation

Building on past work on relaxed greedy algorithms, we consider successively optimizing the ℓ_1 penalized likelihood one term at a time, optimizing choices of α, β and h in the update

$$\hat{f}_k(x) = (1-\alpha)\hat{f}_{k-1}(x) + \beta h(x)$$

for each $k = 1, 2, \ldots$. The result is that it solves the ℓ_1 penalized likelihood optimization, with a guarantee that after k steps we have a k component mixture within order $1/k$ of the optimum. Indeed, one initializes with $\hat{f}_0(x) = 0$ and $v_0 = 0$. Then for each step k, ones optimizes α, β, and h to provide the the kth term $h_k(x)$. At each iteration one loops through the dictionary trying each $h \in \mathcal{H}$, solving for the best associated scalars $0 \leq \alpha \leq 1$ and $\beta \in R$, and picks the h that best improves the ℓ_1 penalized log-likelihood, using $v_k = (1-\alpha)v_{k-1} + |\beta|\ a_{h_k}$ as the updated bound on the variation of $\hat{f}_k$. This is a case of what we call an ℓ_1 *penalized greedy pursuit*. This algorithm solves the penalized log-likelihood problem, with an explicit guarantee on how close we are to the optimum after k steps. Indeed, for any given data set $\underline{X}$ and for all $k \geq 1$,

$$\frac{1}{n}\left[\log\frac{1}{p_{\hat{f}_k}(\underline{X})} + \lambda v_k\right] \ \leq\ \inf_f\left\{\frac{1}{n}\left[\log\frac{1}{p_f(\underline{X})} + \lambda V_f\right] + \frac{2V_f^2}{k+1}\right\},$$

where the infimum is over functions in the linear span of the dictionary, and the variation corresponds to the weighted ℓ_1 norm $\|\theta\|_1 = \sum_{h\in\mathcal{H}}|\theta_h|a_h$, with a_h set to be not less than $\|h\|_\infty$. This inequality shows that $\hat{f}_k$ has penalized log-likelihood within order $1/k$ of the optimum.

This computation bound for ℓ_1 penalized log-likelihood is developed in the Yale Thesis research of one of us, Xi Luo, adapting some ideas from the corresponding algorithmic theory for ℓ_1 penalized least squares from Huang et al (2008). The proof of this computation bound and the risk analysis given above have many aspects in common. So it is insightful to give the proof here.

It is equivalent to show for each f in the linear span that

$$\frac{1}{n}\left[\log\frac{p_f(\underline{X}_n)}{p_{\hat{f}_k}(\underline{X}_n)} + \lambda(v_k - V_f)\right] \;\le\; \frac{2V_f^2}{k+1}.$$

The left side of this desired inequality which we shall call e_k is built from the difference in the criterion values at $\hat{f}_k$ and an arbitrary f. It can be expressed as

$$e_k \;=\; \frac{1}{n}\sum_{i=1}^{n}[f(X_i) - \hat{f}_k(X_i)] + \log\int p_f(x)\exp\{\hat{f}_k(x) - f(x)\} \;+\; \lambda[v_k - V_f],$$

where the integral arising from the ratio of the normalizers for $p_{\hat{f}_k}$ and p_f. Without loss of generality, making $\mathcal{H}$ closed under sign change, we restrict to positive β. This e_k is evaluated with $\hat{f}_k(x) = (1-\alpha)\hat{f}_{k-1}(x) + \beta h(x)$ and $v_k = (1-\alpha)v_{k-1} + \beta a_h$, at the optimized α, β and h, so we have that it is as least as good as at an arbitrary h with $\beta = \alpha v/a_h$ where $v = V_f$. Thus for any h we have that e_k is not more than

$$\frac{1}{n}\sum_{i=1}^{n}[f(X_i) - \bar{\alpha}\hat{f}_{k-1}(X_i) - \alpha v h(X_i)/a_h] + \log\int p_f(x)e^{[\bar{\alpha}\hat{f}_{k-1}(x) + \alpha v h(x)/a_h - f(x)]} \;+\; \bar{\alpha}\lambda[v_{k-1} - v],$$

where $\bar{\alpha} = (1-\alpha)$. Reinterpret the integral using the expectation of $e^{\alpha[vh(x)/a_h - f(x)]}$ with respect to $p(x) \;=\; e^{\bar{\alpha}[f_{k-1}(x) - f(x)]}p_f(x)/c$, where c is its normalizing constant. Accordingly, we add and subtract $\log c = \log\int e^{\bar{\alpha}[f_{k-1}(x)-f(x)]}p_f(x)$ which, by Jensen's inequality using $\bar{\alpha} \le 1$, is not more than $\bar{\alpha}\log\int e^{[f_{k-1}(x)-f(x)]}p_f(x)$. Recognizing that this last integral is what arises in e_{k-1} and distributing f between the terms with coefficients $\bar{\alpha}$ and α, we obtain that e_k is not more than

$$(1-\alpha)e_k + \alpha\frac{1}{n}\sum_{i=1}^{n}[f(X_i) - vh(X_i)/a_h] + \log\int e^{\alpha[vh(x)/a_h - f(x)]}p(x).$$

This inequality holds for all h so it holds in expectation with a random selection in which each h is drawn with probability $a_h|\theta_h|/v$ where the θ_h are the coefficients in the representation $f(x) = \sum_{h\in\mathcal{H}}\theta_h h(x)$ with $v = \sum_h |\theta_h|a_h = V_f$. We may bring this expectation for random h inside the logarithm, and then inside the integral, obtaining an upper bound by Jensen's inequality. Now for each x and random h the quantities $[vh(x)/a_h - f(x)]$ have mean zero and have range of length not more than $2v$ since $a_h \ge \|h\|_\infty$. So by Hoeffding's moment generating function bound, the expectation for random h of $e^{\alpha[vh(x)/a_h - f(x)]}$ is not more than $e^{\alpha^2v^2/2}$. Thus

$$e_k \;\le\; (1-\alpha)e_{k-1} + \alpha^2V_f^2$$

for all $0 \le \alpha \le 1$, in particular with $\alpha = 2/(k+1)$, and $e_0 \le 2V_f^2$, so by induction the result holds

$$e_k \le \frac{2V_f^2}{k+1}.$$

This computation bound as well as its regression counterpart in Huang, Cheang and Barron (2008) holds even for $\lambda = 0$, which shows its relationship to past relaxed greedy algorithm work (by Jones 1992, Barron 1993, Lee, Bartlett and Williamson 1996, Cheang 1998, Cheang and Barron 2001, Li and Barron 2000, Zhang 2003 and Barron, Cohen, Dahmen, and DeVore 2008). These previous results remind us that explicit control on the ℓ_1 norm of the estimator is not necessary for similar conclusions. Instead, one can incorporate a penalty on the the number of terms k rather than their ℓ_1 norm and have fast computations

by traditional relaxed greedy pursuit algorithms with $\lambda = 0$. The conclusion in the cited work is that it yields estimators which perform well as captured by risk bounds based on the best tradeoff between the accuracy of functions in the linear span and their ℓ_1 norm of coefficients. The result stated here for ℓ_1 penalized log-likelihood and in Huang et al (2008) for regression, takes the matter a step further to show that with suitable positive λ the greedy pursuit algorithm solves the ℓ_1 penalized problem.

This computation analysis comfortably fits with our risk results. Indeed, the proof of our main risk conclusion (Theorem 3.1) involves the penalized likelihood ratio $\log \frac{p_{f^*}(X)}{p_{\hat{f}}(X)} + pen(\hat{f})$. Instead of the exact penalized likelihood estimator $\hat{f}$, substitute its k term greedy fit $\hat{f}_k$, Then the computation bound of the current section shows that this penalized likelihood ratio is not more than its corresponding value at any f, with addition of $2V_f^2/(k+1)$. Accordingly, its risk is not more than

$$Ed(f^*, \hat{f}_k) \;\leq\; \min_{f \in \mathcal{F}} \left\{ D(f^*, f) + \frac{\lambda_n V_f}{n} + \frac{2V_f^2}{k+1} \right\} + \frac{C}{n}.$$

Finally, we note an intrinsic connection between the computation analysis and the information-theoretic validity of the penalty for statistical risk. Indeed, inspecting the proof of the computation bound we see that it can be adapted to show that $V^2/(K+1)$ bounds the discrepancy divided by n of an associated greedily obtained f_K, which may be used as a representor of f, rather than the sample average $\tilde{f}$ used in Section 4. Moreover with prescription of α_k and β_k, one again can describe such f_K using $K \log(2p)$ bits. Accordingly, the same analysis used to demonstrate the computation bound also demonstrates the information-theoretic validity of the ℓ_1 penalty.

The key step in our results is demonstration of approximation, computation, or covering properties, by showing that they hold on the average for certain distributions on the dictionary of possibilities. As a reviewer notes, as information-theorists we are predisposed to look for opportunity to provide such an argument by Shannon's pioneering work. One can see other specific precursors for the probabilistic proof argument used here. For the purposes of demonstrating information-theoretically valid penalties for log-likelihood for Rissanen's MDL criterion, the idea for the probabilistic argument came in part from its use in the least squares setting, showing approximation bounds by greedy algorithms, in the line of research initiated by Jones.

References

Banerjee, O., L.E. Ghaoui, and A. d'Aspremont (2007). Model selection through sparse maximum likelihood estimation. To appear in the *Journal of Machine Learning Research.* Available at http://arxiv.org/abs/0707.0704

Barron, A. R. (1985). *Logically Smooth Density Estimation.* Ph. D. Thesis, Department of Electrical Engineering, Stanford University, Stanford, CA.

Barron, A. R. (1988). The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions. University of Illinois Department of Statistics Technical Report #7. Available at www.stat.yale.edu/∼arb4/publications.htm

Barron, A. R. (1990). Complexity Regularization with application to artificial neural networks. In G. Roussas (Ed.) *Nonparametric Functional Estimation and Related Topics.* pp. 561-576. Dordrecht, the Netherlands, Kluwer Academic Publishers.

Barron, A. R. (1991). Approximation and estimation bounds for artificial neural networks. *Computational Learning Theory: Proceedings of the Fourth Annual ACM Workshop*, L. Valiant (ed.). San Mateo, California, Morgan Kaufmann Publ. pp. 243-249.

Barron, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*. Vol. 39, pp. 930-945.

Barron, A. R. (1994). Approximation and estimation bounds for artificial neural networks. *Machine Learning*. Vol. 14, pp. 113-143.

Barron, A. R. (1998). Information-theoretic characterization of Bayes performance and the choice of priors in parametric and nonparametric problems. In A. Dawid, J.M. Bernardo, J.O. Berger and A. Smith (Eds.), *Bayesian Statistics*. Vol. 6, pp. 27-52. Oxford University Press.

Barron, A.R., L. Birgé, and P. Massart (1999). Risk bounds for model selection by penalization. *Probability Theory and Related Fields*. Vol. 113, pp. 301-413.

Barron, A.R., and G. H. L. Cheang (2001). Penalized least squares, model selection, convex hull classes, and neural nets. In M. Verleysen (Ed.). *Proccedings of the 9th ESANN*, pp.371-376. Brugge, Belgium, De-Facto Press.

Barron, A.R., A. Cohen, W. Dahmen, and R. DeVore (2008). Approximation and learning by greedy algorithms. *Annals of Statistics*. Vol. 36, No.1, pp. 64-94.

Barron, A. R., and T. M. Cover (1991). Minimum complexity density estimation. *IEEE Transactions on Information Theory*. Vol. 37, No.4, pp. 1034-1054.

Barron, A. R., and N. Hengartner (1998). Information theory and superefficiency. *Annals of Statistics*. Vol. 26, No.5, pp. 1800-1825.

Barron, A.R., J. Rissanen, and B. Yu (1998). The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*. Vol. 44, No.6, pp. 2743-2760. Special Commemorative Issue: Information Theory: 1948-1998.

Bhattacharyya, A. (1943). On a measure of divergence between two statistical populations defined by probability distributions. *Bulletin of the Calcutta Mathematics Society*. Vol. 35, pp. 99-109.

Birgé, L., and P. Massart. (1993). Rates of convergence for minimum contrast estimators. *Probability Theory and Related Fields*. Vol. 97, 113-150.

Birgé, L., and P. Massart. (1998). Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*. Vol. 4, 329-375.

Bunea, F., and A.B. Tsybakov and M.H. Wegkamp (2006). Aggregation and sparsity via ℓ_1 penalized least squares. In G. Lugosi and H.U. Simon (Eds.), *Proceedings of the 19th Annual Conference on Learning Theory: COLT 2006*. pp. 379-391. Springer-Verlag, Heidelberg.

Bunea, F., and A.B. Tsybakov and M.H. Wegkamp (2007a). Aggregation for Gaussian regression. *Annals of Statistics*. Vol. 35, pp. 1674-1697.

Bunea, F., and A.B. Tsybakov and M.H. Wegkamp (2007b). Sparse density estimation with ℓ_1 penalties. In N. Behouty and C. Gentile (Eds.), *Proceedings of the 20th Annual Conference on Learning Theory: COLT 2007*. pp. 530-543. Springer-Verlag, Heidelberg.

Cheang, G. H. L. (1998). *Neural Network Approximation and Estimation of Functions*. Ph.D. Thesis, Department of Statistics, Yale University.

Chen, S.S. and D.L. Donoho (1994). Basis pursuit. *Proceedings of the Asilomar Conference*. www-stat.stanford.edu/~donoho/Reports/1994/asilomar.pdf

Chen, S.S, D.L. Donoho and M. A. Saunders (1999). Atomic decompositions by basis pursuit. *SIAM Journal on Scientific Computing*. Vol. 20. pp. 33-61.

Chernoff, H. (1952). A measure of asymptotic efficiency of test of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*. Vol. 23, pp. 493-507.

Clarke, B. S., and A. R. Barron (1990). Information-theoretic asymptotics of Bayes methods. *IEEE Transactions on Information Theory*. Vol. 36, No.3, pp. 453-471.

Clarke, B. S., and A. R. Barron (1994). Jeffreys' prior is asymptotically least favorable under entropy risk. *Journal of Statistical Planning and Inference*. Vol. 41, pp. 37-60.

Cover, T. M., and J. Thomas (2006). *Elements of Information Theory*. Second Edition. New York, Wiley-Interscience.

Cox, D. D., and F. O'Sullivan (1990). Asymptotic analysis of penalized likelihood and related estimators. *Annals of Statistics*. Vol. 18, pp. 1676-1695.

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press.

Cucker, F., and S. Smale (2001). On the mathematical foundations of learning. *Bulletin of the American Mathematics Society*. Vol. 39. pp. 1-49.

Davisson, L. (1973). Universal noiseless coding. *IEEE Transactions on Information Theory*. Vol. 19, pp. 783-795.

Davisson, L., and Leon-Garcia (1980). A source matching approach to finding minimax codes. *IEEE Transactions on Information Theory*. Vol. 26, pp. 166-174.

de Montricher, G.M., R.A. Tapia and J. R. Thompson (1975). Nonparametric maximum likelihood estimation of probability densities by penalty function methods. *Annals of Statistics*. Vol. 3, pp. 1329-1348.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *Annals of Statistics*. Vol. 32, pp. 407-499.

Friedman, J., T. Hastie, and R. Tibshirani (2007a). Pathwise coordinate optimization. *Annals of Applied Statistics*. Vol. 1, pp. 302-332.

Friedman, J., T. Hastie, and R. Tibshirani (2007b). Sparse inverse covariance estimation with the lasso. *Biostatistics*. Dec. 12.

Gallager, R. G. (1968). *Information Theory and Reliable Communication*. New York, Wiley.

Gallager, R. G. (1974). Notes on Universal Coding, Supplement #3. MIT course 6.441.

Good, I.J., and R. A. Gaskins (1971). Nonparametric Roughness Penalties for Probability Densities. *Biometrika*. Vol. 58, pp. 255-277.

Good, I.J. and R. A. Gaskins (1980). Density Estimation and Bump-Hunting by the Penalized Likelihood Method Exemplified by Scattering and Meteorite Data. *Journal of American Statistical Association*. Vol. 75, pp. 42-73.

Grünwald, P. (2007). *The Minimum Description Length Principle*. Cambridge, MA, MIT Press.

Haussler, D. (1997). A general minimax result for relative entropy. *IEEE Transactions on Information Theory*. Vol. 43, No.4, pp. 1276-1280.

Haussler, D., and A.R. Barron (1993). How well do Bayes methods work for on-line prediction of + or -1 values? In *Computational Learning and Cognition: Proc. Third NEC Research Symposium*, pp.74-100. SIAM, Philadelphia.

Haussler, D., and M. Opper (1997). Mutual Information, metric entropy, and cumulative relative entropy risk. *Annals of Statistics*. Vol. 25, pp. 2451-2492.

Huang, C., and G.H.L. Cheang, and A.R. Barron (2008). Risk of penalized least squares, greedy selection and ℓ_1-penalization from flexible function libraries. Submitted to *Annals of Statistics*.

Jones, L. K. (1992). A simple lemma on greedy approximation in Hilbert spaces and convergence rates for projection pursuit regression and neural network training. *Annals of Statistics*, Vol. 20, p. 608-613.

Juditsky, A. and A. Nemirovski (2000). Functional aggregation for nonparametric regression. *Annals of Statistics*. Vol. 28, pp. 681-712.

Koh, K., and S.J. Kim and S. Boyd (2007). An interior-point method for large-scale ℓ_1 regularized logistic regression. *Journal of Machine Learning Research*. Vol. 8, pp. 1519-1555.

Kolaczyk, E.D., and R. D. Nowak (2004). Multiscale likelihood analysis and complexity penalized estimation. *Annals of Statistics*. Vol. 32, pp. 500-527.

Kolaczyk, E.D., and R. D. Nowak (2005). Multiscale generalized linear models for nonparametric function estimation. *Biometrika*. Vol. 92, No. 1, pp. 119-133.

Koltchinskii, V. and D. Panchenko (2005). Complexities of convex combinations and bounding the generalization error in classification. *Annals of Statistics*. Vol. 33, pp. 1455-1496.

Lafferty, J. (2007). Challenges in statistical machine learning. Presented at *Information Theory and Applications*, University of California, San Diego, Feb. 2007. Video http://ita.ucsd.edu/workshop/07/talks

Lee, W.S., P. Bartlett, and R. C. Williamson (1996). Efficient agnostic learning of neural networks with bounded fan-in. *IEEE Transactions on Information Theory*. Vol. 42, pp. 2118-2132.

Li, J.Q. (1999). *Estimation of Mixture Models*. Ph.D. Thesis, Department of Statistics, Yale University, New Haven, CT.

Li, J.Q., and A. R. Barron (2000). Mixture density estimation. In S. Solla, T. Leen, and K.-R. Muller (Eds.), *Advances in Neural Information Processing Systems*, Vol. 12, pp. 279-285.

Liang, F., and A. R. Barron (2004). Exact minimax strategies for predictive density estimation, data compression, and model selection. *IEEE Transactions on Information Theory*. Vol. 50, pp. 2708-2726.

Makovoz, Y. (1996). Random approximates and neural networks. *Journal of Approximation Theory*, Vol. 85, pp. 98-109.

Meier, L., and S. van de Geer, and P. Bühlmann (2008). The Group Lasso for logistic regression. *Journal of the Royal Statistics Society, Series B*. Vol. 70.

Modha, D., and E. Masry (1996a). Rates of convergence in density estimation using neural networks. *Neural Computation*. Vol. 8, pp. 1107-1122.

Modha, D., and E. Masry (1996b). Minimum complexity regression estimation with weakly dependent observations. *IEEE Transactions on Information Theory*. Vol. 42, pp. 2133-2145.

Nemirovskii, A.S., B.T. Polyak, and A. B. Tsybakov (1985). Rate of convergence of nonparametric estimates of maximum likelihood type. *Problems in Information Transmission*. Vol. 21, pp. 258-272.

Park, M.Y. and T. Hastie (2007). L_1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B*. Vol. 69, pp.659-677.

Rakhlin, A., D. Panchenko, and S. Mukherjee (2005). Risk bounds for mixture density estimation. *ESAIM: Probability and Statistics*, Vol. 9, pp. 220-229.

Rényi, A. (1960). On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*. Vol. 1, pp. 547-561.

Rissanen, J. (1978). Modeling by the shortest data description. *Automatica*. Vol. 14, pp. 465-471.

Rissanen, J. (1983). A universal prior on integers and estimation by minimum description length. *Annals of Statistics*. Vol. 11, pp. 416-431.

Rissanen, J. (1984). Universal coding, information, prediction and estimation. *IEEE Transactions on Information Theory*. Vol. 30, pp. 629-636.

Rissanen, J. (1986). Stochastic complexity and modeling. *Annals of Statistics*. Vol. 14, pp. 1080-1100.

Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry*. Hackensack, NJ, World Scientific.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*. Vol. 42, No. 1, pp. 40-47.

Shannon, C. (1948). The mathematical theory of communication. *Bell System Technical Journal*. Vol. 27, pp. 379-423,623-656.

Shen, X. (1998). On the method of penalization. *Statistica Sinica*. Vol. 8, pp. 337-357.

Shtarkov, Y. M. (1987). Universal sequential coding of single messages. *Problems of Information Transmission*. Vol. 23, No. 3, pp. 3-17.

Silverman, B. (1982). On the estimation of probability function by the maximum penalized likelihood method. *Annals of Statistics*. Vol. 10, pp. 795-810.

Takeuchi, J., and A. R. Barron (1997a). Asymptotically minimax regret for exponential families. In *Proceedings of the Symposium on Information Theory and its Applications*, pp. 665-668.

Takeuchi, J., and A. R. Barron (1997b). Asymptotically minimax regret for exponential and curved exponential families. Fourteen page summary at www.stat.yale.edu/∼arb4/publications.htm for the presentation at the *1998 International Symposium on Information Theory*.

Takeuchi, J., and A. R. Barron (1998). Asymptotically minimax regret by Bayes mixtures. In *Proceedings of the 1998 International Symposium on Information Theory*.

Takeuchi, J., T. Kawabata, and A. R. Barron (2007). Properties of Jeffreys' mixture for Markov Sources. Accepted to appear in the *IEEE Transactions on Information Theory*.

Tapia, R. A., and J. R. Thompson (1978). *Nonparametric Probability Density Estimation* Baltimore: The Johns Hopkins University Press.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO, *Journal of Royal Statistical Society, Series B*. Vol. 58, pp. 267-288.

van de Geer, S. A. (2008). High-dimensional generalized linear models and the LASSO *Annals of Statistics*. Vol. 36, pp. 614-645.

Wahba, G. (1990). *Spline Models for Observational Data*. Philadelphia, SIAM.

Willett, R., and R. Nowak (2005). Multiscale Poisson intensity and density estimation. www.ece.wisc.edu/%7Enowak/multiscale_poisson.pdf

Wong, W. H., and X. Shen (1995). Probability inequalities for likelihood ratios and convergence rate of sieve estimates. *Annals of Statistics*. Vol. 23, pp. 339-362.

Xie, Q., and A. R. Barron (1997). Minimax redundancy for the class of memoryless sources. *IEEE Transactions on Information Theory*. Vol. 43, pp. 646-657.

Xie, Q., and A. R. Barron (2000). Asymptotically minimax regret for data compression, gambling, and prediction. *IEEE Transactions on Information Theory*. Vol. 46, pp. 431-445.

Yang, Y., and A. R. Barron (1998). An asymptotic property of model selection criteria. *IEEE Transactions on Information Theory*. Vol. 44, pp. 117-133.

Yang, Y., and A. R. Barron (1999). Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*. Vol. 27, pp. 1564-1599.

Zhang, H., G. Wahba, Y. Lin, M. Voelker, R.K. Ferris, B. Klein (2005). Variable selection and model building via likelihood basis pursuit. *Journal of American Statistical Association*. Vol. 99, pp. 659-672.

Zhang, T. (2003). Sequential greedy approximation for certain convex optimization problems. *IEEE Transactions on Information Theory*. Vol. 49, pp. 682-691.

Zhang, T. (2006). From epsilon-entropy to KL-entropy: analysis of minimum information complexity density estimation. *Annals of Statistics*. Vol. 34, pp. 2180-2210.

Zhang, T. (2007). Some sharp performance bounds for least squares regression with ℓ_1 regularization. Manuscript at http://stat.rutgers.edu/∼tzhang/pubs.html

Greedy and Relaxed Approximations to Model Selection: A simulation study

Guilherme V. Rocha and Bin Yu

April 6, 2008

Abstract

The Minimum Description Length (MDL) principle is an important tool for retrieving knowledge from data as it embodies the scientific strife for simplicity in describing the relationship among variables. As MDL and other model selection criteria penalize models on their dimensionality, the estimation problem involves a combinatorial search over subsets of predictors and quickly becomes computationally cumbersome.

Two approximation frameworks are: convex relaxation and greedy algorithms. In this article, we perform extensive simulations comparing two algorithms for generating candidate models that mimic the best subsets of predictors for given sizes (Forward Stepwise and the Least Absolute Shrinkage and Selection Operator - LASSO). From the list of models determined by each method, we consider estimates chosen by two different model selection criteria (AIC_c and the generalized MDL criterion - gMDL). The comparisons are made in terms of their selection and prediction performances.

In terms of variable selection, we consider two different metrics. For the number of selection errors, our results suggest that the combination Forward Stepwise+gMDL has a better performance over different sample sizes and sparsity regimes. For the second metric of rate of true positives among the selected variables, LASSO+gMDL seems more appropriate for very small sample sizes, while Forward Stepwise+gMDL has a better performance for sample sizes at least as large as the number of factors being screened. Moreover, we found that, asymptotically, Zhao and Yu's ((1)) irrepresentibility condition (index) has a larger impact on the selection performance of Lasso than on Forward Stepwise. In what refers to prediction performance, LASSO+AIC_c results in good predictive models over a wide range of sample sizes and sparsity regimes. Last but not least, these simulation results reveal that one method often can not serve for both selection and prediction purposes.

1 Introduction

The practice of statistics often refers to making efficient use of observed data to infer relationships among variables in order to either gain insight into an observed phenomenon (interpretation) or be able to make predictions based on partial information (prediction). In this paper, we focus on models designed to uncover how a dependent or response variable $Y \in \mathcal{Y}$ is affected by a set of p predictor variables $X \in \mathbb{R}^p$. Whether the goal is prediction or interpretation, the important task is to learn some "meaningful" or stable characteristics of the data across different samples of the data.

A traditional approach consists of postulating a class of models $\mathcal{F}$ indexed by a parameter β. An estimate $\hat{\beta}$ is often defined as:

$$\hat{\beta} \quad = \quad \arg\min_{\beta \in \mathcal{F}} \quad \sum_i L(Z_i, X_i^T \beta), \tag{1}$$

where $Z_i = (X_i, Y_i), i = 1, \ldots, n$ denotes then n observed data samples; $Y \in \mathbb{R}^n$ and is a vector containing the observed values for the dependent variable; $X \in \mathbb{R}^{n \times p}$ is a matrix containing the observed

values of the predictors in its rows; β is a model within the postulated class; and L is a loss function measuring the goodness of fit to the data Z for the model indexed by β. In this paper, we restrict attention to loss functions L defined by the *negative log-likelihood* (neg-loglikelihood) of probabilistic models. This framework is general enough to accommodate both regression and classification models and encompasses all Generalized Linear Models (2; 3). In this paper we will focus attention on the standard Gaussian linear regression model:

$$Y = X\beta + \varepsilon, \text{ with } \varepsilon \sim N(0, \sigma^2). \tag{2}$$

The minimization in (1) translates in this case to the L_2-loss:

$$\hat{\beta} \quad = \quad \arg\min_{\beta} \left\{ \|Y - X\beta\|^2 \right\}. \tag{3}$$

Some of the information criteria we will be dealing with below also require an estimate of the variance σ^2. Unless otherwise stated, we use the likelihood estimate:

$$\hat{\sigma}^2 \quad = \quad \frac{\|Y - X\hat{\beta}\|^2}{n}. \tag{4}$$

However reasonable the estimate defined by (1) may be, this approach does not account for the fact that we often ignore what is an appropriate class of models in which to fit the data, that is, $\mathcal{F}$ should also be estimated. On the one hand, if we rely solely on the observable empirical neg-loglikelihood $L(Z, .)$ to decide between two classes of model $\mathcal{F}_1 \subset \mathcal{F}_2$, the larger class will be trivially preferred. On the other hand, the simpler model class $\mathcal{F}_1$ may be more representative of any structure contained the data and less sensitive to noise. The Minimum Description Length (MDL) principle introduced by Jorma Rissanen (4–6) addresses this problem by including the cost of coding the model itself into the picture. As more complex models are costlier to describe, parsimony is now rewarded.

The problem of identifying an adequate model class on which to search for an estimate $\hat{\beta}$ has been recognized since the seventies. It has since motivated developments of model (variable) selection criteria that penalize the neg-loglikelihood by measures of complexity of the model including the Minimum Message Length criterion (MML, 7), C_p (8), Akaike's Information Criterion (AIC, 9), Bayesian Information Criterion (BIC, 10), and various MDL methods (e.g. the generalized MDL criterion, gMDL, 11).

For linear models, many model selection criteria involve a penalty in the dimensionality of the model under evaluation (i.e., number of non-zero terms in $\hat{\beta}$), that is, the selected estimate is of the form:

$$\hat{\beta}(\lambda_n) \quad = \quad \arg\min_{\beta} \left\{ 2L(Z, \beta) + \lambda_n \|\beta\|_0 \right\}, \tag{5}$$

where $\|\beta\|_0 = \#\{j : \beta_j \neq 0\}$ and λ_n is a tuning parameter trading-off summarization performance and "complexity" of the model. We will refer to such penalties as ℓ_0-penalties in what follows. Perhaps the two most popular examples of model selection criteria within this family are AIC (9) and BIC (10) for which $\lambda_n = 2$ and $\lambda_n = \log(n)$ respectively. In this paper, we will be working with two criteria related to AIC and BIC: the AIC_C criterion (12) is a finite-sample corrected version of AIC; and the gMDL criterion (11) that tries to combine the virtues of AIC and BIC.

The strict computation of ℓ_0-penalized estimates leads to a costly combinatorial search over all subsets of the largest model: the best subset search problem is an NP hard problem in the number of predictors p as established by a recent formal proof (13). Exact solutions to this problem are computationally infeasible for modern massive data sets where the number of predictors p is in the orders of thousands as in gene expression data analysis and even millions as in text processing applications. Even if the

computation of models involving all subsets were feasible, it would still be wasteful. As the number of models to be compared is large, many of them are indistinguishable within the precision afforded by the number of samples practically available. As an example, a regression model involving 50 predictors (a modest size for many modern data sets) would require the comparison of $2^{50} \sim 10^{15}$ models.

Computationally feasible approximations to the ℓ_0-penalized estimates are currently a very active and exciting field of research in statistics. In this paper, we perform extensive simulations to compare the prediction and variable selection performance of models picked by AIC_C and gMDL from lists of candidate models generated by algorithms that identify subsets that are "approximately" the best for their dimension.

The first approximation we consider is the greedy Forward Stepwise regression for selecting variables. Forward Stepwise regression is an eminently algorithmic procedure. Starting from the null model, it selects the predictor whose coefficient corresponds to the largest sized term on the loss function gradient and refits the model involving the selected parameters at each step. For linear models and under the squared error loss (L_2-loss), that corresponds to picking the variables most correlated with the residuals at each step (see 14). The second approximation we study is the convex relaxation approach in which the ℓ_0-penalization is replaced by the ℓ_1-norm of the candidate vector of coefficients β. Early examples of the use of ℓ_1-norm as a penalization are the non-negative garrote (15), the Least Absolute Shrinkage and Selection Operator (LASSO, 16) and basis pursuit (17). The soft-thresholding rule used in VisuShrink (18) is also intimately related to the ℓ_1-penalty. This relaxation results in a convex penalization, easing the burden of computing estimates defined as the solution to an optimization problem (19).

The remainder of this paper is organized as follows. Section 2 reviews some of the theoretical results concerning exact ℓ_0 and ℓ_1 penalized estimates. Section 3 presents a brief overview of the greedy (Forward Stepwise) and relaxed (LASSO) regularization paths. There, we also present the selection criteria we will consider in our later simulation experiments. In Section 4, we present our simulation setup and results obtained for the squared error loss. Section 5 presents our conclusions.

2 Properties of ℓ_0 and ℓ_1-penalized estimates

The properties of ℓ_0-penalized estimates are well understood as various theoretical results have been obtained since their introduction in the seventies (e.g. 20–22). Two important examples of ℓ_0-penalized estimates are AIC (9) and BIC (10). These two criteria reflect a tension that exists between prediction performance and model selection accuracy. Well known results show that AIC-type criteria have the property of yielding the minimax-rate optimal of the regression function under the predictive L_2-loss (23–28), while BIC like criteria are consistent in terms of model selection (21).

Penalization by the ℓ_1-norm of β aims at solving an approximate solution to a convex relaxation of the optimization problem (5). The approximate estimate $\hat{\beta}_{LASSO}(\lambda_n)$ is defined as:

$$\hat{\beta}_{LASSO}(\lambda_n) \quad = \quad \arg\min_{\beta} L(Z, \beta) + \lambda_n \|\beta\|_1, \tag{6}$$

where $\|\beta\|_1 = \sum_j |\beta_j|$. Despite its relative youth, ℓ_1-penalized estimation has undergone intensive research in recent years and a series of theoretical results concerning its properties have been achieved (e.g. 29–38; 1; 39–41). Many of these results are either asymptotic in nature or concern the behavior of sparse approximation in the noiseless setting. In the deterministic setting, the use of convex relaxation of the ℓ_0-norm by the ℓ_1-norm was shown to recover the correct sparse representation under incoherence conditions (42; 31; 30; 32; 34).

In what concerns the predictive performance of ℓ_1-penalized estimates, results in (29) establish that, based on observed data, the actual out-of-sample prediction error can be estimated with greater pre-

cision for the non-negative garrote (closely related to ℓ_1-penalized estimates) than for subset selection procedures. As a result, non-negative garrote estimates can attain better predictive models than their ℓ_0 counterparts. As we will see in our experimental section, this seems to carry over to the LASSO.

In terms of model selection, asymptotic results by (38), (1) and (40) establish conditions for model selection consistency for ℓ_1-penalized estimates under the L_2-loss in the non-parametric setting (i.e., $p_n \to \infty$ as $n \to \infty$). Here, we define the *irrepresentability index* as:

$$II(\Sigma, \beta) \quad = \quad 1 - \|\Sigma_{21}\Sigma_{11}^{-1}\text{sign}(\beta)\|_\infty \tag{7}$$

where Σ_{11} is the covariance matrix of the covariates with non-zero coefficients and Σ_{21} is the partition of the covariance matrix of the covariates accounting for the correlation between the irrelevant and relevant covariates. Results from (1) show that a sufficient condition for the LASSO to be consistent in model selection for some sequence λ_n as $n \to \infty$ is that:

$$II(\Sigma, \beta) \quad > \quad 0 \tag{8}$$

Later in this paper, we will be investigating the effect of the irrepresentability index on the model selection performance in finite samples. Results in (39) refine the model selection consistency results for the LASSO by determining at what rates the number of relevant covariates q and the number of measured predictors p can increase as n grows for model selection consistency to be preserved.

3 Approximation algorithms and selection criteria

The strict implementation of model selection criteria of the form shown in (5) requires the computation of estimates for all possible subsets. As mentioned before this is both computationally infeasible and wasteful given the large number of candidates that must be compared. It does, however, suggest that two tasks are involved in the selection of a model: generating a series of candidate models and applying a criterion to pick the "best" among them.

We consider two algorithms (Forward Stepwise and LASSO) for generating candidate models based on approximations to the combinatorial problem (5). For selecting estimates out of the lists of candidates created by these two algorithms, we consider two different criteria: AIC_C (12) and gMDL (11).

Before we proceed, we point out that alternative algorithms for generating candidate models and alternative selection criteria exist. Boosting algorithms (43) are an important tool for generating list of candidate models. For an example of Boosting algorithms applied to model selection, see (44). Cross-validation (45–47) is an important tool for choosing among different models, especially in what refers to prediction. It is, however, limited by its computational cost and often inadequate for model selection purposes (24).

3.1 Description of the path-tracing algorithms

Although the exact solution to problem (5) is a combinatorial problem, a natural greedy approximation suggests itself. At first, initialize a set of active parameters $\mathcal{A}$ to be empty and set $\hat{\beta}_0 = 0$ – the sparsest possible solution. Then repeat the following process until no parameters are left out or a local optimal is attained. Pick the parameter corresponding to the entry in the gradient vector $\nabla_\beta L$ with the largest absolute value. Add the chosen parameter to the set $\mathcal{A}$ and refit the model adjusting the estimates of parameters contained in $\mathcal{A}$ (i.e., set the new estimate to be a vector such that the gradient of all variables in $\mathcal{A}$ are zero). We shall refer to this algorithm as the *Forward Stepwise algorithm* for the remainder of this paper. It has close connections to the orthogonal greedy algorithms from approximation theory (see,

for instance, 48; 49). [1]

The convex relaxation approximation takes a different route. As mentioned above, it replaces the exact solution of the problem (5) by an approximation based on convex relaxation as defined in (6). A series of candidate models is generated by letting λ_n vary over $[0, \infty)$. At a first glance, the convex relaxation approach seems radically different from the Forward Stepwise regression algorithm. However, the homotopy/LARS [2] algorithm introduced in (50; 51) to compute all LASSO candidates reveals a close connection between them. The homotopy/LARS algorithm also starts by setting an active set $\mathcal{A}$ of parameters to be empty and set $\hat{\beta}_0 = 0$. At each step, it then selects the parameters with the highest gradients, computes a direction preserving the gradient with respect to all active parameters equal in size and determines a step size in which one of two events happen. Either the gradient corresponding to an inactive term becomes as high as the ones in $\mathcal{A}$ in which case a new term is added to $\mathcal{A}$; or one of the parameter estimates in $\mathcal{A}$ hits zero in which case it is excluded from $\mathcal{A}$.

In the case of linear models fitted using an L_2-loss, an analysis in (51) gives the computational cost of the k-th interaction of these algorithms in terms of the current size of the active set a_k and the number of observed samples n. At the k-th step, the costlier operation to perform is determining the direction of the next step. To do so, it is necessary to invert the matrix $X'_{\mathcal{A}} X_{\mathcal{A}}$. This can be done efficiently by updating its Cholesky decomposition at each step of the algorithm at a cost in the order of $O(a_k^2 + a_k n)$.

For Forward Stepwise, the entire regularization path has exactly $r = \text{rank}(X) \leq \min\{p, n\}$ steps resulting in a cost of the order of $O(r^3 + r^2 n)$ for the entire Forward Stepwise path. The complete LASSO regularization path, on the other hand, allows variables to be dropped and re-added to the model along the way and hence has a random number of steps. Well behaved data will cause the computational cost of the LASSO and Forward Stepwise path to be roughly the same. In particular, if the positive cone condition in (51) is satisfied, the two paths are known to agree, thus involving approximately the same computational effort. On the other hand, the LASSO path is costlier when a lot of variable droppings take place. In our experience, we have observed more correlated designs to be associated with longer and consequently costlier paths for the LASSO.

3.2 Selection criteria for choosing an estimate from the regularization path

The Forward Stepwise and the LASSO algorithms above generate each a collection of models for us to choose from, which we call their *regularization paths*. We will focus our attention on two different criteria for picking models from the Forward Stepwise and LASSO regularization paths: the AIC_C (12) (corrected AIC) and the gMDL (11) criteria. We decide for these two criteria based on the good results reported in (11), (52; 44), and (53).

The AIC_C was proposed by Sugiura (12) as a finite sample correction for Akaike's AIC (9). The authors have previously used this criterion in the $n < p$ setting with good predictive performance (54). We use it here in place of cross-validation to reduce the computational cost of our experiments. For linear models based on the L_2-loss (Gaussian likelihood for residuals), the AIC_C estimate are defined as:

$$\hat{\beta}_{\text{AIC}_\text{C}} \quad = \quad \arg\min_{\beta \in \text{path}} \left\{ \frac{n}{2} \log\left(\sum_{i=1}^{n} \|Y_i - X_i \beta\|^2 \right) + \frac{1}{2} \cdot \frac{n\left(1 + \frac{K(\beta)}{n}\right)}{1 - \left(\frac{K(\beta)+2}{n}\right)} \right\}.$$

where $K(\beta)$ denotes an effective dimension of the model associated to β.

The second criterion we consider is the gMDL (11) criterion motivated as a data-driven bridging the AIC and BIC. We refer the reader to (11) for more details on the gMDL criterion. For a Gaussian (L_2-

[1] Boosting algorithms in their turn relate to the pure greedy algorithms in approximation theory.

[2] LARS standing for Least Angle Regression and Selection

loss) linear model and again letting $K(\beta)$ again denote an effective dimension of the model associated to β, the gMDL estimate is defined as:

$$\hat{\beta}_{\text{gMDL}} \quad = \quad \arg \min_{\beta \in \text{path}} \text{gMDL}(Z, \beta)$$

with:

$$\text{gMDL}(Z,\beta) = \begin{cases} \log\left(\frac{\|Y - X\beta\|^2}{n - K(\beta)}\right) + \frac{K(\beta)}{2} \log\left(\frac{\frac{\|X\beta\|^2}{K(\beta)}}{\frac{\|Y - X\beta\|^2}{n - K(\beta)}}\right) + \log(n), & \text{if } R^2 > \frac{K(\beta)}{n}, \\ \log\left(\frac{Y'Y}{n}\right) + \frac{1}{2}\log(n), & \text{otherwise.} \end{cases}$$

For both LASSO and the Forward Stepwise one effective dimensionality of the model $K(\beta)$ is given by the number of non-zero terms in β. For the LASSO, this is justified by the unbiased estimate for the degrees of freedom for LASSO estimates introduced in (55).

4 Simulation results

After reviewing the algorithms we will be using and some of the theoretical properties of ℓ_0 and ℓ_1 penalized estimates, we now present the results of our simulations. As seen in Section 3 above, LASSO and Forward Stepwise have some close connections and subtle differences. Natural questions regarding how their differences and similarities translate into selection accuracy and predictive performance arise. Our experiments below are geared to shed some lights on some of these questions.

Throughout this section, we work with the squared error loss (L_2-loss) and use the `lars` package implementation for both the LASSO and Forward Stepwise selection algorithms in `R`.

4.1 Simulation Set-up

The data in our simulations is generated according to:

$$Y \quad = \quad X\beta + \varepsilon,$$

where n observations are available and p predictors can be selected, that is, $Y, \varepsilon \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$. Throughout, $\varepsilon \sim N(0, \mathbb{I}_n)$. The predictors are also Gaussian with $X \sim N(0, \Sigma)$. To avoid systematic biases favoring one or another method, both the covariance matrix of the predictors $\Sigma \in \mathbb{R}^{p \times p}$ and the coefficients of the model $\beta \in \mathbb{R}^p$ are chosen randomly. Σ is given by $\Sigma = \frac{1}{p}W$, where W has a Wishart$(\mathbb{I}_p, \lceil dp \rceil)$ distribution. The higher the multiplier d in the degrees of freedom, the less correlation among the predictors as Σ concentrates around the orthogonal design with increasing d. For the coefficients, we fix the fraction of non-zero coefficients $s \in (0, 1)$ and randomly choose $q = \lfloor sp \rfloor$ coefficients to be non-zero. Conditional on the sparsity structure, the non-zero coefficients are sampled independently from a $N(0, 1)$ distribution and re-normalized to keep the signal to noise ratio fixed at 2.0.

Once the regularization paths are traced according to the Forward Stepwise and LASSO algorithms, model estimates are picked using the AIC_C and gMDL criteria. We also compare the selected models to models chosen from the path based on full information on the model. The *prediction oracle* is defined as the estimate in the path that minimizes the model error $(\hat{\beta} - \beta)'\mathbb{E}(X'X)(\hat{\beta} - \beta)$. The *selection oracle* estimate is the one model in the path minimizing the number of selection errors (i.e., the size of the symmetric difference between the selected set and the true set).

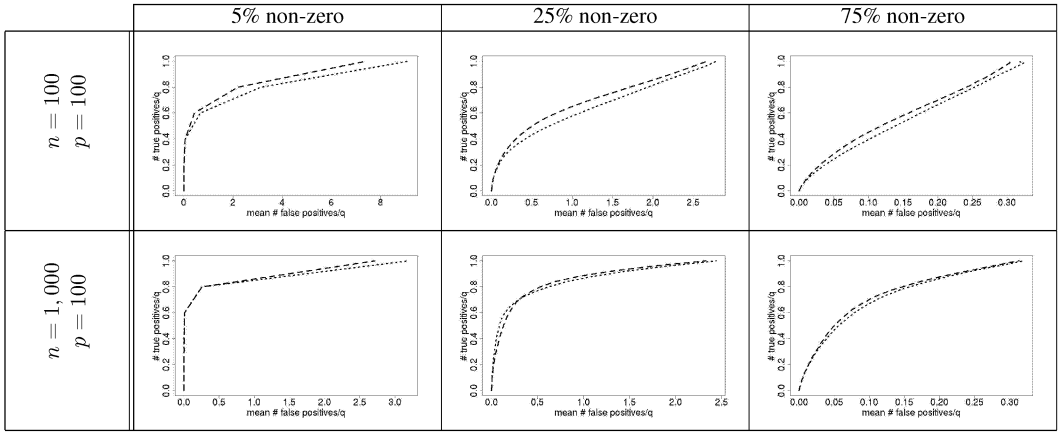


Figure 1: **'Mean ROC curves" for the LASSO (dashed) and Forward Stepwise (dotted):** Within each panel, the relative operating characteristic (ROC) curve shows the mean minimal number of false positives (horizontal axis) needed to achieve a given number of true positives (vertical axis) for both the LASSO (dashed lines) and Forward Stepwise (dotted lines). A selection procedure is better the more its curve approaches the upper left corner of the plot. As we can see here, both the LASSO and Forward Stepwise trade-off between false positives and true positives in a similar fashion for all sample sizes and sparsity levels.

4.2 Model selection results

We first take on the model selection aspects of LASSO and Forward Selection. We start by analyzing how the LASSO and Forward Stepwise trade-off between their ability of detecting true coefficients while keeping irrelevant predictors out of the model.

4.2.1 "Mean" ROC curves for the Forward Stepwise and LASSO

In this first step of our analysis, we compare the relative operating characteristic (ROC) curves for Forward Stepwise and the LASSO. An ROC curve will show the trade-off between the gain of adding a relevant variable and the loss of including an irrelevant variable as we move along the regularization path. By comparing the ROC curves of the LASSO and Forward Stepwise, we have a view of the model selection behavior of the two methods for all possible choices of the tuning parameter λ_n.

To estimate these curves, we fix a number of correctly selected variables and record the mean number of irrelevant variables included in the earliest model in the path containing that many true variables. The estimated curves are shown in Figure 1 for different sample sizes and sparsity levels.

Overall, we see a remarkable similarity in the ROC curves for the LASSO and Forward Stepwise. As a result, the ROC curves suggest that the LASSO and Forward Stepwise have similar behavior in terms of model selection accuracy over a wide range of settings.

4.2.2 The effect of the irrepresentability index and sample size

We now evaluate how much the irrepresentable index (7) affects the ability of the selection oracle to correctly select a model from the LASSO and Forward Stepwise. According to recent theoretical results (38; 1; 40), the presence of a model with all correct variables in the LASSO path is strongly related to

the irrepresentable index. Does the asymptotic results carry over to the small-n-large-p case? And how does the irrepresentable index affect the Forward Stepwise estimates if at all? The results presented in Figure 2 aim at answering these questions.

The two panels on Figure 2 show the minimum number of selection errors (ie, the number of errors committed by the selection oracle) plotted against the irrepresentability index for different sample sizes and sparsity levels as indicated. The results seem to imply that it takes large samples for the irrepresentability index to become a dominant effect on the model selection performance of the LASSO. Another interesting conclusion from Figure 2 is the relative insensitivity of Forward Stepwise to the irrepresentability index, especially in the sparser case. This was a somewhat surprising result given the similarity between the two algorithms. It also suggests that the coherence requirements in (35) as sufficient conditions for Forward Stepwise to recover the sparsest solution are overly restrictive.

4.2.3 Selection Oracle vs. AIC_C and gMDL

We now assess the model selection performance of the AIC_C and gMDL criteria for picking estimates from the LASSO and Forward Stepwise regularization paths. Figures 3 and 4 show the number of selection errors for gMDL and AIC_C and how they compare to the selection oracle for the LASSO and Forward Stepwise at different sparsity levels.

Throughout, gMDL outperforms AIC_C in keeping track of the minimal number of selection errors. Given the model selection consistency (alt. inconsistency) of BIC (alt. AIC) in the parametric case (21), that is not a surprising result: while gMDL strives to combine the virtues of BIC and AIC, the AIC_C simply adjusts the behavior of AIC for finite samples.

Also interesting is the fact that gMDL seems to approach the selection performance of the selection oracle as n increases for Forward Stepwise but not for the LASSO. That suggests that a selection criterion specifically designed for use with the LASSO regularization path can improve upon LASSO estimates picked by gMDL.

4.2.4 Specificity of Forward Stepwise and LASSO estimates

As an important application of variable selection consists of identifying potential relevant factors for further analysis, we now investigate how the Forward Stepwise and LASSO estimates fare in this respect. The important quantity in this case is the proportion of true positive effects among the selected effects. Given the imbalance between the proportion of true positives and true negatives in sparse models, a good performance in terms of number of selection errors does not necessarily translate into performance in terms of correct positive rate. Table 1 reports these results for Forward Stepwise and LASSO estimates picked by gMDL. The results for AIC_C were considerably worse and are not reported.

A high correct positive rate can be achieved by simply over-restricting the estimates. As a control for this, we also report the number of false positives among the selected predictors. A very low number of false positives serves as a warning of over-restriction. Overall, the number of false positives was about the same for Forward Stepwise and LASSO.

When an oracle is available, the correct positive rate for Forward Stepwise is significantly larger for all cases considered. However, when models are picked according to the feasible gMDL, an interesting effect occurs. For smaller samples ($n = 50, p = 100$), the LASSO estimates reach substantially higher correct positive rates than Forward Stepwise. As the sample sizes increase, Forward Stepwise gradually becomes better in the comparison to the LASSO and is preferable for large samples.

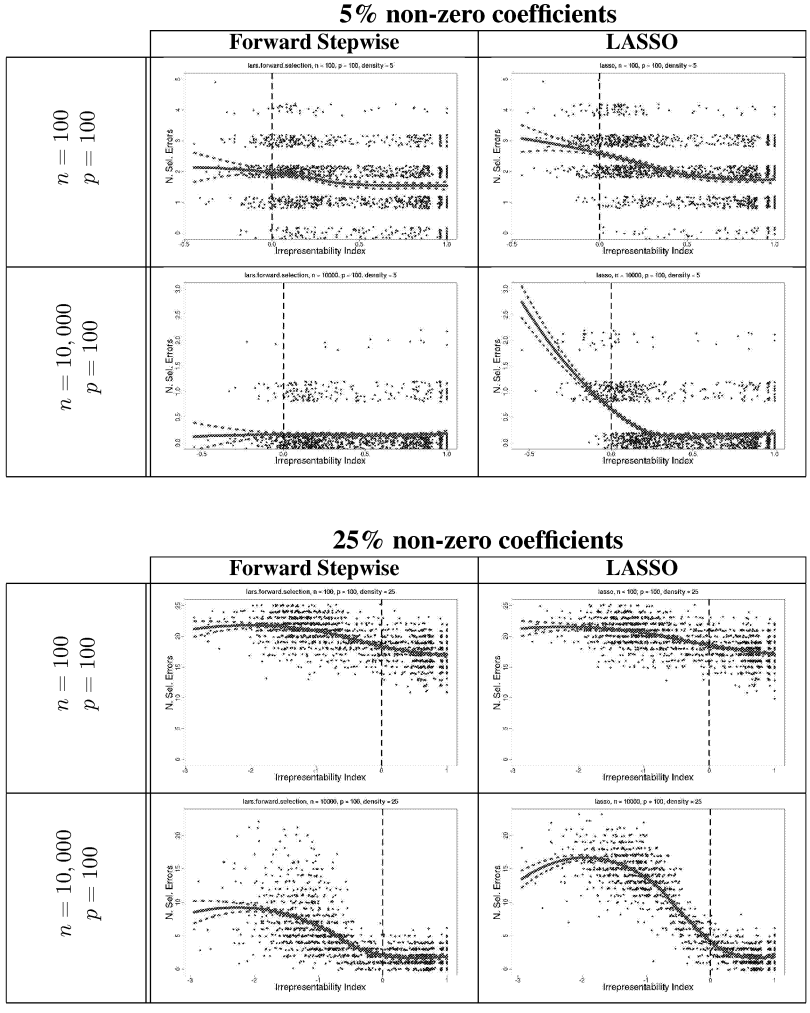


Figure 2: **Number of selection errors for LASSO and Forward Stepwise vs the irrepresentability index:** Each panel shows a plot of the (jittered) selection oracle number of selection errors vs. the irrepresentability index for the approximation and sample size indicated. In small samples, the irrepresentability index does not affect the model selection performance of neither the LASSO nor Forward Stepwise. Asymptotically, the irrepresentability index affects the LASSO more markedly than Forward Stepwise, particularly in the sparsest case.

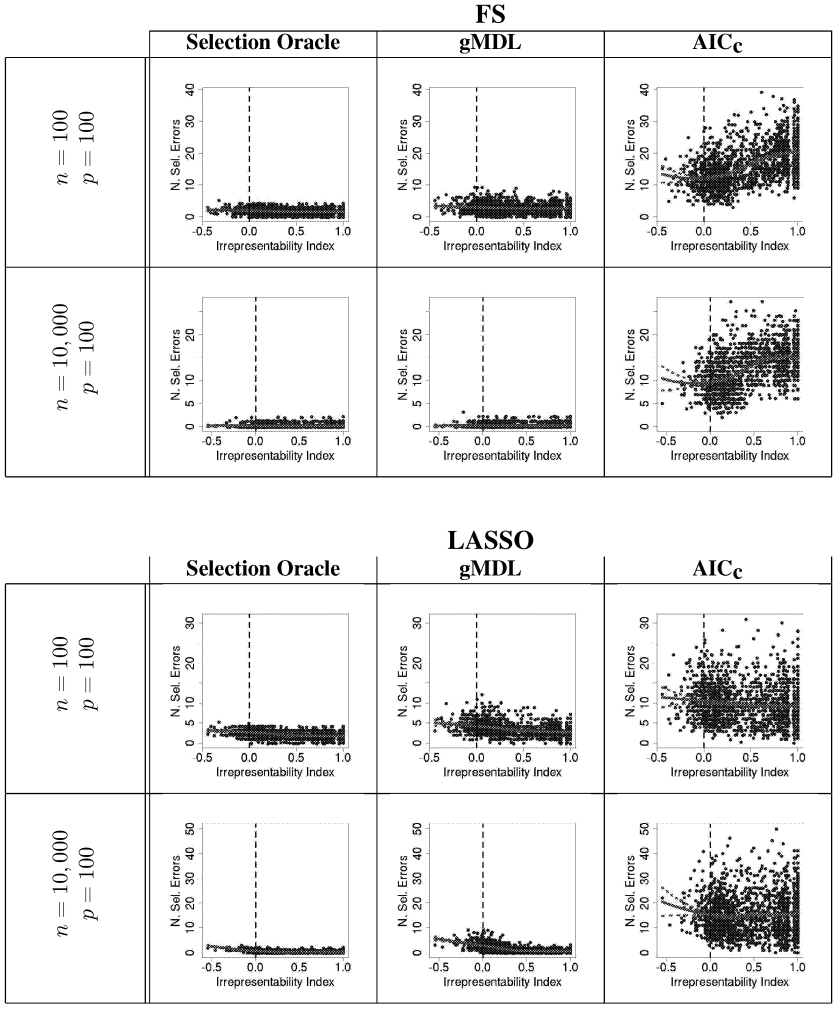


Figure 3: **Number of selection errors under** 5% **non-zero coefficients:** Each panel shows the (jittered) number of selection errors vs. the irrepresentability index for the indicated criterion and sample size. The gMDL criterion had a better performance than AIC_c in terms of number of selection errors for both LASSO and Forward Stepwise and all sample sizes considered. Using gMDL results in a slightly better selection performance for Forward Stepwise in comparison to LASSO.

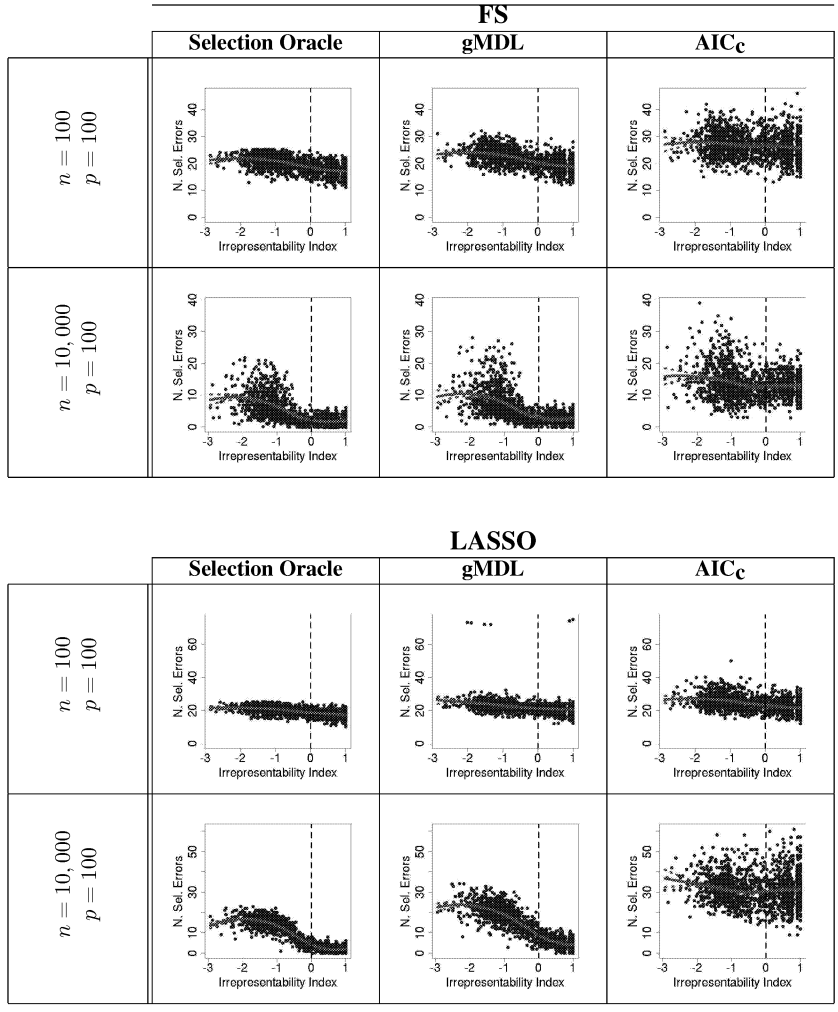


Figure 4: **Number of selection errors under** 25% **non-zero coefficients:** As in Figure 3, each panel shows the (jittered) number of selection errors vs. the irrepresentability index for the indicated criterion and sample size. The gMDL criterion still performs on par or slightly better than AIC_c in terms of number of selection errors for both LASSO and Forward Stepwise and all sample sizes considered. Again, using gMDL results in a slightly better selection performance for Forward Stepwise in comparison to LASSO.

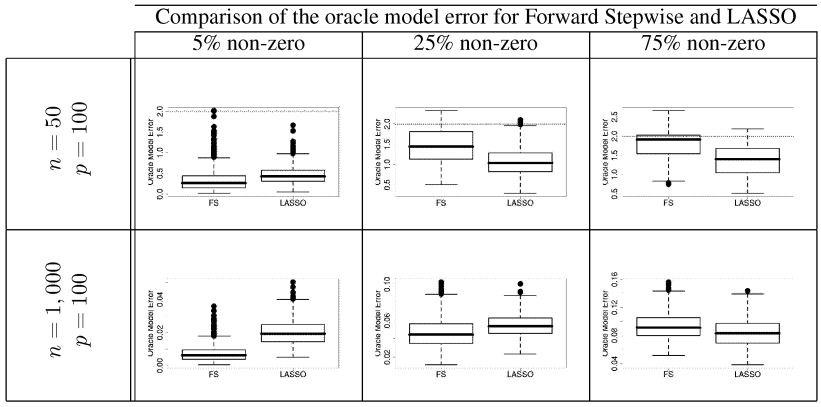


Figure 5: **Oracle model errors for Forward Stepwise and LASSO:** Each panel shows boxplots of the prediction oracle model errors for the Forward Stepwise and LASSO. The dotted lines in the upper panels indicate the model error of the null model (excluding the intercept error). In terms of the oracle model error, LASSO and Forward Stepwise perform similarly. LASSO has a slight advantage in small sample and less sparse settings, while Forward Stagewise seems better for sparser models and large sample sizes. The relative virtues of LASSO and Forward Stepwise for prediction change considerably when an oracle is no longer available (see Figure 6 below).

4.3 Prediction results

We end the exposition of our simulation results by evaluating how the LASSO and Forward Stepwise approximations compare in terms of predictive performance. Figure 5 shows boxplots comparing the model error associated to the LASSO and Forward Stepwise predictive oracles, that is, the models in the regularization path with the minimum model error. The best possible performance depends on the sparsity of the underlying model and the available sample size. In the sparsest case considered, the Forward Stepwise oracle had a better performance than its LASSO counterpart for all sample sizes considered. At less sparse regimes, the LASSO has an advantage for smaller samples, but Forward Stepwise catches up as the sample size increases.

When an oracle is not available and the sample size is small, the AIC_C estimate picked from the LASSO (LASSO+AIC_C estimate) is able to track the model error of the LASSO prediction oracle. The LASSO+AIC_C estimate had a competitive predictive performance across all simulated set-ups. This can be regarded as the LASSO version of earlier experimental and theoretical results (15; 29) for the non-negative garrote estimates. For large sample sizes and very sparse models, however, the Forward Stepwise+gMDL estimate can outperform the LASSO+AIC_C.

5 Discussion/Concluding Remarks

The MDL framework introduced by Jorma Rissanen is an instrumental tool in extracting knowledge from data. However, the high dimensional nature of many modern data sets poses computational challenges due to the combinatorial nature of the optimization problem defining many MDL estimates. A common approach to circumvent this problem consists in applying model selection criteria to a reduced list of candidates generated by algorithms that heuristically identify potentially good models.

In this paper, we present a series of experiments comparing models selected from the regularization

			Correct positive rate				# False positives			
			Sel. Oracle		gMDL		Sel. Oracle		gMDL	
n	p	q	FS	LASSO	FS	LASSO	FS	LASSO	FS	LASSO
50	100	5	98.7	97.4	66.3	75.6	2.49	2.74	2.34	2.73
			0.2	0.2	0.6	0.6	0.027	0.026	0.024	0.026
50	100	25	89.5	83.8	57.3	71.1	21.58	19.97	21.54	22.91
			0.4	0.4	0.6	0.8	0.064	0.088	0.040	0.039
100	100	5	99.1	98.0	80.3	75.2	1.67	2.01	1.63	1.82
			0.1	0.2	0.5	0.6	0.026	0.026	0.025	0.027
100	100	25	87.5	83.0	70.7	74.0	17.94	16.97	18.58	19.18
			0.3	0.3	0.5	0.5	0.092	0.099	0.060	0.082
1000	100	5	99.9	99.2	97.9	81.5	0.50	0.70	0.60	0.58
			0.0	0.1	0.2	0.5	0.017	0.019	0.018	0.018
1000	100	25	92.2	87.1	88.1	69.8	8.023	8.87	8.83	6.23
			0.3	0.3	0.4	0.4	0.109	0.118	0.085	0.072

Table 1: **Proportion of correct positives according to regression type and selection criterion:** If an oracle is available, Forward Stepwise can reach higher proportions of correctly selected variables than LASSO. Between gMDL and AIC_C, gMDL proved better for screening (hence, AIC_C is not shown). LASSO+gMDL is a better screener in small samples and Forward Stepwise+gMDL is a better screener for larger samples. Notice that the number of false positives is roughly the same for LASSO+gMDL and Forward Stepwise+gMDL within each experimental settings (n,p,q).

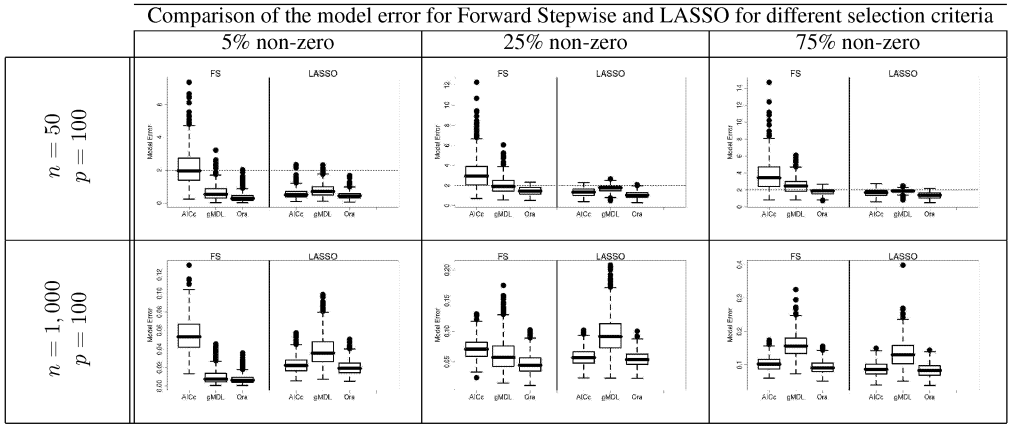


Figure 6: **Model errors for Forward Stepwise and LASSO for gMDL and AIC_C:** Each panel shows a boxplot of the model errors for Forward Stepwise and the LASSO and different selection criteria as indicated (Ora is the predictive oracle). The dotted line shows the model error of the null model. Throughout, the LASSO+AIC_C estimate managed to track the LASSO prediction oracle model error. The gMDL criterion can keep a good track of the oracle model error for Forward Stepwise in the sparsest case. Overall, LASSO+AIC_C have steadier predictive performance: it far exceeds Forward Stepwise+gMDL in the less sparse cases and it performs on par with Forward Stepwise+gMDL in the sparsest case.

path of either greedy (Forward Stepwise) or convex relaxation (LASSO) algorithms and selected by either AIC_C or the gMDL. We compare the selected models according to their prediction and variable selection performances.

In what concerns variable selection accuracy, the list of models generated by Forward Stepwise and the LASSO trade-off very similarly between false negatives and false positives, as evidenced by the experimental mean ROC curves (see Figure 1 for a definition). In terms of the number of variable selection errors, the Forward Stepwise+gMDL estimates seemed to have the best performance over the cases considered. For maximizing the correct positive rate among the selected variables, gMDL had the best results. For sample sizes smaller than the number of predictors being selected, the combination LASSO+gMDL had a better performance. As the sample sizes increased, the combination Forward Stepwise+gMDL achieved the best results.

Still regarding the selection performance of the two methods, our simulations suggest that, in small samples, the irrepresentability index (7) does not have a great influence on the oracle number of selection errors for neither the LASSO nor Forward Stepwise. Asymptotically, however, not even the selection oracle model picked from the LASSO path is model selection consistent for negative values of the irrepresentability index as postulated by theoretical results (38; 1; 40). The models picked from Forward Stepwise by the selection oracle for large samples were less affected by the irrepresentable index especially in the sparser cases. The incoherence conditions used in (35) provide sufficient conditions for the candidates recovered by Forward Stepwise to recover the best subsets, but our results suggest such conditions are overly restrictive.

In terms of prediction, the model error of models picked from the Forward Stepwise and LASSO paths by the prediction oracle performed very similarly. However, when an oracle was not available, the LASSO+AIC_C estimate had a good predictive performance across all settings tested. Such results reproduce for the LASSO, earlier simulation (15) and theoretical (29) findings for the non-negative garrote. They do provide compelling evidence to prefer the LASSO over Forward Stepwise in a reduced sample size situation. In that respect, we identify a minor theoretical gap: do Breiman's theoretical results (29) concerning the stability of the non-negative garrote carry over to the LASSO? Our simulation results seem to suggest so.

Finally, we observe an interesting parallel between the theoretical results for AIC and BIC for the all subsets case and our results. Regardless of the approximation used to obtain a list of candidate models, the AIC_C criterion was the best choice for prediction, whereas gMDL was the best performer for variable selection. Given that AIC_C and gMDL are "closer" to AIC and BIC respectively, it seems plausible that AIC-like (alt. BIC-like) criteria are more suitable for prediction (alt. variable selection) purposes when all subsets are substituted by a list of "approximately" best subsets.

References

[1] P. Zhao and B. Yu, "On model selection consistency of LASSO," Journal of Machine Learning Research, vol. 7, pp. 2541–2563, 2006. [Online]. Available: http://jmlr.csail.mit.edu/papers/volume7/zhao06a/zhao06a.pdf

[2] J. A. Nelder and R. W. M. Wedderburn, "Generalized linear models," Journal of the Royal Statistical Society, Series A, vol. 135, no. 3, pp. 370–384, 1972.

[3] P. McCullagh and J. A. Nelder, Generalized Linear Models. London ; New York: Chapman & Hall, 1989.

[4] J. Rissanen, "Modeling by shortest data description," Automatica, vol. 14, pp. 465–471, 1978.

[5] ——, Stochastic Complexity in Statistical Inquiry, ser. World Scientific Series in Computer Science. Singapore: World Scientific, 1989, vol. 15.

[6] ——, Information and Complexity in Statistical Modeling, ser. Series: Information Science and Statistics. 233 Spring Street, New York, NY 10013, USA: Springer, 2007.

[7] C. S. Wallace and D. M. Boulton, "An information measure for classification," Computer Journal, vol. 11, no. 2, pp. 185–195, 1968. [Online]. Available: http://www.csse.monash.edu.au/~lloyd/tildeMML/Structured/1968-WB-CJ/

[8] C. L. Mallows, "Some comments on C_p," Technometrics, vol. 15, no. 4, pp. 661–675, 1973.

[9] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in 2nd International Symposium on Information Theory, B. N. Petrov and F. Csáki, Eds. Budapest: Akadémia Kiadó, 1973, pp. 267–281.

[10] G. Schwartz, "Estimating the dimension of a model," The Annals of Statistics, vol. 6, pp. 461–464, 1978.

[11] M. Hansen and B. Yu, "Model selection and the principle of minimum description length," Journal of the American Statistical Association, vol. 96, no. 454, pp. 746–774, 2001.

[12] N. Sugiura, "Further analysis of the data by Akaike's Information Criterion and finite corrections," Communications in Statistics, vol. A7, no. 1, pp. 13–26, 1978.

[13] X. Huo and X. S. Ni, "When do stepwise algorithms meet subset selection criteria?" Annals of Statistics, vol. 35, no. 2, pp. 870–887, 2007.

[14] S. Weisberg, Applied Linear Regression. New York: Wiley, 1980.

[15] L. Breiman, "Better subset regression using the nonnegative garrote," Technometrics, vol. 37, no. 4, pp. 373–384, 1995.

[16] R. Tibshirani, "Regression shrinkage and selection via the LASSO," Journal of the Royal Statistical Society, Series B, vol. 58, no. 1, pp. 267–288, 1996.

[17] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," SIAM Review, vol. 43, no. 1, pp. 129–159, 2001.

[18] D. Donoho and I. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," Biometrika, vol. 81, no. 3, pp. 425–455, August 1994.

[19] S. Boyd and L. Vandenberghe, Convex Optimization. Cambridge, UK ; New York: Cambridge University Press, 2004.

[20] E. J. Hannan and B. G. Quinn, "The determination of the order of an autoregression," Journal of the Royal Statistical Society, Series B, vol. 41, no. 2, pp. 190–195, 1979.

[21] R. Nishii, "Asymptotic properties of criteria for selection of variables in multiple regression," The Annals of Statistics, vol. 12, no. 2, pp. 758–765, 1984.

[22] R. Shibata, "An optimal selection of regression variables," Biometrika, vol. 68, no. 1, pp. 45–54, 1981.

[23] ——, "Asymptotic mean efficiency of a selection of regression variables," Annals of the Institute of Statistical Mathematics, vol. 35, no. 3, pp. 415–423, 1983.

[24] K.-C. Li, "Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set," Annals of Statistics, vol. 15, no. 3, pp. 958–975, 1987.

[25] B. T. Polyak and A. Tsybakov, "Asymptotic optimality of the C_p-test for the orthogonal series estimation of regression," Theory of Probability and its Applications, vol. 35, no. 2, pp. 293–, 1991.

[26] J. Shao, "An asymptotic theory for linear model selection (with discussions)," Statistica Sinica, pp. 221–264, 1997.

[27] Y. Yang and A. Barron, "Information theoretic determination of minimax rates of convergence," The Annals of Statistics, vol. 27, no. 5, pp. 1564–1599, 1999.

[28] Y. Yang, "Can the strengths of AIC and BIC be shared?" Biometrika, vol. 101, pp. 937–950, 2003.

[29] L. Breiman, "Heuristics of instability and stabilization in model selection," The Annals of Statistics, vol. 24, no. 6, pp. 2350–2383, 1996.

[30] D. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization," Proceedings of the National Academy of Sciences, vol. 100, no. 5, pp. 2197–2202, 2003.

[31] M. Elad and A. M. Bruckstein, "A generalized uncertainty principle and sparse representation in pairs of bases," IEEE Transactions on Information Theory, vol. 48, no. 9, p. 2558, 2002.

[32] R. Gribonval and M. Nielsen, "Sparse representation in unions of bases," IEEE Transactions on Information Theory, vol. 49, no. 12, p. 3320, December 2003.

[33] D. L. Donoho, "For most large underdetermined systems of equations, the minimal ℓ_1-norm near-solution approximates the sparsest near-solution," 2004, from the author's website. [Online]. Available: http://www-stat.stanford.edu/~donoho/Reports/2004/l1l0approx.pdf

[34] J.-J. Fuchs, "On sparse representations in arbitrary redundant bases," IEEE Transactions on Information Theory, vol. 50, no. 6, p. 1341, June 2004.

[35] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," IEEE Transactions on Information Theory, vol. 50, no. 10, pp. 2231 – 2242, October 2004.

[36] ——, "Recovery of short, complex linear combinations via ℓ_1-minimization," IEEE Transactions on Information Theory, vol. 51, no. 4, p. 1568, 2005.

[37] J. A. Tropp and A. C. Gilbert, "Signal recovery from partial information via orthogonal matching pursuit," IEEE Transactions on Information Theory, 2007.

[38] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the lasso," The Annals of Statistics, vol. 34, no. 3, pp. 1436–1462, 2006.

[39] M. Wainwright, "Sharp thresholds for high-dimensional and noisy recovery of sparsity," Department of Statistics, UC Berkeley, Tech. Rep., 2006. [Online]. Available: http://www.stat.berkeley.edu/tech-reports/709.pdf

[40] H. Zou, "The adaptive LASSO and its oracle properties," Journal of the American Statistical Association, vol. 101, pp. 1418–1429, 2006. [Online]. Available: http://www.stat.umn.edu/~zouhui/pub.htm

[41] E. Candes and T. Tao, "The Danzig Selector: Statistical estimation when p is much larger than n," The Annals of Statistics, 2007.

[42] D. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," IEEE Transactions on Information Theory, vol. 47, no. 7, p. 2845, 2001.

[43] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in Machine Learning: Proceedings of the Thirteenth International Conference, W. W. Cohen and A. Moore, Eds. Pittsburgh, Pennsylvania, USA: Morgan Kaufmann, 1996, pp. 148–156.

[44] P. Bühlmann, "Boosting for high dimensional linear models," The Annals of Statistics, vol. 34, no. 2, pp. 559–583, 2006.

[45] D. M. Allen, "The relationship between variable selection and data augmentation and a method for prediction," Technometrics, vol. 16, pp. 125–127, 1974.

[46] M. Stone, "Cross-validation choice and assessment of statistical predictions," Journal of the Royal Statistical Society, Series B, vol. 36, pp. 111–147, 1974.

[47] S. Geisser, "The predictive sample reuse method with applications," Journal of the American Statistical Association,, vol. 70, pp. 320–328, 1975.

[48] R. A. DeVore and V. N. Temlyakov, "Some remarks on greedy algorithms," Advances in Computational Mathematics, vol. 5, pp. 173–187, December 1996.

[49] V. N. Temlyakov, "Weak greedy algorithms," Advances in Computational Mathematics, vol. 12, pp. 213–227, 2000.

[50] M. Osborne, B. Presnell, and B. A. Turlach, "On the LASSO and its dual," Journal of Computational and Graphical Statistics, vol. 9, no. 2, pp. 319–337, June 2000. [Online]. Available: http://citeseer.ist.psu.edu/osborne99lasso.html

[51] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," The Annals of Statistics, vol. 35, pp. 407–499, 2004.

[52] P. Bühlmann and B. Yu, "Sparse boosting," Journal of Machine Learning Research, vol. 7, pp. 1001–1024, 2006.

[53] C. M. Hurvich, J. S. Simonoff, and C.-L. Tsai, "Smoothing parameter selection in nonparametric regression using an improved akaike information criterion," Journal of the Royal Statistical Society. Series B (Statistical Methodology), vol. 60, no. 2, pp. 271–293, 1998. [Online]. Available: http://links.jstor.org/sici?sici=1369-7412%281998%2960%3A2%3C271%3ASPSINR%3E2.0.CO%3B2-6

[54] P. Zhao, G. Rocha, and B. Yu, "Grouped and hierarchical model selection through composite absolute penalties," Department of Statistics, UC Berkeley, Tech. Rep. 703, 2006. [Online]. Available: http://www.stat.berkeley.edu/users/gvrocha/papers/703.pdf

[55] H. Zou, T. Hastie, and R. Tibshirani, "On the "degrees of freedom" of the LASSO," Stanford University Department of Statistics, Tech. Rep., 2004. [Online]. Available: http://www-stat.stanford.edu/~hastie/Papers/dflasso.pdf

Experimental design and model selection: The example of exoplanet detection

Vijay Balasubramanian[1,2], Klaus Larjo[1] and Ravi Sheth[1*]

[1]*David Rittenhouse Laboratories, University of Pennsylvania, Philadelphia, PA 19104, USA*

[2]*School of Natural Sciences, Institute for Advanced Study, Princeton, NJ 08540, USA*

Abstract

We apply the Minimum Description Length model selection approach to the detection of extra-solar planets, and use this example to show how specification of the experimental design affects the prior distribution on the model parameter space and hence the posterior likelihood which, in turn, determines which model is regarded as most 'correct'. Our analysis shows how conditioning on the experimental design can render a non-compact parameter space effectively compact, so that the MDL model selection problem becomes well-defined.

[*]vijay@physics.upenn.edu, klarjo@physics.upenn.edu, shethrk@physics.upenn.edu

1 Introduction

The Bayesian approach to parametric model selection requires the specification of a prior probability distribution over the parameter space. The Jeffreys' prior, which is proportional to the square root of the determinant of the Fisher information computed in the parameter space, has been shown to be the uniform prior over all *distributions* indexed by the parameters in a parametric family [1]. Geometrically, its integral over a region of the parameter space computes a volume that essentially measures the fraction of statistically distinguishable probability distributions within that region [1]. In this interpretation, the Jeffreys prior distribution

$$\omega(\Theta) = \frac{\sqrt{\det\, J_{ij}(\Theta)}}{\int d^d\Theta \sqrt{\det\, J_{ij}(\Theta)}} d^d\Theta \tag{1}$$

where $\Theta = \{\theta_1, \cdots, \theta_d\}$ simply measures the fractional volume of the small element $d^d\Theta$ relative to total volume of the parametric manifold $V = \int d^d\Theta \sqrt{\det J_{ij}(\Theta)}$. Here J_{ij} is the Fisher information on the parameter space $\Theta \in \mathbb{R}^d$ and $d^d\Theta$ is the standard Riemannian volume element on $\mathbb{R}^d$. The volume V also appears in the Minimum Description Length (MDL) approach to model selection [2, 3], conceptually because it effectively measures how many different distributions are describable by different parameter choices.

An important difficulty in applying the MDL approach to model selection occurs when the parameter space is noncompact and the volume V diverges. In this case, from the Bayesian perspective, a uniform prior on the parameter space does not exist, while from the MDL perspective the number of models that might be describable diverges, leading to problems with the definition of the description length. Of course the parameter space can be cut off by hand, but unless the choice of cut-off is well founded, it can lead to artifacts in the comparison of different model families [4, 5, 6]. Unfortunately in many practical problems the parameter space *is* noncompact and V diverges. For example, in astrophysics, the detection of exoplanets depends on a model of the light coming from the occluded star. This model will contain a non-compact direction representing the orbital period of the planet – see, e.g., [7]. For examples from psychophysics see, e.g., [4].

In this note we argue that merely specifying the experimental set-up – before the measurement of any actual data – influences the prior distribution on the parameter space. This occurs because, given the finite number of measurements in any experiment, many of the probability distributions indexed by a parametric manifold will be statistically indistinguishable. In cases where the parameter space is noncompact, the uniform prior conditioned on the experimental setup can thus become well-defined. In the geometric language of [1], the volume that measures the number of probability distributions in the parametric family that are statistically distinguishable given a *finite* number of measurements can be finite even if the parameter space is non-compact. In effect, specifying the experimental set-up can render the parameter space compact.

Our results illustrate how the choice of experimental set-up influences the measure on the parameter space of a model, thereby affecting which model is regarded as most 'correct'. In

section 2 we briefly review the computation of posterior probabilities, and consider the effect of conditioning on the experimental set-up on the parameter space measure. In section 3 we apply these considerations to a physical problem: the analysis of light-curves of stars with orbiting planets. In this example we see that the volume of the parameter space is rendered effectively finite after the experimental set-up is specified.

2 The effect of experimental design on the parameter space measure

2.1 Review

Suppose one is interested in some physical phenomenon, and has made N relevant measurements: $Y = \{y_1, \ldots, y_N\}$. Further suppose that there are two different parametric models, A and B, that aim to describe the phenomenon in question. The basic question to be answered is which of the two models is the better one, considering the experimental data Y. The probability-theoretic answer to this question is to compute the posterior probabilities $P(A|Y)$ and $P(B|Y)$, which we can write using the Bayes Rule as

$$P(A|Y) = \frac{P(A)}{P(Y)} \int \omega(\Theta) P(Y|\Theta), \tag{2}$$

where $\Theta = (\theta_1, \ldots, \theta_d) \in \mathbb{R}^d$ is the vector of variables parametrising A, and $\omega(\Theta)$ is the volume form associated to the measure on the parameter space, which we will define shortly. A corresponding expression can also be written for $P(B|Y)$. Since we wish to compare $P(A|Y)$ and $P(B|Y)$, we can ignore the common factor $P(Y)$, and we will assume $P(A) = P(B)$ and drop this factor as well. Thus the only remaining ingredient to be defined is the volume form $\omega(\Theta)$; we simply quote the result from [1]: the volume form that gives equal weight to all statistically distinguishable distributions in the parametric family is

$$\omega(\Theta) = \frac{\sqrt{\det\ J_{ij}(\Theta)}}{\int d^d\Theta \sqrt{\det\ J_{ij}(\Theta)}} d^d\Theta, \tag{3}$$

where $J_{ij}(\Theta)$ is the *Fisher information matrix*, defined as the second derivative of the Kullback–Leibler distance $D(\Theta_p||\Theta_q)$:

$$J_{ij}(\Theta_p) = \partial_{\theta_i}\partial_{\theta_j} D(\Theta_p||\Theta_p + \Phi)|_{\Phi=0}, \tag{4}$$

$$D(\Theta_p||\Theta_q) = \int d\vec{x}\ \Theta_p(\vec{x}) \ln \frac{\Theta_p(\vec{x})}{\Theta_q(\vec{x})}. \tag{5}$$

where $d\vec{x}$ is the integration measure over the sample space $\{\vec{x}\}$, and $\Theta_p(\vec{x})$ is the distribution function associated to the values of the parameters $(\theta_1^p, \ldots, \theta_d^p)$. Now we have defined everything needed to compute the posterior probabilities, and we illustrate the formalism by applying it to the analysis of light-curves.

Using this, we can compute the Fisher information matrix by computing the Kullback–Leibler distance between two nearby points and Taylor expanding:

$$\begin{aligned} D(\Theta_0||\Theta_q) &= \int d^N\vec{y}\ \Theta_0(\vec{y})\ \ln\frac{\Theta_0(\vec{y})}{\Theta_q(\vec{y})} \\ &\approx -\int d^N\vec{y}\ \Theta_0(\vec{y})\ln\frac{\Theta_0(\vec{y})+\partial_{\theta_i}\Theta_0(\vec{y})\Delta\theta_i+\partial_{\theta_i}\partial_{\theta_j}\Theta_0(\vec{y})\Delta\theta_i\Delta\theta_j}{\Theta_0(\vec{y})} \\ &\approx -\int d^N\vec{y}\left(\partial_{\theta_i}\Theta_0(\vec{y})\Delta\theta_i+\partial_{\theta_i}\partial_{\theta_j}\Theta_0(\vec{y})\Delta\theta_i\Delta\theta_j-\frac{1}{2}\frac{(\partial_{\theta_i}\Theta_0(\vec{y}))(\partial_{\theta_j}\Theta_0(\vec{y}))}{\Theta_0(\vec{y})}\Delta\theta_i\Delta\theta_j\right) \\ &= \frac{1}{2}\underbrace{\int d^N\vec{y}\frac{(\partial_{\theta_i}\Theta_0(\vec{y}))(\partial_{\theta_j}\Theta_0(\vec{y}))}{\Theta_0(\vec{y})}}_{\equiv J_{ij}(\Theta_0)}\Delta\theta_i\Delta\theta_j. \end{aligned} \tag{6}$$

On the third line, the terms linear in Θ_0 vanish, as exchanging the order of integration and derivation, the integral of Θ_0 will yield a constant 1, which then differentiates to zero.

2.2 Effect of the experimental set-up

The measure (3) is independent of the experimental data Y and is constructed under the assumption that the entire sample space can be measured by the observer. However, in real experiments, instrumental and design limitations only allow observation of some subset M of the sample space. Thus an observation either results in no detected outcome, or in a measurement $y_i \in M$. Thus the effective predicted distribution of measured outcomes is not the $\Theta(\vec{y})$, but rather

$$\Theta(\vec{y}) = \begin{cases} \Theta(\vec{y}), & \text{for } \vec{y}\in M, \\ \Theta^{\text{Out}}, & \text{no measured outcome,} \end{cases} \tag{7}$$

where $\Theta^{\text{Out}} \equiv \int_{\vec{y}\notin M} d\vec{y}\ \Theta(\vec{y})$. We will argue that if the models in the asymptotic regions of a noncompact parameter space differ in their predictions mostly outside the observable region M, the Fisher information for the effective distributions (7) can decay sufficiently quickly to render the volume $V = \int d^d\Theta\ \sqrt{\det J_{ij}(\Theta)}$ finite. In this section we will give one set of sufficient conditions for this to happen and in Sec. 3 we will give a detailed example.

Consider a model, specified by parameters $\vec{\theta} = (\theta_1, \ldots, \theta_d) \in \mathbb{R}^d$, and a distribution $\Theta_{\vec{\theta}}(\vec{x})$, with $\vec{y} \in \mathbb{R}^n$. We will slightly simplify notation simply referring to the distribution as $\Theta(\vec{y})$ and understanding the implicit parameter dependence. Let us use spherical coordinates in the parameter space $\mathbb{R}^d$ with ρ being the radial coordinate, i.e. $(\theta_1, \ldots, \theta_d) \to (\rho, \varphi_1, \ldots, \varphi_{d-1})$. Also consider an experimental set-up that can only make measurements inside some compact region $M \subset \mathbb{R}^n$. Thus,the probability of no measurement being registered by this experiment is $\Theta^{\text{Out}} \equiv \int_{\vec{y}\notin M} d\vec{y}\ \Theta(\vec{y})$.

Our first assumption is a smoothness condition, so that inside the region M the distribution does not fluctuate too much as one approaches the asymptotics of parameter space:

$$\left|\partial_i\Theta(\vec{y})|_{\vec{y}\in M}\right| \le \delta(\rho), \quad \text{for large } \rho,\ i = 1, \ldots, d, \tag{8}$$

where $\delta(\rho)$ goes to zero as ρ goes to infinity; we will later specify the exact scaling needed. Intuitively, this condition says that as the parameter $\rho \to \infty$, the models do not differ too much

inside the observable part of the sample space M. This allows us to estimate

$$\left|\partial_i \Theta^{\text{Out}}\right| = \left|\partial_i (1 - \int_{y \in M} d\vec{y}\ \Theta(\vec{y}))\right| \leq \text{Vol}(M)\delta, \tag{9}$$

where $\text{Vol}(M)$ denotes the volume of the compact region M.

Secondly we assume that inside M, the distributions $\Theta(\vec{y})$ do not decay too quickly as $\rho \to \infty$. Intutitively, since any experiment will only measure a finite amount of data (say N points), if the probability of a single measurement lying inside M is significantly less than $1/N$, then the experimental set-up will not detect anything. Thus we will require

$$\Theta(\vec{y})\mid_{\vec{y} \in M} > \epsilon(\rho), \quad \text{for large } \rho, \tag{10}$$

where again we will later specify the scaling of $\epsilon(\rho)$ with ρ.[1]

Using these assumptions, we can establish an upper bound for the Fisher information (6):

$$\begin{aligned} |J_{ij}| &\leq \left|\int_{\vec{y} \in M} \frac{\partial_i \Theta(\vec{y}) \partial_j \Theta(\vec{y})}{\Theta(\vec{y})}\right| + \left|\frac{\partial_i \Theta^{\text{Out}} \partial_j \Theta^{\text{Out}}}{\Theta^{\text{Out}}}\right| \\ &< \delta^2 \left|\int_{\vec{y} \in M} \frac{1}{\Theta(\vec{y})}\right| + \text{Vol}(M)^2\delta^2 \leq \text{Vol}(M)\frac{\delta^2}{\epsilon} + \text{Vol}(M)^2\delta^2 \sim \text{Vol}(M)\frac{\delta^2}{\epsilon}. \end{aligned} \tag{11}$$

Thus the determinant of the Fisher information scales as

$$\sqrt{\text{Det } J_{ij}} \sim \left(\frac{\delta^2}{\epsilon}\right)^{\frac{d}{2}}, \tag{12}$$

and for the integral V to be finite one must have suppression stronger than $\sqrt{\text{Det } J_{ij}} \sim \rho^{-d}$. Thus the integral converges if δ is suppressed more strongly than

$$\delta(\rho) < \frac{\sqrt{\epsilon(\rho)}}{\rho}. \tag{13}$$

From the experimental set-up one can estimate how $\epsilon(\rho)$ scales with ρ, which then determines how $\delta(\rho)$ needs to scale for the integral to converge. This is thus a sufficient condition for rendering the parameter space effectively finite.

It is worth stressing that, following the above analysis, any method of deciding the validity of a model is impacted by the choice of the experiment in a completely computable way, and this should be taken into account when designing experiments.

3 The probability of exo-planet detection

3.1 Model for exo-planets

Consider a star orbited by a planet so that the planet periodically passes between the star and Earth. The light output (light-curve) of such a star is a constant line, with a small periodic dip

[1] This condition can be relaxed by recognizing that if $\Theta(\vec{y})|_{\vec{y} \in M}$ decays too quickly as $\rho \to \infty$, then the models in the asymptotic region of the parameter space make no measurable predictions for experiments designed with a finite number of measurements. The example in the Sec. 3 will illustrate such a scenario.

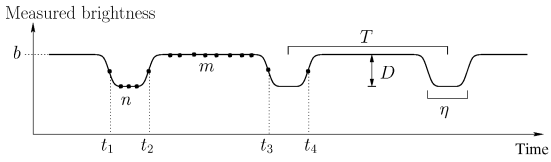


Figure 1: An example of a light-curve.

when the planet is eclipsing part of the star. One model for such a light-curve was proposed in [7] as

$$y(T, D, \eta, \tau, b; t) = b - \frac{D}{2}\left[\tanh c(\tilde{t} + \frac{1}{2}) - \tanh c(\tilde{t} - \frac{1}{2})\right], \tag{14}$$

where

$$\tilde{t} = \frac{T \sin \frac{\pi(t-\tau)}{T}}{\pi\eta}. \tag{15}$$

An example light-curve is shown in figure 1; T is the period of the planet; η is the duration of the transit, i.e. how long the planet eclipses the star; D is the depth of the dip in the curve; b is the total observed brightness of the star; and τ is a phase parameter specifying when the planets transit occurs. Finally, c is a constant parameter specifying the sharpness of the edges of the light-curve, expected to be fairly large as the transition between transit/no-transit is relatively quick. The assumption $c \gg 1$ greatly simplifies our analysis, and is not physically very restrictive.

The parameter space for this model is clearly non-compact as T can range to infinity. However, we will argue that the space is effectively rendered compact after the experimental set-up is specified. To be precise, the parameter space is[2]:

$$T \in [0, \infty), \quad D \in [0, b], \quad \tau \in [0, T], \quad \eta \in [0, \delta T], \quad b \in [0, b_{max}], \tag{16}$$

where δ is a small number that we will estimate, and the maximal brightness b_{max} is naturally given by the brightness of Sirius, the brightest star visible from Earth. Assuming a circular orbit as in Figure 2, the ratio of the transit time to the period of the planet is given by

$$\frac{\eta}{T} \approx \frac{2r/v_{\text{planet}}}{2\pi R/v_{\text{planet}}} = \frac{1}{\pi}\frac{r}{R}.$$

For the currently known transiting exo-planets this ratio is around ~ 0.1 [8], although for a typical system one expects it to be smaller as large planets orbiting close to the star are easier to observe, which favors largest values of the ratio. For an elliptical orbit, the answer will differ by an $\mathcal{O}(1)$ factor, but will have the same dependence on r/R. Thus, η will always be a small fraction of T.

Now we can write down the probability density for measuring values $\vec{y} = (y_1, \ldots, y_N)$ for the light-curve at times $(t_1, \ldots, t_N)$ with the light-curve specified by parameters

[2]Note that we consider c to be a constant, not a parameter.

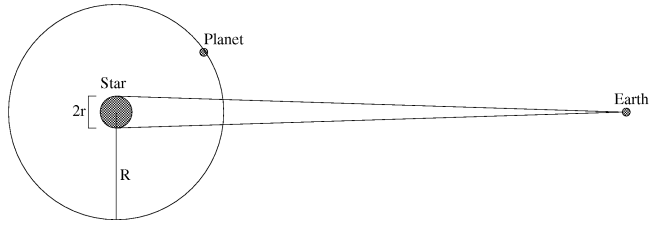


Figure 2: The basic set-up: an extra-solar planet orbiting a star of radius r with an average distance R.

$(\theta_1^0, \theta_2^0, \theta_3^0, \theta_4^0, \theta_5^0) = (T, D, \eta, \tau, b)$ as

$$\Theta_0(\vec{y}) = \prod_{k=1}^{N} \frac{1}{\sqrt{2\pi\sigma_k}} e^{-\frac{(y_k - y_0(\theta_i^0; t_k))^2}{2\sigma_k^2}} = (2\pi\sigma)^{-\frac{N}{2}} e^{-\frac{1}{2\sigma^2}\sum_{k=1}^{N}(y_k - y_0(\theta_i^0; t_k))^2}, \tag{17}$$

where we have assumed that the uncertainty in each measurement is Gaussian, and further we have chosen the standard deviation to be equal for all measurements for simplicity. Using (17) in the formula (6), we see that the the integrals in the Fisher information are Gaussian in y_k; thus we can compute them analytically to get

$$J_{ij} = \frac{1}{\sigma^2} \sum_{k=1}^{N} \partial_{\theta_i} y(\theta; t_k) \partial_{\theta_j} y(\theta; t_k). \tag{18}$$

This is our key formula, and we shall spend the next subsection analysing its properties.

3.2 Finiteness of light-curve parameter space

We now wish to apply the general arguments of section 2 to the exo-planet system. Consider an experimental set-up that can barely measure two periods, and then consider shortening the experiment slightly so that only one dip is detected; this is depicted in figure 1. To be precise, the shorter set-up measures the beginning and end of a transit at t_1 and t_2, n points in between, and m points after the transit. The longer set-up makes measurements at the same times, and additionally at times t_3 and t_4, detecting the second transit. In the next subsections we will show that $J_{\text{short}} \lll J_{\text{long}}$, indicating that detecting the second dip is of fundamental importance to experimental design; without the second dip the experimental set-up can't differentiate models with large enough T. This renders the parameter space effectively finite, as an experiment can not differentiate between models that have period T larger than the duration of the experiment.

3.2.1 Effect of measuring a second transit on $\det(J)$

In this subsection we will give an estimate for the magnitude of the determinant of the Fisher information, and show how it is affected by the inclusion of the second transit in the data. In subsequent subsections we will exactly compute the determinant for a few specific experimental set-ups.

From (18) and the definition of a determinant, we see that in each term of the determinant each parameter θ_i appears exactly twice in the derivatives, i.e. each term is of the form

$$J_{i_1 j_1} J_{i_2 j_2} J_{i_3 j_3} J_{i_4 j_4} J_{i_5 j_5} \sim \frac{1}{\sigma^{10}} \partial_T y \; \partial_T y \; \partial_D y \; \partial_D y \; \partial_\eta y \; \partial_\eta y \; \partial_\tau y \; \partial_\tau y \; \partial_b y \; \partial_b y. \tag{19}$$

As a rough estimate of the size the determinant, we investigate how large terms of this type can be. The derivatives are

$$\partial_T y(\theta, t) = \frac{cD}{2} \frac{f(\tilde{t})}{\pi\eta} \left(\sin \frac{\pi(t-\tau)}{T} - \frac{\pi(t-\tau)}{T} \cos \frac{\pi(t-\tau)}{T} \right), \tag{20}$$

$$\partial_D y(\theta, t) = \frac{1}{2} \left(\tanh c(\tilde{t} - \frac{1}{2}) - \tanh c(\tilde{t} + \frac{1}{2}) \right), \tag{21}$$

$$\partial_\tau y(\theta, t) = -\frac{cD}{2} \frac{f(\tilde{t})}{\eta} \cos \frac{\pi(t-\tau)}{T}, \tag{22}$$

$$\partial_\eta y(\theta, t) = -\frac{cD}{2} \frac{\tilde{t} f(\tilde{t})}{\eta}, \qquad \partial_b y(\theta, t) = 1, \tag{23}$$

$$\text{with} \qquad f(\tilde{t}) \equiv \tanh^2 c(\tilde{t} + \frac{1}{2}) - \tanh^2 c(\tilde{t} - \frac{1}{2}). \tag{24}$$

From (24) we see that $f(\tilde{t}) \neq 0$ only when $\tilde{t} \approx \pm\frac{1}{2}$, again assuming large c. This tells us that the measurements that contribute most to the Fisher information are the ones on the edges of the dips[3] , i.e. at times t_1, t_2, t_3 and t_4 in figure 1. We write the condition $|\tilde{t}| \approx \frac{1}{2}$ as

$$\left| \sin \frac{\pi(t-\tau)}{T} \right| = \frac{\pi}{2} \frac{\eta}{T}, \tag{25}$$

and note that the ratio of transit time to period is very small, $\frac{\eta}{T} \ll 1$. This gives us the solutions

$$\frac{t-\tau}{T} \approx n \pm \frac{\eta}{2T}, \tag{26}$$

where n is an integer indexing the number of the dip, with $n = 0$ denoting the solitary dip if only one is present in the data.

We wish to estimate the ratio of the determinants of the Fisher information by an order of magnitude estimate

$$\frac{J_{\text{short}}}{J_{\text{long}}} \sim \frac{J^s_{i_1 j_1} J^s_{i_2 j_2} J^s_{i_3 j_3} J^s_{i_4 j_4} J^s_{i_5 j_5} \; |_{\max}}{J^l_{i_1 j_1} J^l_{i_2 j_2} J^l_{i_3 j_3} J^l_{i_4 j_4} J^l_{i_5 j_5} \; |_{\max}}, \tag{27}$$

where both the numerator and the denominator are of the form (19), and according to the argument above the maximal contributions come from the edge measurements. From (21-23) we see that the derivatives with respect to D, η, τ and b are all periodic at the edges: $|\partial_{\theta_i} y(\theta, t_1)| = \ldots = |\partial_{\theta_i} y(\theta, t_4)|$ for $\theta_i \neq T$, and thus will cancel in the ratio (27).

It is crucial that $\partial_T y$, however, is not periodic due to the second term in (20). At the first dip, $t_1, t_2 = \pm\frac{\eta}{2T}$, we expand (20) to find

$$\partial_T y(t_1) \approx \partial_T y(t_2) \approx \frac{\pi^2 cD}{48} \frac{\eta^2}{T^3}, \tag{28}$$

[3]This statement is somewhat subtle, and we will discuss this matter in more detail in section 3.2.3; for our current purposes it is sufficiently accurate.

while at the second dip, $t_3, t_4 = 1 \pm \frac{\eta}{2T}$, the contribution is

$$|\partial_T y(t_3)| \approx |\partial_T y(t_4)| \approx \frac{cD}{2\eta}, \tag{29}$$

ignoring signs that are irrelevant for this estimate. Thus we see that the Fisher information increases strongly as the second dip is included:

$$\frac{J_{\text{short}}}{J_{\text{long}}} \sim \frac{(\partial_T y(t_{1,2}))^2}{(\partial_T y(t_{1,2}))^2 + (\partial_T y(t_{1,2})\partial_T y(t_{3,4})) + (\partial_T y(t_{3,4}))^2} \sim \left(\frac{\eta}{T}\right)^6 \lll 1, \tag{30}$$

where we ignored order one coefficients. This is an explicit example of how our arguments from section 2 work for a realistic model: when an experimental set-up does not have the capability to detect two dips, it becomes impossible to determine the period, and consequently the Fisher information is very small (or vanishing) compared to an experiment that is able to detect two dips and determine the period more accurately. For any given experiment of finite duration Δt, the Fisher Information will decline with T when $T \gg \Delta t$ effectively rendering the parameter space compact.

3.2.2 The tail $T \to \infty$

To verify our claim that the parameter space is really rendered compact we need to show that $\det J \to 0$ strongly enough as T is taken to infinity. It is easy enough to find the T-scaling of the derivatives (20-23); $\partial_T y$ scales as T^{-3}, while the others stay finite in the large T limit. Thus, as seen from (19), the determinant will scale as

$$\sqrt{\det J} \sim \partial_T y \sim \frac{1}{T^3}, \tag{31}$$

which shows that that the parameter space measure vanishes fast enough for large T to render the parameter space volume finite.

3.2.3 Explicit computation of Det(J_{ij}) for specific experimental set-ups

While the order of magnitude estimate of the previous subsection offers an intuitive reason as to why the Fisher information decreases sharply when the number of peaks detected falls below two, it is still instructive to explicitly compute the determinant in a few experimental set-ups.

Detecting two dips: Let us first consider the case J_{long} from section 3.2, i.e. measurements at times indicated in figure 1. Using the derivatives (20-23) one can write down the Fisher information matrix (18) as

$$J_{ij}^{\text{long}} = \begin{pmatrix} 2(T_1^2 + T_3^2) & -T_1 & 2T_1X & -4T_3X & 2T_1 \\ -T_1 & 1+n & -2X & 0 & -(2+n) \\ 2T_1X & -2X & 4X^2 & 0 & 4X \\ -4T_3X & 0 & 0 & 16X^2 & 0 \\ 2T_1 & -(2+n) & 4X & 0 & 4+n+m \end{pmatrix}, \tag{32}$$

where for brevity we defined

$$T_1 \equiv \partial_T y(t_1) = \partial_T y(t_2) = \frac{cD\pi^2}{48}\frac{\eta^2}{T^3}, \quad T_3 \equiv \partial_T y(t_3) = -\partial_T y(t_4) = \frac{cD}{2\eta}, \tag{33}$$

$$X \equiv -\frac{cD}{4\eta} = \partial_\eta y(t_{1,2,3,4}) = -\frac{\partial_\tau y(t_{1,3})}{2} = \frac{\partial_\tau y(t_{2,4})}{2}. \tag{34}$$

In computing this matrix we used that $f(\tilde{t}) = 0$ for $\tilde{t} \neq \pm\frac{1}{2}$, which is true up to corrections of order e^{-c}, as seen from (24); for this reason one does not need to specify the exact times of the n measurements during the dip, or the m measurements outside the dip, as up to e^{-c} corrections they all contribute equally. The determinant of the Fisher information is simple,

$$\mathrm{Det}(J_{ij}^{\mathrm{long}}) = 64nmX^4(T_1^2 + T_3^2) \approx 64nmX^4T_3^2. \tag{35}$$

This result explains the subtlety referred to earlier: although measurements at the edges contribute the most to the Fisher information, if one only has measurements at the edges ($n = m = 0$) the Fisher information actually vanishes. Physically this is easy to interpret, as only measuring the edges $t_1, \ldots, t_4$ will yield four points lying on a line, and thus they cannot be used to determine any information about the curve; other data points are needed to 'anchor' the data.

Detecting only one dip: Similarly one can compute the Fisher information in the 'short' experimental set-up, where measurements are made at the same times as before, except not at t_3 and t_4. This yields

$$J_{ij}^{\mathrm{short}} = \begin{pmatrix} 2T_1^2 & -T_1 & 2T_1X & 0 & 2T_1 \\ -T_1 & \frac{1}{2}+n & -X & 0 & -(1+n) \\ 2T_1X & -X & 2X^2 & 0 & 2X \\ 0 & 0 & 0 & 8X^2 & 0 \\ 2T_1 & -(1+n) & 2X & 0 & 2+n+m \end{pmatrix}, \tag{36}$$

and perhaps surprisingly the determinant vanishes: $\mathrm{Det}(J_{ij}^{\mathrm{short}}) = 0$, up to tiny e^{-c} corrections. This indicates that the estimate in section 3.2.1 was an overestimate[4]: terms in the determinant of J^{short} are of the magnitude estimated, but the determinant is arranged in such a way that the terms cancel to a high accuracy, and the compactness of the parameter space is strengthened.

4 Discussion

Our analysis has shown how the specification of an experimental design affects the measure on model parameter spaces in MDL model selection (or equivalently the prior probability distribution on parameters in the Bayesian approach). Interestingly, the finite number of measurements within a bounded sample space in any practical experiment can effectively render a non-compact parameter space compact thereby leading to a well-defined prior distribution (3). Our analysis

[4] As the estimate illustrates an intuitive reason why the appearance of the second peak is so important, we decided to include it.

could be turned around to design experiments to discriminate well between models in some chosen region of the parameter space by ensuring that the Fisher information (18) is large in the desired region. It would also be useful to determine general conditions under which experimental design effectively makes model parameter spaces compact, perhaps following the arguments of Sec. 2.

Acknowledgments: This paper was written in honor of Jorma Rissanen's 75th birthday and his many seminal achievements in statistics and information theory. VB and KL were partially supported by the DOE under grant DE-FG02-95ER40893, and KL was also partly supported by a fellowship from the Academy of Finland. VB was also partly supported as the Helen and Martin Chooljian member at the Institute for Advanced Study.

References

[1] V. Balasubramanian, "Statistical Inference, Occam's Razor and Statistical Mechanics on The Space of Probability Distributions," [arXiv:cond-mat/9601030], V. Balasubramanian, "A Geometric Formulation of Occam's Razor for Inference of Parametric Distributions," [arXiv:adap-org/9601001].

[2] J. Rissanen, "Modeling by shortest data description", Automatica, 14:1080-1100, 1978.

[3] J. Rissanen, "Fisher information and stochastic complexity", IEEE Trans. Inform.Theory, 42:40-47, 1996.

[4] I.J. Myung, V. Balasubramanian and M.A. Pitt, "Counting Probability Distributions: Differential Geometry and Model Selection", Proceedings of the National Academy of Science, 97(21) 11170–11175, 2000.

[5] F. Liang and A. R. Barron, "Exact minimax strategies for predic- tive density estimation, data compression, and model selection", IEEE Transactions on Information Theory 50, 2708-2726, 2004.

[6] Chapter 11 of P.D. Grünwald, *The Minimum Description Length Principle*, MIT Press, June 2007.

[7] P. Protopapas, R. Jimenez and C. Alcock, "Fast identification of transits from light-curves," Mon. Not. Roy. Astron. Soc. **362**, 460 (2005) [arXiv:astro-ph/0502301].

[8] http://obswww.unige.ch/ pont/simpleTABLE.dat

Enumerative Coding for Tree Sources

Álvaro Martín* Gadiel Seroussi † Marcelo Weinberger ‡

To Jorma, our favorite Bayesian, happy birthday!

Abstract

Efficient enumerative coding for tree sources is, in general, surprisingly intricate—a simple uniform encoding of type classes, which is asymptotically optimal in expectation for many classical models such as FSMs, turns out not to be so in this case. We describe an efficiently computable enumerative code that is universal in the class of tree sources in the sense that, for a string emitted by an unknown source supported on a known tree, the expected normalized code length of the encoding approaches the entropy rate of the source with a convergence rate $(K/2)(\log n)/n$, where K is the number of free parameters of the source. Based on recent results characterizing type classes for tree sources, the code consists of the index of the sequence in the tree-type class, and an efficient description of the class itself using a non-uniform encoding of selected symbol and string counts. The results are extended to a twice-universal setting, where the tree underlying the source is unknown, and is estimated from the input sequence.

1 Introduction

Jorma's journey through universal coding for parametric model classes went from the crude, original version of the two–part codes of [1], to the Normalized Maximum Likelihood (NML) codes of [2], which he first analyzed in [3] by establishing a connection between the two methods. The transition from two–part codes to the NML code unveils, in a sense, an obvious reliance on *enumerative coding* methods [4], via the *method of types* [5]. In the method of types, the set of sequences of a given length n over a finite alphabet $\mathcal{A}$ is partitioned into *type classes*, where two sequences belong to the same class if and only if they are assigned the same probability for any choice of the model parameters. Since all sequences in a type class are equiprobable, the universal probability assignment problem can be reduced to optimally assigning probabilities to type classes.

This is indeed the case for the NML code, which can be interpreted as a description of the type, generated by assigning to it a probability proportional to its ML probability, followed by an enumeration of the sequences in the type class. In two–part codes, instead, the ML parameter estimate of the model is quantized, thereby merging type classes, which are assigned a probability using the same quantized parameter estimate. The two–part

*Instituto de Computación, Universidad de la República, Montevideo, Uruguay. Supported by Grant PDT - S/C/IF/63/147. E-mail: almartin@fing.edu.uy

†Hewlett-Packard Laboratories. 1501 Page Mill Road Palo Alto, CA 94304, U.S.A., and Facultad de Ingeniería, Universidad de la República, Montevideo, Uruguay. E-mail: gseroussi@ieee.org

‡Hewlett-Packard Laboratories, Palo Alto, CA 94304, U.S.A. E-mail: marcelo.weinberger@hp.com

code of [1] is inherently incomplete since, once the quantized estimate is described, it is redundant to allocate coding space to all the n–tuples; rather, only the sequences in the corresponding type classes need to be assigned codes. Moreover, this two–part code assigns a uniform probability over the merged type classes. These drawbacks are addressed in [3] via yet another interpretation of the NML code.

Unfortunately, implementing the NML code is difficult even for the simplest model classes. Other universal methods, based, for example, on the Krichevskii-Trofimov sequential probability assignment [6], are computationally efficient, and also assign the same code length to all the sequences of a given type. They do not, however, provide a separate and identifiable description of the type. In this paper, we are interested in *universal enumerative codes* that possess both qualities: they provide a separate description of the type class of the encoded sequence, and this description can be efficiently computed. By "universal" we mean codes whose normalized average length differs from the entropy rate of the source by a term of the form $K\frac{\log n}{2n}$, where K is the number of free statistical parameters of the source, in accordance to Rissanen's lower bound [1, Theorem 1]. By "efficient computation" we mean one whose encoding running time is polynomial in the length of the input sequence, and with code construction time that is also polynomial in the number of free parameters of the source.[1]

For exponential (or asymptotic exponential) families [7, 8] satisfying some mild regularity conditions, most type classes have, to first approximation, the same ML probability [9, Appendix A]. This observation leads to enumerative source codes that are universal (in expectation), and for which uniform coding is used both for the set of type classes and for the set of sequences of each type. In particular, for finite–state machine (FSM) models [10], such a code can be efficiently implemented. Indeed, for an FSM F with a finite set of states S and alphabet of size $|\mathcal{A}| = \alpha$, there are[2] $\Theta\left(n^{|S|(\alpha-1)}\right)$ type classes of sequences of length n [11, attributed to N. Alon]. Thus, a uniform encoding of the type class gives a normalized cost of up to $|S|(\alpha-1)\frac{\log n}{n} + O(\frac{1}{n})$ bits. Moreover, by means of Stirling's formula and bounding the expectation as in [12], one obtains, for the type class of a random sequence, that the expected normalized logarithm of the class size is upper-bounded by $\mathcal{H} - |S|(\alpha-1)\frac{\log n}{2n} + O(\frac{1}{n})$, where $\mathcal{H}$ is the entropy rate of the source. Thus, the term subtracted from $\mathcal{H}$ compensates for half the cost of describing the type class, yielding an overall penalty of $|S|(\alpha-1)\frac{\log n}{2n} + O(1/n)$ bits over $\mathcal{H}$, which is optimal by [1]. Since the description of the type class is not difficult in this case, and efficient methods exist for enumerating the class, the resulting enumerative code satisfies our requirements.

The question arises: Is a similar technique applicable to useful model classes which do not necessarily induce an asymptotic exponential family? In this paper, we address this question for the popular *tree models* [12, 13], which have proven very valuable as modeling tools in data compression and other applications in information theory and statistics (cf. [12,

[1] Our complexity requirement will focus on the description of the type class, as there are known efficient methods to do the enumeration of the class itself for most cases of interest. Notice that, since the number of types in the cases of interest is generally exponential in the number of free parameters of the source, a construction of the NML code relying on the computation of the ML probability of each type would be very inefficient.

[2] We use conventional asymptotic notation: $O(f(n))$ denotes a function $g(n)$ such that $0 \le g(n) \le cf(n)$ for a positive constant c and sufficiently large n, $\Theta(f(n))$ a function $g(n)$ such that $g(n) = O(f(n))$ and $f(n) = O(g(n))$, and $o(f(n))$ a function $g(n)$ such that $\lim_{n\to\infty} g(n)/f(n) = 0$. All logarithms will be to base two. To simplify discussions, we will sometimes ignore fractional parts of code lengths, referring, for example, to $\log n$ bits instead of the more precise "at most $\lceil \log n \rceil$ bits." This loose convention will be immaterial to the main asymptotic results of the paper.

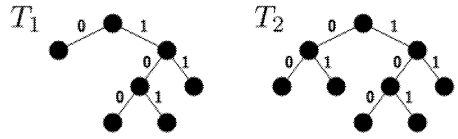


Figure 1: Tree models over $\mathcal{A} = \{0, 1\}$

13, 14, 15, 16]). Tree models offer a compact representation which, in real life applications, can often model finite–memory processes with a significantly smaller number of parameters compared to fully parametrized Markov models. These savings translate to a lower *model cost* [1] and a faster convergence to optimal performance (e.g., compression ratio).

In contrast to the FSM case, for a tree model with a set of states S_T, which in general does not induce an asymptotic exponential family of distributions [8], the number of type classes grows polynomially as n^k, but the exponent k may be larger than $(\alpha - 1)|S_T|$. The asymptotic number of type classes in this case is fully characterized in [17], generalizing the corresponding result for FSMs. A simple example that illustrates this phenomenon is easy to construct: Consider the trees T_1 and T_2 of Figure 1. It can be shown that each type class of T_1 is partitioned into up to a constant number of type classes in T_2. Since the tree T_2 has the "FSM property" (in the sense that the occurrence of a symbol in a state defines the next state), the number of type classes, by the above discussion on FSMs, is $\Theta(n^5)$ for both trees, even though T_1 has four states. Since, as noted, the partitions are essentially the same, lower-bounding the average code length appropriately, we can readily see that an enumerative scheme using a uniform code for the type classes would not exploit the reduction in the number of model parameters of T_1 with respect to T_2. Notice, however, that a type class with significantly different conditional empirical distributions for states 00 and 01 of T_2 will have small probability under the model T_1 for *any* choice of model parameters, which suggests that the savings in code length might be recovered with a *non-uniform* code for the type classes.

In this paper, we construct such non-uniform code, leading to an efficient enumerative coding scheme which is universal (in expectation) for tree sources. While in the context of the two–part code of [1], the use of non-uniform codes as in [3] has only a third order effect (not affecting the rate of convergence to the entropy), here the non-uniformity is crucial to achieve second–order optimality as defined by [1]. Furthermore, in the twice–universal setting, in which a tree is not given and optimality is rather required for *any* possible tree, we show that, by suitably estimating a tree from the data, the sequences in the aforementioned "atypical" type classes for each given tree would in fact estimate a different tree. These type classes can thus be discarded from the coding space, leading to a twice–universal enumerative code for sequences in the class of tree models.

Our implementation of the second part of the enumerative code, namely, the index of x^n in its type class, will be based on the enumeration of tree type classes from [17]. As for the first part, namely, the description of the type class, a key building block in our scheme will be a collection of codes for encoding counts of occurrences of certain patterns within the input sequence. In the example of Figure 1, given the empirical conditional distribution, $\hat{p}$, in state 0 of T_1, and the number of occurrences, n_s, of the pattern 00, we can estimate the number of occurrences, $n_s^{(a)}$, of symbol a in state 00 of T_2 as $\hat{n}_s^{(a)} = n_s \hat{p}(a)$. If n_s and $\hat{p}$ have already been described to the decoder, we can then encode the difference $n_s^{(a)} - \hat{n}_s^{(a)}$ by assigning high probability to small absolute differences. This observation will be generalized

to encode, efficiently, a collection of pattern counts that uniquely determines the type class of a sequence.

The rest of the paper is organized as follows. After introducing our notation and formal setting in Section 2, in Section 3 we introduce variable length codes for symbol counts, following the observation above. These codes, which are based on Golomb codes [18] and dubbed *symbol count codes* (*SCCs*), will help us obtain, in Section 4, a precise asymptotic estimate for the size of the type class of a sequence with respect to a given tree. This estimate complements the exact combinatorial characterization of the type class in [17]. In Section 5 we generalize the construction of SCCs to codes for pattern counts, dubbed *string count codes* (*SCCs**), which yield an efficient description of the type class and, combined with the result of Section 4, lead to a universal enumerative code. The SCCs* also yield an alternative construction of enumerative codes, taken with respect to an extension of the original tree T, and where the increment in the cost of describing the larger model is compensated exactly by the reduction in the expected size of the class, maintaining an expected code length that is still optimal with respect to T. In particular, such a code can be derived from the *FSM closure* [15] of T using known techniques for enumerating sequences in FSM type classes. Finally, in Section 6 we present two approaches for the twice-universal setting. The first approach is a standard plug-in scheme where the tree is first estimated, and then the previously derived universal enumerative code is applied as-is, using the estimated tree in lieu of the true one. In the second approach, we take advantage of the observation above that for any given tree, there will be sequences that are "atypical" for the tree, and will not estimate it regardless of the source parameters. Thus, the enumerative code is significantly simplified in the twice-universal setting by excluding such sequences from the coding space for the estimated tree.

2 Preliminaries

Let $\mathcal{A}$ be an alphabet of $\alpha \geq 2$ symbols, and let λ denote the empty string. We denote by u_j^k the string $u_j u_{j+1} \dots u_k$ over $\mathcal{A}$, with $u_j^k = \lambda$ when $j > k$, and we omit the subscript when $j = 1$. For a string $u = u^k$, we let $\overline{u} = u_k u_{k-1} \dots u_1$ denote its reverse, $|u| = k$ its length, and we write $\mathrm{hd}(u) = u_1$, and $\mathrm{tail}(u) = u_2^k$; $|S|$ also denotes the cardinality of a set S. Concatenation of u and v is denoted uv, and $u \preceq v$ (resp. $u \prec v$) denotes the prefix (resp. proper prefix) relation.

Our models will be based on full α-ary trees[3] (or simply, *trees*), in which each edge is labeled with a symbol in $\mathcal{A}$, and each node with the string formed by concatenating the edge labels on the path from the root (labeled by λ) to the node. We identify a tree T with its set of nodes, and each node with its label, e.g., $u \in T$ indicates that there is a node of T labeled u. Leaves of T are called *states*, and the set of states is denoted S_T. For a sufficiently long sequence x^n, we refer to the (unique) prefix of $\overline{x^n}$ in S_T, denoted $\sigma(x^n)$, as the state *selected* by x^n (the dependence of σ on T will be assumed from the context). For the purpose of selecting states, we assume that x^n is preceded by an arbitrary *fixed* semi-infinite string $x_{-\infty}^0$. This convention uniquely determines, for any given tree, an *initial state* s_0 "selected" by λ, and guarantees that any (short) sequence selects a state. Thus, x^n uniquely determines a *state sequence* $\sigma(\lambda){=}s_0, \sigma(x^1), \sigma(x^2), \dots, \sigma(x^n)$, with $\sigma(x^n)$ referred to as the *final state* of x^n with respect to T. If $\sigma(x^i) = s$, we say that the symbol x_{i+1} *occurs in state* s. The notion of occurrence is extended to arbitrary strings, namely, if for

[3] A α-ary tree is full if and only if every internal node has exactly α children.

$u = u^k$ and some index i, $0 \le i < n$, we have $x_{i-k+1}^i = u$, we say that x_{i+1} *occurs in context* u in x^n (notice that if x_{i+1} occurs state s, then it occurs in context $\overline{s}$).

Trees do not necessarily define a *next-state* function, since the occurrence of a symbol in a state does not necessarily determine the following state. In the tree T_1 of Figure 1, for example, the occurrence of symbol 1 in state 0 does not determine whether the next state would be 100 or 101. When a next state function exists for a tree T we say that T *is FSM*. It is shown in [15] that T is FSM if and only if every suffix of a state of T belongs to T. A tree T' is a *refinement* of T if $T \subseteq T'$. If a state s of T is an internal node of a refinement T', we say that T' refines s. The tree T_F obtained from T by adding all the suffixes of states of T is called its *FSM closure*; T_F is the minimal FSM refinement of T [15]. In Figure 1, T_2 is the FSM closure of T_1.

A *tree source* is determined by a tree T and a set of $|S_T|$ conditional probability distributions $\mathrm{P}_T(\,\cdot\,|\,s\,)$, $s \in S_T$, over $\mathcal{A}$. This induces, for each length $n \ge 0$, a natural probability assignment for a random variable X^n over $\mathcal{A}^n$ ("emitted" by the source), given by

$$\mathrm{P}_T(X^n = x^n) = \prod_{i=1}^{n} \mathrm{P}_T\left(x_i|\sigma(x^{i-1})\right), \quad n \ge 0 \quad (\,\mathrm{P}_T(\lambda) \stackrel{\Delta}{=} 1\,). \tag{1}$$

Different trees and sets of conditional probabilities can generate the probability assignment (1); we refer to these as *tree models* for the source. A tree model (T, P_T) is *minimal* if for every internal node u of T there exist states uv and uw such that $\mathrm{P}_T(\cdot|uv) \not\equiv \mathrm{P}_T(\cdot|uw)$. We will loosely use the symbol T to refer both to a tree model and to its underlying tree, the conditional distributions being understood from the context. All expectations in the sequel will be with respect to $\mathrm{P}_T(\,\cdot\,)$.

For a sequence x^n, a string s, and a symbol $a \in \mathcal{A}$, define

$$n_s^{(a)}(x^n) = |\{\, i \,:\, 0 \le i < n,\ x_{i-k+1}^i = \bar{s},\ x_{i+1} = a\,\}|\,,$$

namely, the number of occurrences of a in context $\bar{s}$ (or, if $s \in S_T$, in state s) in x^n. Define also $n_s(x^n) = \sum_{a\in\mathcal{A}} n_s^{(a)}(x^n)$, the number of occurrences of context $\bar{s}$ (or state s) in x^{n-1}. (We omit the dependence of counts on x^n when clear from the context.) Notice that we also have $n_s = \sum_{a\in\mathcal{A}} n_{sa}$. Furthermore, denoting by $\mathbf{i}(u)$ and $\mathbf{f}(u)$ the indicator functions of the predicates $\overline{u} = x_{-|u|+1}^0$ and $\overline{u} = x_{n-|u|+1}^n$ (i.e., x_1 occurs in context $\overline{u}$ and $\overline{u}$ occurs at the end of x^n), respectively, we have $n_{as} + \mathbf{f}(as) = n_s^{(a)} + \mathbf{i}(as)$ for every $a \in \mathcal{A}$. To simplify expressions, we will use a generic constant δ to account for border adjustments due to terms of the form $\mathbf{i}(u)$ and $\mathbf{f}(u)$. In coding situations these terms will be known to the decoder, and in any case border effects will have no bearing on the asymptotic results.

From (1), using a simple algebraic argument, it follows that for sequences x^n and y^n, we have $\mathrm{P}_T(x^n) = \mathrm{P}_T(y^n)$ for all choices of the distributions $P(\,\cdot\,|s)$, $s \in S_T$, if and only if $n_s^{(a)}(x^n) = n_s^{(a)}(y^n)$ for all $s \in S_T$, $a \in \mathcal{A}$. Thus, in the case of tree models (as in other cases of interest), the notion of type defined in probabilistic terms in Section 1 admits a combinatorial characterization. For a tree T, and a sequence x^n, we denote by $\mathcal{N}(T, x^n)$ the collection of counts $\{\, n_s^{(a)} \,\}_{s\in S_T,\, a\in\mathcal{A}}$. The *type class* of x^n with respect to T can then be defined as

$$\mathcal{T}(T, x^n) = \{\, y^n \in \mathcal{A}^n \,:\, \mathcal{N}(T, y^n) = \mathcal{N}(T, x^n)\,\}\,.$$

A tree T and a sequence x^n determine a probability distribution, $\hat{P}_T$, defined by the empirical conditional probabilities $\hat{P}_T(a|s) = n_s^{(a)}(x^n)/n_s(x^n)$ (as before, we omit the dependence of the distribution $\hat{P}_T$ on x^n when clear from the context); $\hat{P}_T(x^n)$ is the maximum likelihood probability of x^n under T.

An *enumerative (source) code* for T is comprised of two parts: a description of $\mathcal{N}(T, x^n)$, and an index of x^n within $\mathcal{T}(T, x^n)$. By using a trivial bound $|\mathcal{T}(T, x^n)| \leq \prod_{s \in S_T} \frac{n_s!}{\prod_{a \in \mathcal{A}} n_s^{(a)}!}$, Stirling's formula, and the application of bounds on expectations from [12], we obtain the following lemma, which bounds the length of the second part of the enumerative code.

Lemma 1 *Let T be a tree source with entropy rate $\mathcal{H}$ and all conditional probabilities nonzero. Then,*

$$\frac{1}{n} E\left[\log |\mathcal{T}(T, x^n)|\right] \leq \mathcal{H} - |S_T|(\alpha - 1)\frac{\log n}{2n} + O\left(\frac{1}{n}\right). \tag{2}$$

Denote by $\mathcal{N}_b(T, x^n)$, $b \in \mathcal{A}$, the collection of counts $\{n_s^{(a)}\}_{s \in S_T,\, a \in \mathcal{A} \setminus \{b\}}$. The lemma below follows from simple linear algebra arguments on the set of state-transition equations for an FSM. The proof is omitted.

Lemma 2 *If a tree T is FSM, then for any $b \in \mathcal{A}$, $\mathcal{N}_b(T, x^n)$ and the final state, $\sigma(x^n)$, of x^n in T completely determine $\mathcal{N}(T, x^n)$.*

A generalization of Lemma 2 for unrestricted tree models can be found in [17].

3 Non-uniform codes for symbol counts

Since all the counts in $\mathcal{N}_b(T, x^n)$ are upper-bounded by n, it follows from Lemma 2 that $(\alpha-1)|S_T| \log n + \log |S_T|$ bits suffice to describe $\mathcal{T}(T, x^n)$ when T is FSM. Therefore, by (2), an enumerative code for an FSM tree T, based on uniform coding of the type class, is universal (in expectation), with an optimal normalized redundancy of $|S_T|(\alpha-1)\frac{\log n}{2n}$ bits. As discussed in Section 1, however, for a general tree model, an enumerative code with a uniform encoding of type classes may be suboptimal. This motivates the discussion in this section, where we present a class of non-uniform codes, dubbed SCCs (*symbol count codes*) for describing symbol counts $n_w^{(a)}$ in certain contexts w. These codes will be used as a tool to bound the expected size of the tree type classes in Section 4, and, with a further generalization, to construct the actual non-uniform code for $\mathcal{N}(T, x^n)$ in Section 5. Together, both contributions will lead to a universal, efficiently computable enumerative code for general tree sources.

Consider a symbol a and a fixed context w such that $s \prec w$ for some state s of T. Define

$$z_{w,a} = n_w^{(a)} - \frac{n_s^{(a)}}{n_s} n_w\,, \quad Z_{w,a} = |z_{w,a}|\,, \quad \text{and} \quad \mathrm{sg}_{w,a} = \begin{cases} 1 & z_{w,a} > 0\,, \\ 0 & \text{otherwise}\,. \end{cases}$$

As customary, denote by $\lfloor z \rfloor$ (resp. $\lceil z \rceil$) the largest (resp. smallest) integer satisfying $\lfloor z \rfloor \leq z \leq \lceil z \rceil$. Given $\mathrm{sg}_{w,a}$, $\lfloor Z_{w,a} \rfloor$, n_s, $n_s^{(a)}$, and n_w, it is possible to reconstruct $n_w^{(a)}$ as

$$n_w^{(a)} = \begin{cases} \lfloor Z_{w,a} \rfloor + \left\lceil \frac{n_s^{(a)}}{n_s} n_w \right\rceil\,, & \mathrm{sg}_{w,a} = 1\,, \\ \left\lfloor \frac{n_s^{(a)}}{n_s} n_w \right\rfloor - \lfloor Z_{w,a} \rfloor\,, & \text{otherwise}\,. \end{cases} \tag{3}$$

Hence, if n_s, $n_s^{(a)}$, and n_w are known by a decoder, encoding $\mathrm{sg}_{w,a}$ and $\lfloor Z_{w,a} \rfloor$ suffices to describe $n_w^{(a)}$. For $n_w > 0$ we will encode $\lfloor Z_{w,a} \rfloor$ using a Golomb code of parameter $\lceil \sqrt{n_w} \rceil$.

Specifically, we use a unary code for the integer division $\lfloor Z_{w,a} \rfloor^{\top} = \lfloor Z_{w,a} \rfloor / \lceil \sqrt{n_w} \rceil$, using $\lfloor Z_{w,a} \rfloor^{\top} + 1$ bits, and encode $\lfloor Z_{w,a} \rfloor^{\perp} = \lfloor Z_{w,a} \rfloor \mod \lceil \sqrt{n_w} \rceil$ uniformly with $\log \lceil \sqrt{n_w} \rceil$ bits. When $n_w = 0$, we have $n_w^{(a)} = 0$ for all a, and encoding $n_w^{(a)}$ is not necessary; we define $\lfloor Z_{w,a} \rfloor^{\top} = 0$ in this case, to simplify discussions on expectations.

The intuition behind this code, which we denote $C_{w,a}\left(n_w^{(a)}\right)$, is that we can think of $Z_{w,a}$ as the absolute difference between the true value of $n_w^{(a)}$ and the estimate $\frac{n_s^{(a)}}{n_s} n_w$ that the decoder could guess from the known counters n_s, $n_s^{(a)}$ and n_w for a typical sequence of the source (T, P_T). The probability of the estimate $\frac{n_s^{(a)}}{n_s} n_w$ differing from the true value decays exponentially fast, which leads to a constant expectation of $\lfloor Z_{w,a} \rfloor^{\top}$, as stated in the following lemma.

Lemma 3 *Let T be a tree source with all conditional probabilities in states of T different from zero. Then, the expectation $E\left[\, \lfloor Z_{w,a} \rfloor^{\top} \right]$ is upper-bounded by a constant independent of n.*

For conciseness, the proof of Lemma 3 is omitted here. Although somewhat technical, it follows rather straightforwardly using a large deviations argument based on [19, Theorem 2].

By Lemma 3, the expected length of the unary part of the Golomb code used in SCCs is upper-bounded by a constant. On the other hand, when $n_w > 0$, a uniform encoding of $\lfloor Z_{w,a} \rfloor^{\perp} = \lfloor Z_{w,a} \rfloor \mod \lceil \sqrt{n_w} \rceil$ takes $\log \lceil \sqrt{n_w} \rceil$ bits. The code length of this uniform part can be upper-bounded by $\log\left(\sqrt{n_w} + 1\right) \leq 1 + \frac{1}{2} \log n$.

The foregoing discussion yields the following corollary to Lemma 3.

Corollary 1 *The expected length of the code $C_{w,a}\left(n_w^{(a)}\right)$ is upper-bounded by $\frac{1}{2} \log n + O(1)$.*

Thus, under appropriate conditions, SCCs can code occurrence counts using, on average, half the length of the naive encoding.

4 The expected size of $\mathcal{T}(T, X^n)$.

In this section we study the asymptotic behavior of $E\,|\mathcal{T}(T, X^n)|$, thus estimating the expected length of a uniform encoding of the index of x^n within its type class. The index can be computed and uniformly encoded, efficiently, using the combinatorial characterization of the tree type class in [17]. When T is FSM, the bound in (2) for $\frac{1}{n}\mathrm{E}\,[\log |\mathcal{T}(T, x^n)|\,]$ is tight, and as mentioned, it leads readily, by means of Lemma 2, to the optimality of enumerative coding where type classes are encoded uniformly. For general trees however, the coefficient $\frac{1}{2}|S_T|(\alpha - 1)$ in the negative term of order $\frac{\log n}{n}$ in (2) is not the best one can obtain, and a larger coefficient is necessary to offset a corresponding length increase in the type class description part. We derive a tighter bound later in this section.

We first show how an economic description of $\mathcal{N}(T_F, x^n)$, the type class of x^n with respect to the FSM closure of T, can be obtained by means of SCCs from one of $\mathcal{N}(T, x^n)$. This, together with the bound (2) applied to T_F, and a coding argument, will lead to the desired tight bound. The description of $\mathcal{N}(T, X^n)$ itself will be discussed in Section 5, and will require additional tools.

Let $\mathcal{I}(T) = T \setminus S_T$ denote the set of internal nodes of T. We say that a state $s \in S_T$ is *forgetful* if $as \in \mathcal{I}(T)$ for *all* $a \in \mathcal{A}$. In a forgetful state, a next-state transition cannot be determined for *any* occurring symbol $a \in \mathcal{A}$. A tree with no forgetful states is called *canonical*. It can readily be shown that, by sequentially refining forgetful states until no such state is left, a tree T is brought to a unique minimal canonical extension [17], which will be denoted $T_{\mathbf{c}}$. In a sense, forgetful states are the farthest away from satisfying the FSM property, and $T_{\mathbf{c}}$ brings a tree T "closer" to its FSM closure T_F. Thus, we have $T \subseteq T_{\mathbf{c}} \subseteq T_F$. It is also readily verified that if an extension step on a forgetful state s transforms, say, a tree T' into T'', then $\mathcal{N}(T', x^n)$ and the final state of x^n in T'' completely determine $\mathcal{N}(T'', x^n)$. Thus, $\mathcal{N}(T_{\mathbf{c}}, x^n)$ is fully determined by $\mathcal{N}(T, x^n)$ and $\sigma_{\mathbf{c}}(x^n)$, the final state of x^n in $T_{\mathbf{c}}$, leading to the following result.

Lemma 4 *$\mathcal{N}(T_{\mathbf{c}}, x^n)$ can be reconstructed from a description of $\mathcal{N}(T, x^n)$ and a description, of length $\log |S_{T_{\mathbf{c}}}|$, of $\sigma_{\mathbf{c}}(x^n)$.*

We define the *state transitions graph* $G_T{=}(V_T, E_T)$ of a tree T, with vertex set $V_T{=}S_T$, and edge set E_T comprising all state pairs (u, v) such that some sequence causes a direct transition from u to v in T. It is readily verified that E_T is given by

$$E_T = \{\, (u, v) \,:\, u \preceq \text{tail}(v) \text{ or } \text{tail}(v) \prec u \,\} \,. \tag{4}$$

Lemma 5 ([17]) *Let $G_{T_{\mathbf{c}}}{=}(V_{T_{\mathbf{c}}}, E_{T_{\mathbf{c}}})$ be the state transitions graph of $T_{\mathbf{c}}$. The number of type classes of sequences of length n with respect to T is $\Theta\left(n^{|E_{T_{\mathbf{c}}}|-|V_{T_{\mathbf{c}}}|}\right)$.*

The value in the exponent of the bound of Lemma 5, $|E_{T_{\mathbf{c}}}| - |V_{T_{\mathbf{c}}}|$, will arise also in the coefficient of the term of order $\log n/n$ in the normalized expectation of $\log |\mathcal{T}(T, X^n)|$, which we will show to be upper-bounded by $\mathcal{H} - \frac{|E_{T_{\mathbf{c}}}|-|V_{T_{\mathbf{c}}}|}{2n} \log n + O(1/n)$. When T is FSM, T is also canonical, and there are exactly α edges departing from each state in T. Thus, in this case, we have $|E_{T_{\mathbf{c}}}| - |V_{T_{\mathbf{c}}}| = |S_T|(\alpha - 1)$, in agreement with Lemma 1 and Lemma 2.

For a node $t \in \mathcal{I}(T_F) \setminus \mathcal{I}(T_{\mathbf{c}})$ we define the (FSM) *over-refinement* of t as $\kappa_t = |\{a \in \mathcal{A} : at \notin \mathcal{I}(T_{\mathbf{c}})\}|$. The name given to κ_t stems from the fact that it counts symbols for which an extension from t was not needed in order to determine a next-state transition in $T_{\mathbf{c}}$, yet it was added in the process of constructing the FSM closure T_F. The total *over-refinement* of T is now defined as

$$\kappa_T = \sum_{t \in \mathcal{I}(T_F) \setminus \mathcal{I}(T_{\mathbf{c}})} (\alpha - 1)(\kappa_t - 1) \,. \tag{5}$$

The following lemma connects κ_T with $|E_{T_{\mathbf{c}}}| - |V_{T_{\mathbf{c}}}|$ and $|E_{T_F}| - |V_{T_F}| = |S_{T_F}|(\alpha - 1)$. The proof, which is omitted here, follows essentially from the definitions in (4) and (5).

Lemma 6 *Let S_{T_F} denote the set of states of T_F. Then,*

$$\kappa_T = -(|E_{T_{\mathbf{c}}}| - |V_{T_{\mathbf{c}}}|) + |S_{T_F}|(\alpha - 1) \,. \tag{6}$$

As discussed above, the exponent in the asymptotic growth of the number of type classes in T and T_F is given by $|E_{T_{\mathbf{c}}}| - |V_{T_{\mathbf{c}}}|$ and $|S_{T_F}|(\alpha - 1)$ respectively. Hence, by Lemma 6, there is, asymptotically, a factor of n^{κ_T} more type classes in T_F than in T, which suggests that roughly κ_T counts of $\log n$ bits each would suffice to describe $\mathcal{N}(T_F, x^n)$ from $\mathcal{N}(T, x^n)$. The next lemma confirms this intuition, and establishes the fact that the counts can be encoded, on average, with a cost of at most $\frac{1}{2} \log n$ bits each.

Lemma 7 *Given $\mathcal{N}(T, x^n)$, the collection $\mathcal{N}(T_F, x^n)$ can be described with SCC encodings of κ_T counts $n_w^{(a)}$ as discussed following (3), plus a constant number of bits used to describe $\sigma_{\mathbf{F}}(x^n)$, the final state of x^n in T_F.*

Proof. Since $\sigma_{\mathbf{F}}(x^n)$ determines $\sigma_{\mathbf{c}}(x^n)$, by Lemma 4, the decoder can reconstruct $\mathcal{N}(T_{\mathbf{c}}, x^n)$ from $\mathcal{N}(T, x^n)$. Let $\Delta = \mathcal{I}(T_F) \backslash \mathcal{I}(T_{\mathbf{c}})$. We will describe $n_{tc}^{(a)}$ for every child tc of $t \in \Delta$ and every $a \in \mathcal{A}$, proceeding in ascending order of length of t. We claim that proceeding this way we are sure that n_{tc} is known (has been described) when describing $n_{tc}^{(a)}$, and, thus, the latter can be encoded using SCCs. Indeed, if $t = bu \in \Delta$, with $b \in \mathcal{A}$ and $u \in T_{\mathbf{c}}$, then $n_{bu} = n_u^{(b)} + \delta$ is known. Otherwise, if $t = bu$ but $u \notin T_{\mathbf{c}}$, then u is an internal node of T_F shorter than bu, and thus $n_{buc} = n_{uc}^{(b)} + \delta$ is known for every child buc of bu. Thus, the claimed order of description is satisfied. Consider now a node $tc \in \Delta$, $c \in \mathcal{A}$. For every symbol a such that at is an internal node of $T_{\mathbf{c}}$, $atc \in T_{\mathbf{c}}$ and therefore $n_{tc}^{(a)} = n_{atc} + \delta$ is known, and requires no further description. We take now a symbol b such that bt is not an internal node of $T_{\mathbf{c}}$, and, for each child tc of t with $c \neq b$, we describe $\kappa_t - 1$ counts $n_{tc}^{(a)}$, $a \neq b$ such that at is not an internal node of $T_{\mathbf{c}}$. We can then compute $n_{tc}^{(b)} = n_{tc} - \sum_{a \neq b} n_{tc}^{(a)}$ and then $n_{tb}^{(a)} = n_t^{(a)} - \sum_{c \neq b} n_{tc}^{(a)}$ for every $a \in \mathcal{A}$. Overall, we obtain $\mathcal{N}(T_F, x^n)$ from $\mathcal{N}(T_{\mathbf{c}}, x^n)$ and $\sigma_{\mathbf{F}}(x^n)$ by providing $(\alpha - 1)(\kappa_t - 1)$ counts for each $t \in \Delta$, each of which can be encoded using SCCs. □

In the following theorem, we apply the results of Lemma 7 and Corollary 1 in a coding argument, to obtain the desired upper bound on the expectation of $|\mathcal{T}(T, X^n)|$.

Theorem 1 *Let X^n be a random sequence emitted by a tree source T with entropy rate $\mathcal{H}$ and all conditional probabilities different from zero. Then,*

$$\frac{1}{n} E\left[\log |\mathcal{T}(T, X^n)|\right] \leq \mathcal{H} - \frac{|E_{T_{\mathbf{c}}}| - |V_{T_{\mathbf{c}}}|}{2n} \log n + O\left(\frac{1}{n}\right). \tag{7}$$

Proof. A sequence in $\mathcal{T}(T, x^n)$ can be encoded by describing the subset $\mathcal{T}(T_F, x^n)$ to which the sequence belongs, and then, uniformly, its index within that subset. Hence, by Lemma 7, and Corollary 1, we have

$$\mathrm{E}\left[\log |\mathcal{T}(T, X^n)|\right] \leq \mathrm{E}\left[\log |\mathcal{T}(T_F, X^n)|\right] + \frac{1}{2}\kappa_T \log n + O(1), \tag{8}$$

the left-hand side of (8) being a lower bound on the expected length of any such description, since the sequences in the type class are equiprobable. Normalizing, and applying Lemma 1 to T_F, we obtain

$$\frac{1}{n}\mathrm{E}\left[\log |\mathcal{T}(T_F, X^n)|\right] \leq \mathcal{H} - \frac{|S_{T_F}|(\alpha - 1)}{2n} \log n + O\left(\frac{1}{n}\right),$$

which, together with (8) and Lemma 6, yields

$$\frac{1}{n}\mathrm{E}\left[\log |\mathcal{T}(T, X^n)|\right] \leq \mathcal{H} - \frac{|E_{T_{\mathbf{c}}}| - |V_{T_{\mathbf{c}}}|}{2n} \log n + O\left(\frac{1}{n}\right). \tag{9}$$

□

Theorem 1 complements, by providing an asymptotic interpretation, the exact combinatorial characterization of the tree type class in [17]. While the combinatorial characterization is instrumental in implementing the enumerative code, the asymptotic result helps us estimate the average code length.

5 Encoding the type class

In this section, we present an efficiently computable description of the type class $\mathcal{T}(T, x^n)$ (or, equivalently, the counts $\mathcal{N}(T, x^n)$), which, together with the enumeration of the type class, will yield the sought universal enumerative code for tree sources. We assume, throughout, that the tree is not trivial, i.e., $|S_T| > 1$.

In analogy to Lemma 2 for FSMs, it is shown in [17] that $\mathcal{N}(T, x^n)$ can be described with $|E_{T_\mathbf{c}}| - |V_{T_\mathbf{c}}|$ symbol counts. Using $\log n$ bits to encode each, and bounding the expectation of $\frac{1}{n} \log |\mathcal{T}(T, X^n)|$, where X^n is a random sequence emitted by T, as in Theorem 1, we obtain an upper bound of the form $\frac{|E_{T_\mathbf{c}}|-|V_{T_\mathbf{c}}|}{2n} \log n + O(\frac{1}{n})$ on the normalized "redundancy" of the code length over the source entropy rate $\mathcal{H}$. In general, however, $|E_{T_\mathbf{c}}| - |V_{T_\mathbf{c}}|$ may be strictly larger than $|S_T|(\alpha - 1)$, and such code may be suboptimal. Therefore, a tighter description of the type class is needed. We will show that encoding $\mathcal{N}_b(T, x^n)$ (which is generally insufficient to characterize the type class) uniformly with $|S_T|(\alpha - 1)$ counts of $\log n$ bits each, we can complete the description of $\mathcal{N}(T, x^n)$ by encoding an additional $|E_{T_\mathbf{c}}| - |V_{T_\mathbf{c}}| - |S_T|(\alpha - 1)$ counts requiring on average $\frac{1}{2} \log n + O(1)$ bits of description each. Together with the bound of Theorem 1, this reduction in code length for the additional counts will result in an optimal expected normalized redundancy of $\frac{|S_T|(\alpha-1)}{2n} \log n + O(1/n)$ bits over $\mathcal{H}$.

Let h and d denote, respectively, the minimal and maximal depth of leaves in T. For $h \le m \le d$, let $T^{[m]}$ denote the truncation of T to depth m, and let $T_\mathbf{c}^{[m]}$ denote the canonical tree of $T^{[m]}$. Notice that, by the definition of a forgetful state, no state of maximal depth of $T^{[m]}$ is refined in $T_\mathbf{c}^{[m]}$, and therefore $T_\mathbf{c}^{[m]}$ has the same depth as $T^{[m]}$. We denote by $S_\mathbf{c}^{[m]}$ the set of states of $T_\mathbf{c}^{[m]}$, and by $\sigma_\mathbf{c}^{[m]}(u)$ the state selected by u in $T_\mathbf{c}^{[m]}$. Algorithm *EncodeTypeClass*, shown in Figure 2, lists the main steps in the proposed encoding of $\mathcal{N}(T, x^n)$. The algorithm starts by encoding, uniformly, the counts in $\mathcal{N}_b(T, x^n)$ with $\log n$ bits per count, and the final state of x^n in $T_\mathbf{c}$, using a constant number of bits. It then iterates to describe, incrementally, each count set $\mathcal{N}(T_\mathbf{c}^{[k+1]}, x^n)$ given a previously described set $\mathcal{N}(T_\mathbf{c}^{[k]}, x^n)$, for $h+1 \le k < d$. Notice that since $T^{[h]}$ is a full balanced tree, $T^{[h+1]}$ is FSM, and, hence, $T_\mathbf{c}^{[h+1]} = T^{[h+1]}$. By Lemma 2, $\mathcal{N}(T^{[h+1]}, x^n)$ is completely determined by $\mathcal{N}_b(T, x^n)$ and the given final state. Thus, a decoder can reconstruct $\mathcal{N}(T_\mathbf{c}^{[h+1]}, x^n)$ from the information provided in Step 1 of EncodeTypeClass, and can recover $\mathcal{N}(T_\mathbf{c}^{[d]}, x^n)$ from the information encoded in the loop of Steps 3–4. Since $T_\mathbf{c}^{[d]}$ is a refinement of T, this is sufficient to reconstruct $\mathcal{N}(T, x^n)$.

Algorithm EncodeTypeClass(T, x^n)

---

1. Encode $\mathcal{N}_b(T, x^n)$ and the final state of x^n in $T_\mathbf{c}$.
2. Set $h = \min\{|s| : s \in S_T\}$ and $d = \max\{|s| : s \in S_T\}$.
3. For $k = h + 1$ to $d - 1$
4. Encode $\mathcal{N}(T_\mathbf{c}^{[k+1]}, x^n)$ given $\mathcal{N}(T_\mathbf{c}^{[k]}, x^n)$.

Figure 2: Encoding of $\mathcal{N}(T, x^n)$

Clearly, the crucial step in EncodeTypeClass is the encoding of the refinement of counters from $T_\mathbf{c}^{[k]}$ to $T_\mathbf{c}^{[k+1]}$ in Step 4. For $u, v \in \mathcal{A}^*$, we denote by $N_{u,v}$ the number of times a transition from context $\overline{u}$ to context $\overline{v}$ occurs in x^n. In particular, when u and v are states

of a tree, $N_{u,v}$ denotes the number of times state v is selected immediately after state u. Notice that, for a state t, we have

$$n_t = \sum_{s \in S_T} N_{t,s} = \sum_{s \in S_T} N_{s,t} + \delta \,. \tag{10}$$

Our implementation of Step 4 will amount to describing all state transition counts $N_{s,t}$, with $s, t \in S_{\mathbf{c}}^{[k+1]}$. This set of counts is sufficient to determine $\mathcal{N}(T_{\mathbf{c}}^{[k+1]}, x^n)$ as we have $n_s^{(a)} = \sum_{au \in S_{\mathbf{c}}^{[k+1]}} N_{s,au}$. However, not all the counters in the set will be explicitly described, since some will be derivable from $\mathcal{N}(T_{\mathbf{c}}^{[k]}, x^n)$ and earlier portions of $\mathcal{N}(T_{\mathbf{c}}^{[k+1]}, x^n)$. The crux of the encoding is to find a minimum subset of transition counts, and the order in which they are described, that suffice to determine *all* of them, and, at the same time, can be described economically.

The following lemma presents conditions that will allow a further simplification in the description of $\mathcal{N}(T_{\mathbf{c}}^{[k+1]}, x^n)$. The proof is straightforward, and is omitted here.

Lemma 8 *For k such that $h+1 \le k \le d-1$, let t be a state of $T_{\mathbf{c}}^{[k+1]}$ such that $\mathrm{tail}(t) \notin \mathcal{I}(T_{\mathbf{c}}^{[k+1]})$. Then, all transitions into t depart from a unique state $s = \sigma_{\mathbf{c}}^{[k+1]}(\overline{\mathrm{tail}(t)})$. Thus, from (10), $N_{s,t} = n_t + \delta$, and encoding $N_{s,t}$ is equivalent to encoding n_t.*

In our implementation of Step 4 of EncodeTypeClass, all explicit encodings will be for counts $N_{s,t}$ chosen with t satisfying the conditions of Lemma 8, and the following additional condition: there exists a state s' *of* T such that $s' \prec \mathrm{tail}(t)$. It turns out that the SCCs of Section 3, which are defined for individual symbols, do not suffice to implement these encodings optimally. (They did suffice in Section 4, when used to encode symbol counts for nodes that refined the original tree T; here, however, we need to encode counts for nodes of T.) Therefore, next, we generalize SCCs to *string count codes* (*SCCs**), which, as their name suggests, are defined on string counts rather than counts of individual symbols.

Consider a fixed string $u = u^q$ such that $s' \prec \overline{u^{q-1}}$ for some $s' \in S_T$. We define a code for $n_{\overline{u}}$, and analyze its expected code length. Let $l = \max\{\, j : 1 \le j < q-1,\ \overline{u^j} \in T \,\}$. Since $\underline{u_1} \in T$, l is well defined. Also, all proper prefixes u^i of u, $l < i < q$, are sufficiently long for $\overline{u^i}$ to determine a state in T. For $l < i < q$, define the following short-hand notations:

$$s_i = \sigma(u^i)\,, \qquad n_i = n_{s_i}\,, \qquad n_i^\alpha = n_{s_i}^{(u_{i+1})}\,, \qquad m_i = n_{\overline{u^i}}\,. \tag{11}$$

The occurrence of context u^q in x^n under the above conditions implies the occurrence of the state sequence $\{s_i\}$, $l{+}1 \le i \le q{-}1$, with symbol u_{i+1} occurring in state s_i (except possibly in border situations, which we shall ignore). In the language of [17], this is a *forced state sequence*. For example, in the tree T_1 of Figure 1, an occurrence of state 100 must be preceded by two occurrences of state 0, with forced occurrences of the symbols 0 and 1, respectively. This knowledge is exploited in the following manner: given m_{l+1}, the number of times context u^{l+1} occurs within x^n, we can estimate $n_{\overline{u}}$ by $m_{l+1} \prod_{i=l+1}^{q-1} \frac{n_i^\alpha}{n_i}$. Define (with a slight abuse of notation previously defined for SCCs),

$$z_u = n_{\overline{u}} - m_{l+1} \prod_{i=l+1}^{q-1} \frac{n_i^\alpha}{n_i}\,, \quad Z_u = |z_u|\,, \quad \text{and} \quad \mathrm{sg}_u = \begin{cases} 1 & z_u > 0\,, \\ 0 & \text{otherwise}\,. \end{cases}$$

In the encoding of $n_{\overline{u}}$ with SCCs*, denoted $C_u^\star(n_{\overline{u}})$, we encode $\lfloor Z_u \rfloor$ using a Golomb code of parameter $\lceil \sqrt{m_{l+1}} \rceil$, namely, a unary code for the integer quotient $\lfloor Z_u \rfloor^\top =$

$\lfloor Z_u \rfloor / \lceil \sqrt{m_{l+1}} \rceil$, using $\lfloor Z_u \rfloor^\top + 1$ bits, concatenated with a uniform code for $\lfloor Z_u \rfloor^\perp = \lfloor Z_u \rfloor \mod \lceil \sqrt{m_{l+1}} \rceil$ using $\log\left(\lceil \sqrt{m_{l+1}} \rceil\right)$ bits. When $n_i = 0$ for some $l < i < q$, or when $m_{l+1} = 0$, we also have $n_{\overline{u}} = 0$, and no encoding is necessary; we define $\lfloor Z_u \rfloor^\top = 0$ in this case. The results below are analogues of Lemma 3 and Corollary 1.

Lemma 9 *Let T be a tree source with all conditional probabilities different from zero. Then, the expectation $E\left[\lfloor Z_u \rfloor^\top\right]$ is upper-bounded by a constant independent of n.*

Corollary 2 *The expected code length of $C_u^\star(n_{\overline{u}})$ is upper-bounded by $\frac{1}{2}\log n + O(1)$.*

Corollary 2 shows that SCCs* provide an efficient way to encode certain string occurrence counts. Lemma 8 provides a way to recover transition counts $N_{s,t}$ from string counts n_t, and certain subsets of these transition counts are sufficient to reconstruct $\mathcal{N}(T_{\mathbf{c}}^{[k+1]}, x^n)$. We will next show that it is possible to select a minimal set of counts for strings t satisfying Lemma 8, and the order in which they are described, to obtain a complete description of $\mathcal{N}(T_{\mathbf{c}}^{[k+1]}, x^n)$ using SCCs*. We distinguish between states of $T_{\mathbf{c}}^{[k+1]}$ that are added to $T_{\mathbf{c}}^{[k]}$ for they belong to $T^{[k+1]}$, and states that arise in $T_{\mathbf{c}}^{[k+1]}$ in order to take $T^{[k+1]}$ to canonical form. Let $U'_{k+1} = (T_{\mathbf{c}}^{[k+1]} \backslash T^{[k+1]}) \backslash T_{\mathbf{c}}^{[k]}$ be the set of nodes of $T_{\mathbf{c}}^{[k+1]}$ which are not in the original tree $T^{[k+1]}$ and are not in $T_{\mathbf{c}}^{[k]}$. Let U_{k+1} be the set of parent nodes of elements of U'_{k+1}, $U_{k+1} = \{z : za \in U'_{k+1}, a \in \mathcal{A}\}$. The following lemma and corollary are instrumental in identifying an appropriate set of counts to describe $\mathcal{N}(T_{\mathbf{c}}^{[k+1]}, x^n)$.

Lemma 10 *For $h+1 \le k \le d-1$, we have $U_{k+1} \subseteq S_{\mathbf{c}}^{[k]}$.*

Corollary 3 *For $h+1 \le k \le d-1$, $T_{\mathbf{c}}^{[k+1]}$ refines states of $T_{\mathbf{c}}^{[k]}$ by at most one level.*

The proof of Lemma 10 is deferred to Appendix A. The proof of Corollary 3 then follows readily.

All the encodings in Step 4 of EncodeTypeClass will be done through the auxiliary procedure P shown in Figure 3. We denote by $S_{\mathbf{c}}^{[m]}(r)$ the set of states of $T_{\mathbf{c}}^{[m]}$ that are children of r, i.e., of the form ra, $a \in \mathcal{A}$. Given a node r and a symbol c, $\mathrm{P}(r, c)$ describes $N_{s,t}$ for every $s \in S_{\mathbf{c}}^{[k+1]}(r)$, and every state $t \in S_{\mathbf{c}}^{[k+1]}$ such that $c = \mathrm{hd}(t)$. We assume (and will later verify) that when the procedure is called, these states t satisfy the conditions of Lemma 8, so that $N_{s,t} = n_t + \delta$. In the procedure, b is a fixed but arbitrary symbol from $\mathcal{A}$, and $\mathcal{A}_b = \mathcal{A} \setminus \{b\}$. Decoding steps are shown in brackets, to verify the losslessness of the code.

Notice the use of SCCs* in Steps 2 and 9. In Step 2, $t = crd$ is a state of $T_{\mathbf{c}}^{[k+1]}$ and, thus, $|cr| < k+1$. Hence, r is a leaf of $T_{\mathbf{c}}^{[k]}$ with $|r| < k$, which implies that there exists a state in T that is a proper prefix of rd for all $d \in \mathcal{A}$, which is a condition for SCCs* to be applicable. In the case of Step 9, it can be shown that the required condition is guaranteed by Step 5, using similar arguments. Furthermore, in the application of SCCs* in Steps 2 and 9, all states s_i in the definition (11), have a length smaller than $k+1$. Thus, n_i, n_i^α, and m_{l+1} are known from $\mathcal{N}(T_{\mathbf{c}}^{[k]}, x^n)$.

We are now ready to present the full implementation of Step 4 of EncodeTypeClass, which is shown as Procedure *RefineTypeClass* in Figure 4. Procedure RefineTypeClass selects transition counts and an order of description that allows Procedure P (and, thus, SCCs*) to be used, and, as we shall prove, such that the total number of counts that

Procedure P(r, c)
Assumption: $cr \in \mathcal{I}(T_{\mathbf{c}}^{[k+1]})$, so Lemma 8 holds when coding n_t.

1. If r and cr are leaves of $T_{\mathbf{c}}^{[k]}$ but internal nodes of $T_{\mathbf{c}}^{[k+1]}$
   /* *From Corollary 3,* $rd \in S_{\mathbf{c}}^{[k+1]}$ *and* $crd \in S_{\mathbf{c}}^{[k+1]}$ $\forall d \in \mathcal{A}$. */
2. Use SCCs* to encode $\alpha{-}1$ counts $n_t = n_{crd}$, $d \in \mathcal{A}_b$.
3. [Reconstruct $n_{crb} = n_{cr} - \sum_{d \neq b} n_{crd}$
   and $N_{s,t}{=}n_{crd} + \delta$ for all $d{\in}\mathcal{A}$, with $s = rd$, $t = crd$].
4. else, for each $s \in S_{\mathbf{c}}^{[k+1]}(r)$
5. If cs is an internal node of $T_{\mathbf{c}}^{[k+1]}$
6. Let W' be the set of states $W' = \{csv \in T_{\mathbf{c}}^{[k+1]} \backslash T_{\mathbf{c}}^{[k]}\}$.
7. Let $W = \{w : wa \in W'\}$ be the parent nodes of W'.
8. For each $csu \in W$ /* $csu \in T_{\mathbf{c}}^{[k]}$ *by Corollary 3.* */
9. Use SCCs* to encode $\alpha{-}1$ counts $n_t = n_{csud}$, $d \in \mathcal{A}_b$.
10. [Reconstruct $n_{csub} = n_{csu} - \sum_{d \neq b} n_{csud}$
    and $N_{s,t} = n_{csud} + \delta$ for all $d \in \mathcal{A}$, with $t = csud$.]
11. For each state $t = csv \notin W'$ of $T_{\mathbf{c}}^{[k+1]}$
12. [Reconstruct $N_{s,t} = n_{csv} + \delta$, from $\mathcal{N}(T_{\mathbf{c}}^{[k]}, x^n)$.]
13. else /* cs *is a leaf of* $T_{\mathbf{c}}^{[k+1]}$ *by the assumptions.* */
    /* *Either* $cs \in T_{\mathbf{c}}^{[k]}$, *or* $s \in T_{\mathbf{c}}^{[k]}$. *Otherwise, by Corollary 3, their respective parents* cr *and* r, *would belong to* $S_{\mathbf{c}}^{[k]}$ *and Step 1 would have not branched to 4.* */
14. [Reconstruct $N_{s,t} = n_{cs}{+}\delta$, from $\mathcal{N}(T_{\mathbf{c}}^{[k]}, x^n)$, for $t{=}cs$]

Figure 3: Encoding of state transition counts

Procedure RefineTypeClass

1. For each $r \in R_{k+1}$ taken in ascending order of length $|r|$
2. If $r \in U_{k+1}$ /* *This implies also* $r \in S_{\mathbf{c}}^{[k]}$. */
3. Take $d \in \mathcal{A}$ such that $dr \notin \mathcal{I}(T_{\mathbf{c}}^{[k]})$. /* *Such* d *must exist; otherwise,* r *would be a forgetful state of* $T_{\mathbf{c}}^{[k]}$. */
4. Use P(r, c) to describe $N_{s,csu}$ for all $s \in S_{\mathbf{c}}^{[k+1]}(r)$, $c \in \mathcal{A}_d$.
5. [Let $N_{s,ds} = n_s^{(d)} = n_s - \sum_{c \neq d} n_s^{(c)}$, for all $s \in S_{\mathbf{c}}^{[k+1]}(r)$.]
6. else, If the children of r belong to $T_{\mathbf{c}}^{[k]}$
7. For each $s \in S_{\mathbf{c}}^{[k+1]}(r)$, and $c \in \mathcal{A}$, such that $cs \notin T_{\mathbf{c}}^{[k+1]}$
8. Let $s' = \sigma_{\mathbf{c}}^{[k+1]}(\overline{cs})$.
9. [Take $N_{s,s'} = n_s^{(c)}$, known from $\mathcal{N}(T_{\mathbf{c}}^{[k]}, x^n)$.]
10. For each $s \in S_{\mathbf{c}}^{[k+1]}(r)$, and $c \in \mathcal{A}$, such that $cs \in T_{\mathbf{c}}^{[k+1]}$
11. Use P(r, c) to describe counts $N_{s,csu}$.
12. else, for each $s \in S_{\mathbf{c}}^{[k+1]}(r)$
13. [For $a \in \mathcal{A}_b$, $s' = \sigma_{\mathbf{c}}^{[k+1]}(\overline{as})$, let $N_{s,s'} = n_s^{(a)}$.]
14. [For $s' = \sigma_{\mathbf{c}}^{[k+1]}(\overline{bs})$, let $N_{s,s'} = n_s - \sum_{a \neq b} n_s^{(a)}$.]

Figure 4: Coding and decoding of $\mathcal{N}(T_{\mathbf{c}}^{[k+1]}, x^n)$ from $\mathcal{N}(T_{\mathbf{c}}^{[k]}, x^n)$

are actually encoded is precisely $|E_{T_\mathbf{c}}| - |V_{T_\mathbf{c}}| - |S_T|(\alpha - 1)$, as needed to achieve optimal expected redundancy.

We define $R_i = \{r \in \mathcal{A}^* : ra \in S_\mathbf{c}^{[i]} \text{ for some } a \in \mathcal{A}\}$, namely, the set of parent nodes of states of $T_\mathbf{c}^{[i]}$. Procedure RefineTypeClass iterates over nodes $r \in R_{k+1}$, and for each $s \in S_\mathbf{c}^{[k+1]}(r)$, it describes all potentially nonzero state transition counts $N_{s,t}$ with $t \in S_\mathbf{c}^{[k+1]}$. The correctness of the procedure is established in the following lemma. The code length is analyzed later in Lemma 12.

Lemma 11 *Assuming that $\mathcal{N}_b(T, x^n)$, $\mathcal{N}(T_\mathbf{c}^{[k]}, x^n)$, and $\sigma_\mathbf{c}(x^n)$ are known, Procedure RefineTypeClass correctly encodes $\mathcal{N}(T_\mathbf{c}^{[k+1]}, x^n)$.*

Proof. The algorithm iterates over all nodes $r \in R_{k+1}$. Then, in each of the three cases distinguished by the conditions of Steps 2, 6, 12, its computations (possibly involving the use of Procedure P) allow the the decoder to recover the counts for all state transitions that depart from every child of r that is a state of $T_\mathbf{c}^{[k+1]}$. What we need to show is that required conditions at various points of the computation are satisfied, and, in particular, that the assumptions of P are satisfied when the procedure is invoked. The losslessness of P itself was established in Figure 3 and its discussion.

In Step 3 we ask for a symbol d such that $dr \notin \mathcal{I}(T_\mathbf{c}^{[k]})$. The condition $r \in U_{k+1}$ satisfied in Step 2, together with Lemma 10, implies that r is a state of $T_\mathbf{c}^{[k]}$. If $dr \in \mathcal{I}(T_\mathbf{c}^{[k]})$ for all $d \in \mathcal{A}$, then r would be a forgetful state, contradicting the definition of $T_\mathbf{c}^{[k]}$. Hence, the symbol d called for in Step 3 must exist.

In Steps 4 and 11, Procedure P is used to encode state transition counts departing from a state s with a symbol c, requiring $cs \in T_\mathbf{c}^{[k+1]}$, so that the encoded counts are of the form $N_{s,csu}$ for some string u. We claim that the condition $cs \in T_\mathbf{c}^{[k+1]}$ is satisfied in Step 4. Indeed, since $r \in U_{k+1}$ (as tested in Step 2), and, as argued above, r is a state of $T_\mathbf{c}^{[k]}$, its refinement in $T_\mathbf{c}^{[k+1]}$ must have been part of the process of taking $T^{[k+1]}$ to canonical form. Therefore, r is forgetful in $T^{[k+1]}$, and, thus, cr is an internal node of $T^{[k+1]}$, which implies that cr is also an internal node of $T_\mathbf{c}^{[k+1]}$. In Step 11, the condition $cs \in T_\mathbf{c}^{[k+1]}$ is also satisfied since it is imposed in Step 10.

In Step 5 we compute $N_{s,ds}$ as $n_s^{(d)}$, implicitly assuming that ds is a state of $T_\mathbf{c}^{[k+1]}$. As argued above for cr, dr must be an internal node of $T_\mathbf{c}^{[k+1]}$. However dr is not an internal node of $T_\mathbf{c}^{[k]}$ by definition in Step 3, thus by Corollary 3, dr is a leaf of $T_\mathbf{c}^{[k]}$ and ds is a leaf of $T_\mathbf{c}^{[k+1]}$.

The computations in Steps 5 and 14 require the knowledge of n_s, the occurrence count of a state $s \in S_\mathbf{c}^{[k+1]}$. We claim that when the algorithm takes an element r of R_{k+1} in an iteration of the loop in Step 1, the decoder knows n_s for every $s \in S_\mathbf{c}^{[k+1]}(r)$. Consider a state $s \in S_\mathbf{c}^{[k+1]}(r)$, and let $v = \text{tail}(s)$. If v is an internal node of $T_\mathbf{c}^{[k+1]}$, $v \in T_\mathbf{c}^{[k]}$ by Corollary 3, and with $a = \text{hd}(s)$ we have $n_s = n_v^{(a)} + \delta$, which is known, given the final state in $T_\mathbf{c}$. If otherwise v is not an internal node of $T_\mathbf{c}^{[k+1]}$, all transitions into s come from a single state s' of $T_\mathbf{c}^{[k+1]}$ by Lemma 8, and we have $n_s = N_{s',s} + \delta$. In Lemma 8, s' is determined by $\overline{v}$ as the unique state $s' \preceq v$. Thus, s' is shorter than s, and, so, $N_{s',s}$ has already been computed in a previous iteration of the loop in Step 1, as the elements of R_{k+1} are taken in ascending order of length. Thus, the claim is proven, showing the validity of the computations in Steps 5 and 14.

In Step 6 we branch on whether the children of r belong to $T_{\mathbf{c}}^{[k]}$ or not. In the latter case, the algorithm skips to Step 12, and for every $s \in S_{\mathbf{c}}^{[k+1]}(r)$ we have that $s \notin T_{\mathbf{c}}^{[k]}$, and $s \notin U'_{k+1}$, since $r \notin U_{k+1}$ in Step 2. Hence, by definition of U'_{k+1}, all states $s \in S_{\mathbf{c}}^{[k+1]}(r)$ belong to $T^{[k+1]} \backslash T_{\mathbf{c}}^{[k]}$, thus $|s| = k+1$ and $\overline{as}$ is sufficiently long to determine a state in $T_{\mathbf{c}}^{[k+1]}$ for every symbol a. This validates the definition of s' in Steps 13 and 14, as well as the use of $\mathcal{N}_b(T, x^n)$ to determine $n_s^{(a)}$ in Step 13. □

We next analyze the expected length of the code defined by the Procedure EncodeTypeClass of Figure 2. Clearly, we require $|S_T|(\alpha - 1)\log n + O(1)$ bits to describe $\mathcal{N}_b(T, x^n)$, and the final state of x^n in $T_{\mathbf{c}}$, in the first step. We are interested now in the number of counts that are actually encoded by Procedure P as the algorithm iterates through the loop in Steps 3–4 of EncodeTypeClass. By carefully following the different cases managed by the algorithm, we will show that the number of counts given to refine the counters from $T_{\mathbf{c}}^{[k]}$ to $T_{\mathbf{c}}^{[k+1]}$ in Step 4 of EncodeTypeClass equals

$$\left(|E_{T_{\mathbf{c}}^{[k+1]}}| - |V_{T_{\mathbf{c}}^{[k+1]}}|\right) - \left(|E_{T_{\mathbf{c}}^{[k]}}| - |V_{T_{\mathbf{c}}^{[k]}}|\right) - (\alpha - 1)\left(|S_{T^{[k+1]}}| - |S_{T^{[k]}}|\right) , \qquad (12)$$

where we recall that $G_{T_{\mathbf{c}}^{[m]}} = \left(V_{T_{\mathbf{c}}^{[m]}}, E_{T_{\mathbf{c}}^{[m]}}\right)$ is the state transitions graph of $T_{\mathbf{c}}^{[m]}$. Adding over all the iterations in the loop of EncodeTypeClass, Equation (12) gives rise to a telescopic summation, which collapses to yield the following lemma.

Lemma 12 *The number of counts encoded by EncodeTypeClass as the algorithm iterates through the loop in Steps 3-4 is* $(|E_{T_{\mathbf{c}}}| - |V_{T_{\mathbf{c}}}|) - (\alpha - 1)|S_T|$

The full proof of Lemma 12 is presented in Appendix B. A simplified outline for canonical trees follows, which nevertheless contains most of the main ideas.

Assume that T is canonical. It is readily verified that then, all trees $T^{[h+1]} \dots T^{[d]}$ are canonical, and all nodes added to $T_{\mathbf{c}}^{[k]}$ to form $T_{\mathbf{c}}^{[k+1]}$ do in fact belong to T and have depth $k+1$. For a state s of $T_{\mathbf{c}}^{[k]}$, consider the set of edges in $E_{T_{\mathbf{c}}^{[k]}}$ and $E_{T_{\mathbf{c}}^{[k+1]}}$ that depart from s (or the children of s if it is refined in $T_{\mathbf{c}}^{[k+1]}$). Consider also the set of descendants from s in $V_{T_{\mathbf{c}}^{[k+1]}}$, i.e., $\{s\}$ when s is not refined, or $\{sb : b \in \mathcal{A}\}$ otherwise. Suppose first that s remains a state in $T_{\mathbf{c}}^{[k+1]}$. The set of edges in $E_{T_{\mathbf{c}}^{[k]}}$ that depart from s is only altered in $E_{T_{\mathbf{c}}^{[k+1]}}$ if $csu \in S_{\mathbf{c}}^{[k]}$ is refined in $T_{\mathbf{c}}^{[k+1]}$ for some u. In this case, there is an increment of $\alpha - 1$ in the number of edges from s to the children of csu in $E_{T_{\mathbf{c}}^{[k+1]}}$ with respect to the single edge from s to csu in $E_{T_{\mathbf{c}}^{[k]}}$. On the other hand the number of descendants from s in $V_{T_{\mathbf{c}}^{[k+1]}}$ is not altered with respect to $V_{T_{\mathbf{c}}^{[k]}}$ as s is not refined. Thus, we have a contribution of $\alpha - 1$ to the difference $\left(|E_{T_{\mathbf{c}}^{[k+1]}}| - |V_{T_{\mathbf{c}}^{[k+1]}}|\right) - \left(|E_{T_{\mathbf{c}}^{[k]}}| - |V_{T_{\mathbf{c}}^{[k]}}|\right)$, which is exactly the number of counts described by Procedure P in Step 9 (we rule out counts given in Step 2 of P since, by our assumption of T being canonical, only nodes at depth k are refined going from $T_{\mathbf{c}}^{[k]}$ to $T_{\mathbf{c}}^{[k+1]}$; since r and rc have different lengths, the condition in Step 1 cannot hold).

Suppose now that s is refined with a full complement of children in $T_{\mathbf{c}}^{[k+1]}$. This causes an increment of $\alpha - 1$ in the number vertices in $V_{T_{\mathbf{c}}^{[k+1]}}$ with respect to $V_{T_{\mathbf{c}}^{[k]}}$. On the other hand, since T is canonical, s must be at level k and therefore there are α edges in $E_{T_{\mathbf{c}}^{[k]}}$

departing from s, and also α edges in $E_{T_{\mathbf{c}}^{[k+1]}}$ departing from each of the α children of s. Thus, we have an increment of $\alpha^2-\alpha$ in the number of edges in $E_{T_{\mathbf{c}}^{[k+1]}}$ with respect to $E_{T_{\mathbf{c}}^{[k]}}$, and an increment of $\alpha-1$ in the number of vertices in $V_{T_{\mathbf{c}}^{[k+1]}}$ with respect to $V_{T_{\mathbf{c}}^{[k]}}$, yielding a contribution of $(\alpha-1)^2$ to the difference $\left(|E_{T_{\mathbf{c}}^{[k+1]}}| - |V_{T_{\mathbf{c}}^{[k+1]}}|\right) - \left(|E_{T_{\mathbf{c}}^{[k]}}| - |V_{T_{\mathbf{c}}^{[k]}}|\right)$. Since T is canonical, and s is refined in $T_{\mathbf{c}}^{[k+1]}$, the conditions in Steps 2 and 6 of RefineTypeClass do not hold, and all transition counts departing from the children of s are reconstructed in Steps 13 and 14. Thus, we do not need to describe counts in these cases, and since we have a total of $(|S_{T^{[k+1]}}| - |S_{T^{[k]}}|)/(\alpha-1)$ states in this situation, the negative term $-(\alpha-1)\,(|S_{T^{[k+1]}}| - |S_{T^{[k]}}|)$ in (12) arises.

EnumCodeT(T, x^n)

---

1. Encode $\mathcal{N}(T, x^n)$ using EncodeTypeClass
2. Encode the index of x^n within $\mathcal{T}(T, x^n)$

Figure 5: Universal enumerative code for tree sources

All the components of a universal enumerative code for tree sources are now in place, and the overall scheme is summarized in Figure 5. The enumeration of the type class for Step 2 is studied in [17]. By Lemma 12 and the discussion preceding Figure 2, Algorithm EncodeTypeClass for Step 1 gives an expected code length of

$$\begin{aligned} E\,|\text{EncodeTypeClass}(T, x^n)| &= \\ &= (\alpha-1)|S_T|\log n + \Big(\,(|E_{T_{\mathbf{c}}}| - |V_{T_{\mathbf{c}}}|) - (\alpha-1)|S_T|\,\Big)\frac{1}{2}\log n + O(1) \qquad (13) \\ &= \frac{1}{2}(\alpha-1)|S_T|\log n + \frac{1}{2}\,(|E_{T_{\mathbf{c}}}| - |V_{T_{\mathbf{c}}}|)\log n\,, \qquad (14) \end{aligned}$$

where the first term in (13) comes from the number of bits used to encode $\mathcal{N}_b(T, x^n)$, and the second term from SCCs*, which, by Lemma 12, are used $(|E_{T_{\mathbf{c}}}| - |V_{T_{\mathbf{c}}}|) - (\alpha-1)|S_T|$ times taking, by Corollary 2, an average of at most $\frac{1}{2}\log n + O(1)$ bits each. Moreover, a straightforward analysis of the computation shows that EncodeTypeClass can be executed in time polynomial (at most quadratic) in $|S_T|$. Normalizing, and applying Theorem 1 to bound the expected code length of Step 2, we arrive at Theorem 2 below, which summarizes the main result of the paper.

Theorem 2 *Let T be a tree source with entropy rate $\mathcal{H}$ and all conditional probabilities different from zero. Then, EnumCodeT can be efficiently implemented, and its code length, $L(X^n)$, for a random sequence X^n emitted by T satisfies*

$$E\left[\frac{L(X^n)}{n}\right] = \mathcal{H} + \frac{|S_T|(\alpha-1)\log n}{2n} + O(\frac{1}{n})\,.$$

The foregoing results yield the following corollary, which will be useful to derive an alternative enumerative coding strategy.

Corollary 4 *The type class of x^n relative to the FSM closure, T_F, of T, can be described using, on average, $\frac{1}{2}(\alpha-1)\,(|S_T| + |S_{T_F}|)\log n + O(1)$ bits.*

Proof. By Lemma 7, given a description of $\mathcal{N}(T, x^n)$, we can obtain one of $\mathcal{N}(T_F, x^n)$ with a cost of $\frac{1}{2}\kappa_T \log n + O(1)$ additional bits on average. Combining with the result of Lemma 6, and accounting for the cost of the description of $\mathcal{N}(T, x^n)$ from (14), yields the claimed result. □

The result of Corollary 4 suggests the following alternative enumerative coding strategy: To encode x^n, encode $\mathcal{N}(T_F, x^n)$ as described in the corollary, and then describe the index of x^n in an enumeration of $\mathcal{T}(T_F, x^n)$. Using the bound of Lemma 1, applied to T_F, for the expected code length of this index, we obtain an expected total normalized code length of $\mathcal{H} + \frac{1}{2}(\alpha - 1)|S_T|\frac{\log n}{n} + O(1/n)$ bits. Notice that although the enumeration is done on $\mathcal{T}(T_F, x^n)$, the expected redundancy is still optimal with respect to the smaller model T.

6 Twice-Universal Coding

In this section we switch to a twice-universal setting in which the actual tree T is unknown. Our first approach follows a conceptually simple, standard plug-in strategy in which we estimate $\hat{T}$ and then use EnumCodeT with $\hat{T}$ as if it were the true tree underlying the model. Later, we will demonstrate an alternative approach in which EnumCodeT can be greatly simplified for the twice-universal setting. We consider a class of penalized maximum likelihood tree model estimators. Specifically, given a sequence x^n, we assign to a tree T a cost $K(T, x^n) = -\log \hat{P}_T(x^n) + C(T)\log n$ where the *penalization function* $C(T)$ is increasing with $|S_T|$. We have

$$K(T, x^n) = -\sum_{s \in S_T, a \in \mathcal{A}} n_s^{(a)} \log \frac{n_s^{(a)}}{n_s} + C(T)\log n\,.$$

The tree model estimate $\hat{T}(x^n)$ for x^n is defined as the tree that minimizes the cost function $K(T, x^n)$ over all possible trees, that is,

$$\hat{T}(x^n) = \arg\min_T \{K(T, x^n)\}\,. \tag{15}$$

Efficient algorithms are known for finding the minimizing tree $\hat{T}(x^n)$; see, e.g., [20].

We define the code *Twice-EnumCodeT* algorithmically in Figure 6.

Twice-EnumCodeT(x^n)

1. Compute the estimate $\hat{T}(x^n)$ of T.
2. Describe $\hat{T}$ to the decoder.
3. Encode x^n using EnumCodeT with respect to the tree $\hat{T}$.

Figure 6: Twice universal enumerative code for tree sources

Using a natural code [20] for describing the full tree $\hat{T}(x^n)$, Step 2 of Twice-EnumCodeT requires one bit per node. To estimate the cost of Step 3, we must analyze the code length of EnumCodeT when applied to $\hat{T}(x^n)$ rather than T. The analysis will rely on upper bounds on the probabilities of over-estimation and under-estimation of T, which are stated in the two lemmas below. Similar bounds are well known for several estimators, and proofs for the lemmas can be readily adapted from [13].

Lemma 13 *Let T be a tree source and consider a penalization function of the form $C(T) = \beta|S_T|$ with $\beta > \frac{\alpha(\alpha+1)+1}{\alpha-1}$. Let $O^n \subset \mathcal{A}^n$ be the set of strings for which a state of T is refined by the estimated tree $\hat{T}$. Then $P_T\{O^n\} \leq |S_T| n^{-\gamma}$ with $\gamma = \beta(\alpha-1) - \alpha(1+\alpha) - 1 > 0$.*

Lemma 14 *Let (T, P_T) be a minimal tree model and consider a penalization function of the form $C(T) = \beta|S_T|$. Let $U^n \subset \mathcal{A}^n$ be the set of sequences whose estimated tree $\hat{T}$ has a state that is refined by T. Then $P_T\{U^n\} \leq R 2^{-nD}$ for positive constants R, D, and sufficiently large n .*

From Lemma 13 and Lemma 14 it follows that we can choose β to make the contribution of sequences with estimated tree $\hat{T} \neq T$ to the expected code length negligible, as long as the code length is upper-bounded by a polynomial in n. We verify this fact next. In Twice-EnumCodeT we describe $\mathcal{N}(\hat{T}, x^n)$ by encoding the final state of x^n with respect to $\hat{T}_{\mathbf{c}}$, encoding $\mathcal{N}_b(\hat{T}, x^n)$ with $|S_{\hat{T}}|(\alpha-1)$ counts of $\log n$ bits each, and finally giving an additional set of $|E_{\hat{T}_{\mathbf{c}}}| - |V_{\hat{T}_{\mathbf{c}}}| - |S_{\hat{T}}|(\alpha-1)$ counts described with SCCs*, which take $O(\sqrt{n})$ bits each. Thus, the complete description of $\mathcal{N}(\hat{T}, x^n)$ takes $O\left(\left(|E_{\hat{T}_{\mathbf{c}}}| - |V_{\hat{T}_{\mathbf{c}}}|\right)\sqrt{n}\right)$ bits. From the definition of a forgetful state, it is readily verified that $|E_{\hat{T}_{\mathbf{c}}}| - |V_{\hat{T}_{\mathbf{c}}}| \leq |E_{\hat{T}}| - |V_{\hat{T}}|$, and from the definition of $|E_{\hat{T}}|$ it is not difficult to see that $|E_{\hat{T}}| = O(|S_{\hat{T}}|)$. Hence, the cost of describing the type class of x^n with respect to $\hat{T}$ is $O(|S_{\hat{T}}|\sqrt{n})$. Since the index of x^n within its class takes no more than n bits, we upper-bound the total code length of Twice-EnumCodeT by $O\left(|S_{\hat{T}}|\sqrt{n} + n\right)$. Finally, noticing that we must have $|S_{\hat{T}}| = O(n/\log n)$, since otherwise $\hat{T}$ would be dominated by a single-state tree in (15), we obtain the desired polynomial bound on the total code length, leading to the following result.

Theorem 3 *Let T be a tree source with entropy rate $\mathcal{H}$ and with all conditional probabilities different from zero. Taking a penalization function $C(T) = \beta|S_T|$ with β sufficiently large, the normalized expected code length of Twice-EnumCodeT is*

$$E\left[\frac{L(X^n)}{n}\right] = \mathcal{H} + \frac{|S_T|(\alpha-1)\log n}{2n} + O(1/n)\,.$$

In the rest of the section we present an alternative code EnumCodeT$'$ which is a simplification of EnumCodeT, applicable when the target tree is an estimate $\hat{T}$, as in Step 3 of Twice-EnumCodeT. Recall that a fundamental tool in EnumCodeT is the use of SCCs* codes, a generalization of SCCs. The latter rely on the quantities $\Delta(w, s, a) = \left|n_w^{(a)} - \frac{n_s^{(a)}}{n_s} n_w\right| n_w^{-\frac{1}{2}}$, for properly defined strings s, w and symbol a, being small with high probability. In other words, sequences with large values $\Delta(w, s, a)$ have small probability under *any* model parameter, and we would expect an estimate $\hat{T}(x^n) \neq T$ for such sequences. The following lemma formalizes these claims.

Lemma 15 *Let $\hat{T} \triangleq \hat{T}(x^n)$ be a tree model estimate for x^n, and let s and w be strings such that $s \in S_{\hat{T}}$, $s \prec w$, and $n_s > 0$. Then, for any refinement T' of $\hat{T}$ that contains w, we have*

$$\left|n_w^{(a)} - \frac{n_s^{(a)}}{n_s} n_w\right| \leq \sqrt{2(C(T') - C(\hat{T}))n_s \ln n}\,.$$

Proof.

Since $C(T)$ is increasing in the number of states of T, it is sufficient to consider the case in which T' is the smallest refinement of $\hat{T}(x^n)$ that contains w, i.e., the tree that results from refining $\hat{T}(x^n)$ by adding $w'b$ for all proper prefixes w' of w and all symbols $b \in \mathcal{A}$. Let $W = \{su : su \in S_{T'}\}$. Since $\hat{T}(x^n)$ minimizes the cost function $K(T, x^n)$, we have

$$-\sum_{t\in S_{\hat{T}},a\in\mathcal{A}} n_t^{(a)} \log \frac{n_t^{(a)}}{n_t} + C(\hat{T}) \log n \leq -\sum_{t\in S_{T'},a\in\mathcal{A}} n_t^{(a)} \log \frac{n_t^{(a)}}{n_t} + C(T') \log n\,. \tag{16}$$

Therefore,

$$-\sum_{t\in S_{\hat{T}},a\in\mathcal{A}} n_t^{(a)} \log \frac{n_t^{(a)}}{n_t} + \sum_{t\in S_{T'},a\in\mathcal{A}} n_t^{(a)} \log \frac{n_t^{(a)}}{n_t} \leq (C(T') - C(\hat{T})) \log n\,, \tag{17}$$

which reduces to

$$-\sum_{a\in\mathcal{A}} n_s^{(a)} \log \frac{n_s^{(a)}}{n_s} + \sum_{su\in W,a\in\mathcal{A}} n_{su}^{(a)} \log \frac{n_{su}^{(a)}}{n_{su}} \leq (C(T') - C(\hat{T})) \log n\,. \tag{18}$$

Since $n_s > 0$, we further obtain

$$-\sum_{a\in\mathcal{A}} \frac{n_s^{(a)}}{n_s} \log \frac{n_s^{(a)}}{n_s} + \sum_{su\in W} \frac{n_{su}}{n_s} \sum_{a\in\mathcal{A}} \frac{n_{su}^{(a)}}{n_{su}} \log \frac{n_{su}^{(a)}}{n_{su}} \leq (C(T') - C(\hat{T})) \frac{\log n}{n_s}\,. \tag{19}$$

Let $\hat{p}(\cdot|s)$ be the probability mass function (PMF) over $\mathcal{A}$ given by $\hat{p}(a|s) = \frac{n_s^{(a)}}{n_s}$ and analogously for $su \in W$, $\hat{p}(a|su) = \frac{n_{su}^{(a)}}{n_{su}}$. Consider also a PMF $\hat{p}(\cdot)$ over W given by $\hat{p}(su) = \frac{n_{su}}{n_s}$. Let X, Y be random variables such that Y takes values in W with $Y \sim \hat{p}(\cdot)$, and X takes values in $\mathcal{A}$ with conditional distribution $P(X = a|Y = su) = \hat{p}(a|su)$. Then, the marginal distribution of X is $X \sim \hat{p}(\cdot|s)$, and the joint distribution of X and Y is

$$P(X = a, Y = su) = P(X = a|Y = su)P(Y = su) = \hat{p}(a|su)\hat{p}(su) = \frac{n_{su}^{(a)}}{n_{su}} \frac{n_{su}}{n_s} = \frac{n_{su}^{(a)}}{n_s}\,.$$

From Equation (19),

$$I(X;Y) = H(X) - H(X|Y) \leq (C(T') - C(\hat{T})) \frac{\log n}{n_s}\,.$$

Let Q be a joint distribution given by the product of the marginal distributions of X, Y, i.e., $Q(X = a, Y = su) = P(X = a)P(Y = su) = \frac{n_s^{(a)}}{n_s} \frac{n_{su}}{n_s}$. Then, by Pinsker's inequality [21, Lemma 12.6.1], we have

$$\frac{1}{2\ln 2} \|P - Q\|_1^2 \leq D(P||Q) = I(X;Y) \leq (C(T') - C(\hat{T})) \frac{\log n}{n_s}\,. \tag{20}$$

Therefore,

$$\left(\sum_{a\in\mathcal{A},su\in W} |P(a, su) - Q(a, su)| \right)^2 \leq 2(C(T') - C(\hat{T})) \frac{\ln n}{n_s}\,, \tag{21}$$

which takes the form

$$\sum_{a\in\mathcal{A},su\in W}\left|\frac{n_{su}^{(a)}}{n_s}-\frac{n_s^{(a)}}{n_s}\frac{n_{su}}{n_s}\right|\le\sqrt{2(C(T')-C(\hat{T}))\frac{\ln n}{n_s}}\,. \tag{22}$$

In particular, taking only the term corresponding to $su=w$ in the summation on the left hand side of (22), we conclude that

$$\left|n_w^{(a)}-\frac{n_s^{(a)}}{n_s}n_w\right|\le\sqrt{2(C(T')-C(\hat{T}))n_s\ln n}\,. \tag{23}$$

□

With a linear penalization function of the form $C(T)=\beta|S_T|$, we have $C(T')-C(\hat{T})=\beta(|S_{T'}|-|S_{\hat{T}}|)\le\beta\alpha|w|$, implying the following corollary.

Corollary 5 *Let $\hat{T}(x^n)$ be a tree model estimate for x^n with a penalization function $C(T)=\beta|S_T|$, $s\in S_{\hat{T}}$, $s\prec w$, and $n_s>0$. Then,*

$$|z_{w,a}|=\left|n_w^{(a)}-\frac{n_s^{(a)}}{n_s}n_w\right|\le\sqrt{\beta\alpha|w|n_s\ln n}\,.$$

It follows from Corollary 5 that, when considering coding with respect to a tree $\hat{T}$ estimated with a linear penalization function, it may be advantageous to replace the use of SCCs with a uniform coding of $z_{w,a}$ in the range $\left(-\sqrt{\beta\alpha|w|n_s\ln n},\sqrt{\beta\alpha|w|n_s\ln n}\right)$. The code length obtained would be $\frac{1}{2}\log n_s+o(\log n)$, which is of similar main order as the expected code length of SCCs (cf. Corollary 1). Notice, however, that the upper bound here is *pointwise*, and not just in expectation. The same idea can be generalized to SCCs* by means of Lemma 16 below, for which we recall the definitions from (11).

Lemma 16 *Let $\hat{T}$ be a tree model estimate for x^n with a penalization function $C(T)=\beta|S_T|$, and let u^q be a string such that an SCC* code is applicable in $\hat{T}$. Then,*

$$|z_u|=\left|n_{\overline{u}}-m_{l+1}\prod_{i=l+1}^{q-1}\frac{n_i^\alpha}{n_i}\right|\le q^{3/2}\sqrt{\beta\alpha n\ln n}\,. \tag{24}$$

Proof. We prove the result by repeatedly applying Corollary 5 as follows. Notice that $m_i=n_{\frac{(u_i)}{u^{i-1}}}$. Then, by Corollary 5, we have $\left|m_i-\frac{n_{i-1}^\alpha}{n_{i-1}}m_{i-1}\right|\le\sqrt{\beta\alpha|u|n\ln n}$ for all $l<i\le q$ (we extend the definition of m_i for $i=q$ as $m_q=n_{\overline{u^q}}$). Applying Corollary 5 again and grouping terms, we further obtain $\left|m_i-\frac{n_{i-1}^\alpha}{n_{i-1}}\frac{n_{i-2}^\alpha}{n_{i-2}}m_{i-2}\right|\le\left(1+\frac{n_{i-1}^\alpha}{n_{i-1}}\right)\sqrt{\beta\alpha|u|n\ln n}$. Starting from $m_i=m_q=n_{\overline{u^q}}$, and repeatedly applying the same arguments, we finally bound $|z_u|$ and the claim of the lemma follows readily. □

Using a uniform encoding for the numbers z_u leads, by (24), to an analogue of Corollary 2 in the twice-universal setting. As before, we note that the bound here is pointwise, as opposed to in expectation as in Corollary 2. The results take advantage of the idea suggested in Section 1, namely, that for some type classes of $\hat{T}$, no sequence in the

class will estimate $\hat{T}$. Therefore, these "atypical" classes can be excluded from the coding space. To implement this idea, we define the code EnumCodeT$'$ exactly as EnumCodeT, but replacing the use of SCCs*, $C_u^\star(n_{\overline{u}})$, by a uniform encoding of z_u in the range $\left(-|u|^{3/2}\sqrt{\beta\alpha n \ln n}, |u|^{3/2}\sqrt{\beta\alpha n \ln n}\right)$. In EnumCodeT, SCCs* are applied to strings u that are prefixes of states of the canonical extension of the tree. Hence, $|u|$ is bounded by the depth of $\hat{T}(x^n)$. For sequences that estimate $\hat{T}(x^n) = T$, $|u|$ is bounded by the depth of T, and the uniform encoding of z_u, by (24), takes $\frac{1}{2}\log n + O(\log\log n)$ bits. Thus, when $\hat{T}(x^n) = T$, the upper bounds of the expected code lengths of EnumCodeT$'$ and EnumCodeT differ by $O(\log\log n)$ bits. As a result, using essentially the same arguments as in Theorem 3, we can prove the following theorem for *Twice-EnumCodeT$'$*, the code obtained by substituting EnumCodeT$'$ for EnumCodeT in Twice-EnumCodeT.

Theorem 4 *Let T be a tree source with entropy rate $\mathcal{H}$ and with all conditional probabilities nonzero. Estimating $\hat{T}(x^n)$ with a penalization function $C(T) = \beta|S_T|$, with β sufficiently large, the normalized expected code length of Twice-EnumCodeT$'$ is*

$$E\left[\frac{L(X^n)}{n}\right] = \mathcal{H} + \frac{|S_T|(\alpha-1)\log n}{2n} + O\left(\frac{\log\log n}{n}\right).$$

A Proof of Lemma 10

Proof. Suppose the claim of Lemma 10 is not true and let $z \in \mathcal{A}^*$, $c \in \mathcal{A}$ such that $zc \in U_{k+1}$ is of maximal length among those elements of U_{k+1} which are not leaves of $T_{\mathbf{c}}^{[k]}$. Since the children of zc were added to $T_{\mathbf{c}}^{[k+1]}$ for zc was forgetful, we know that azc is an internal node of $T_{\mathbf{c}}^{[k+1]}$ for every $a \in \mathcal{A}$. If $azc \in U_{k+1}$, azc is a leaf of $T_{\mathbf{c}}^{[k]}$ for azc is longer than zc. Therefore, az is an internal node of $T_{\mathbf{c}}^{[k]}$. If $azc \notin U_{k+1}$, $azcd \notin U'_{k+1}$ for any $d \in \mathcal{A}$. Hence, by definition of U'_{k+1}, either $azcd \in T_{\mathbf{c}}^{[k]}$ or $azcd \in T^{[k+1]}$. In any case we have that $azc \in T_{\mathbf{c}}^{[k]}$, i.e., az is an internal node of $T_{\mathbf{c}}^{[k]}$. We conclude that az is an internal node of $T_{\mathbf{c}}^{[k]}$ for every $a \in \mathcal{A}$, thus $zc \in T_{\mathbf{c}}^{[k]}$ by definition of canonical tree. Further, since $zc \in U_{k+1}$, $zcd \in U'_{k+1}$ for some $d \in \mathcal{A}$, thus $zcd \notin T_{\mathbf{c}}^{[k]}$ and zc is a state of $T_{\mathbf{c}}^{[k]}$, a contradiction. □

B Proof of Lemma 12

Proof. We will equate the number of counts given in each iteration of the loop to the increment in $|E_{T_{\mathbf{c}}^{[k+1]}}| - |V_{T_{\mathbf{c}}^{[k+1]}}|$ with respect to $|E_{T_{\mathbf{c}}^{[k]}}| - |V_{T_{\mathbf{c}}^{[k]}}|$. For $u \in \mathcal{A}^*$ we define $V_i(u) = \{uv \in S_{\mathbf{c}}^{[i]} : v \in \mathcal{A} \cup \{\lambda\}\}$, $A_i(u) = \{(uv, w) \in E_{T_{\mathbf{c}}^{[i]}} : v \in \mathcal{A} \cup \{\lambda\}\}$ and, for $a \in \mathcal{A}$, $A_i^{(a)}(u) = \{(uv, w) \in A_i(u) : a = \mathrm{hd}(w)\}$. Notice that since $T_{\mathbf{c}}^{[k+1]}$ refines states of $T_{\mathbf{c}}^{[k]}$ in at most one level, $A_k(r)$ and $V_k(r)$ exhaust $E_{T_{\mathbf{c}}^{[k]}}$ and $V_{T_{\mathbf{c}}^{[k]}}$ respectively as r varies along R_{k+1}. Of course, $A_{k+1}(r)$ and $V_{k+1}(r)$ do also exhaust $E_{T_{\mathbf{c}}^{[k+1]}}$ and $V_{T_{\mathbf{c}}^{[k+1]}}$ respectively as r varies along R_{k+1}.

We claim that the number of counts given by an invocation to $\mathrm{P}(r, c)$ equals $|A_{k+1}^{(c)}(r)| - |A_k^{(c)}(r)|$. When the condition of Step 1 holds true, we describe $\alpha - 1$ counts. There are α edges from children of r to children of cr in $A_{k+1}(r)$ and one edge from r to cr in $A_k(r)$, i.e., the number of given counts coincides with the increment $|A_{k+1}^{(c)}(r)| - |A_k^{(c)}(r)|$. Now, when

the condition of Step 5 is satisfied, we have for each $csu \in W$, that there is an increment of $\alpha-1$ in the number of edges that depart from s to children of csu in $A_{k+1}(r)$ with respect to the one single edge from $\sigma_{\mathbf{c}}^{[k]}(\overline{s})$ to csu in $A_k(r)$. The increment coincides with the number of counts given in Step 9. On the other hand, in Step 12, where csv is a state of $T_{\mathbf{c}}^{[k+1]}$ which is not in W', csv is also a state of $T_{\mathbf{c}}^{[k]}$. There is one edge from s to csv in $A_{k+1}(r)$ and also one edge from $\sigma_{\mathbf{c}}^{[k]}(\overline{s})$ to csv in $A_k(r)$. Thus, there is no increment in the number of edges. Finally, in Step 14, we have that cs is a leaf of $T_{\mathbf{c}}^{[k+1]}$. Then, as mentioned, either $cs \in T_{\mathbf{c}}^{[k]}$ or $s \in T_{\mathbf{c}}^{[k]}$, for otherwise, by Corollary 3, their parents cr, r, would belong to $T_{\mathbf{c}}^{[k]}$ and the condition of Step 1 would hold true. When $cs \in T_{\mathbf{c}}^{[k]}$, there is one edge from s to cs in $A_{k+1}(r)$ and also one edge from $\sigma_{\mathbf{c}}^{[k]}(\overline{s})$ to cs in $A_k(r)$. If, on the other hand, $s \in T_{\mathbf{c}}^{[k]}$, there is one edge from s to cs in $A_{k+1}(r)$ and also one edge from s to $\sigma_{\mathbf{c}}^{[k]}(\overline{cs})$ in $A_k(r)$. The claim is proved.

We now analyze RefineTypeClass. When $r \in U_{k+1}$, the number of counts described in Step 4 is $|A_{k+1}(r) \backslash A_{k+1}^{(d)}(r)| - |A_k(r) \backslash A_k^{(d)}(r)|$. For the symbols $d \in \mathcal{A}$ of Step 3, we have that dr is not an internal node of $T_{\mathbf{c}}^{[k]}$ but, since $r \in U_{k+1}$, dr is an internal node of $T_{\mathbf{c}}^{[k+1]}$. Then, dr is a leaf of $T_{\mathbf{c}}^{[k]}$ and the full set of children of dr are leaves of $T_{\mathbf{c}}^{[k+1]}$. There are α edges from the children of r to the children of dr in $A_{k+1}^{(d)}(r)$ and one single edge from r to dr in $A_k^{(d)}(r)$. Hence, the number of counts described in Step 4 is $|A_{k+1}(r)| - |A_k(r)| - (\alpha - 1)$. Since $|V_{k+1}(r)| - |V_k(r)| = \alpha - 1$ we have that the total number of counts described is $(|A_{k+1}(r)| - |V_{k+1}(r)|) - (|A_k(r)| - |V_k(r)|)$. We now consider the case where $r \notin U_{k+1}$ and the children of r belong to $T_{\mathbf{c}}^{[k]}$. When $cs \notin T_{\mathbf{c}}^{[k+1]}$, the decoder computes state transition counts in Step 9. In this case, for every state $s \in S_{\mathbf{c}}^{[k+1]}$ child of r, there is one edge from s to $s' = \sigma_{\mathbf{c}}^{[k+1]}(\overline{cs})$ in $A_{k+1}(r)$ and, since also $s \in S_{\mathbf{c}}^{[k]}$, there is also one edge from $\sigma_{\mathbf{c}}^{[k]}(\overline{s})$ to $\sigma_{\mathbf{c}}^{[k]}(\overline{cs})$ in $A_k(r)$. Hence, $|A_{k+1}^{(c)}(r)| = |A_k^{(c)}(r)|$. For the remaining values of c, we use $|A_{k+1}^{(c)}(r)| - |A_k^{(c)}(r)|$ counts in Step 11. Since the children of r belong to $T_{\mathbf{c}}^{[k]}$, we have $|V_{k+1}(r)| = |V_k(r)|$. Thus, the total number of counts is $(|A_{k+1}(r)| - |V_{k+1}(r)|) - (|A_k(r)| - |V_k(r)|)$. Finally, when the algorithm skips to Step 12, we have that all states s which are children of r do not belong to $T_{\mathbf{c}}^{[k]}$, and do not belong to U'_{k+1}, for $r \notin U_{k+1}$ in Step 2. Hence, by the definition of U'_{k+1}, all states s which are children of r belong to $T^{[k+1]} \backslash T_{\mathbf{c}}^{[k]}$ and, thus, $|s| = k+1$ and $\overline{as}$ is sufficiently long to determine a state in $T_{\mathbf{c}}^{[k+1]}$ for every symbol a. There are α edges in $A_{k+1}(r)$ departing from each of the α children of r, for a total of α^2 edges in $A_{k+1}(r)$. On the other hand, r is a leaf of maximal length in $T_{\mathbf{c}}^{[k]}$ and therefore there are α edges departing from r in $A_k(r)$. Since $|V_{k+1}(r)| = \alpha$ and $|V_k(r)| = 1$, we have $(|A_{k+1}(r)| - |V_{k+1}(r)|) - (|A_k(r)| - |V_k(r)|) = (\alpha - 1)^2$.

Over all, the number of counts described is $\left(|E_{T_{\mathbf{c}}^{[k+1]}}| - |V_{T_{\mathbf{c}}^{[k+1]}}|\right) - \left(|E_{T_{\mathbf{c}}^{[k]}}| - |V_{T_{\mathbf{c}}^{[k]}}|\right) - B_{k+1}(\alpha-1)^2$ where B_{k+1} is the number of elements r in R_{k+1} with $|r| = k$. Clearly, the children in $T_{\mathbf{c}}^{[k+1]}$ of such elements of R_{k+1} are the nodes in $T^{[k+1]} \backslash T^{[k]}$ and we can write $B_{k+1} = \left(|S_{T^{[k+1]}}| - |S_{T^{[k]}}|\right)/(\alpha - 1)$. It follows that the number of counts described can be written as $\left(|E_{T_{\mathbf{c}}^{[k+1]}}| - |V_{T_{\mathbf{c}}^{[k+1]}}|\right) - \left(|E_{T_{\mathbf{c}}^{[k]}}| - |V_{T_{\mathbf{c}}^{[k]}}|\right) - (\alpha - 1)\left(|S_{T^{[k+1]}}| - |S_{T^{[k]}}|\right)$.

The total number of counts described by EncodeTypeClass is, therefore,

$$\sum_{k=h+1}^{d-1} \left(|E_{T_{\mathbf{c}}^{[k+1]}}| - |V_{T_{\mathbf{c}}^{[k+1]}}|\right) - \left(|E_{T_{\mathbf{c}}^{[k]}}| - |V_{T_{\mathbf{c}}^{[k]}}|\right) - (\alpha - 1)\left(|S_{T^{[k+1]}}| - |S_{T^{[k]}}|\right) , \qquad (25)$$

where we recall that $d = \max\{|s| : s \in S_T\}$ and $h = \min\{|s| : s \in S_T\}$. The telescopic sum in (25) reduces to

$$(|E_{T_{\mathbf{c}}}| - |V_{T_{\mathbf{c}}}|) - \left(|E_{T_{\mathbf{c}}^{[h+1]}}| - |V_{T_{\mathbf{c}}^{[h+1]}}|\right) - (\alpha - 1)\left(|S_T| - |S_{T^{[h+1]}}|\right) . \tag{26}$$

Now, since $T^{[h+1]}$ is FSM, $V_{T_{\mathbf{c}}^{[h+1]}} = V_{T^{[h+1]}} = S_{T^{[h+1]}}$, $E_{T_{\mathbf{c}}^{[h+1]}} = E_{T^{[h+1]}}$, and $|E_{T^{[h+1]}}| = \alpha|S_{T^{[h+1]}}|$, so that $|E_{T_{\mathbf{c}}^{[h+1]}}| - |V_{T_{\mathbf{c}}^{[h+1]}}| = (\alpha - 1)|S_{T^{[h+1]}}|$ and (26) becomes

$$(|E_{T_{\mathbf{c}}}| - |V_{T_{\mathbf{c}}}|) - (\alpha - 1)|S_T| . \tag{27}$$

□

References

[1] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Trans. Inform. Theory*, vol. 30, pp. 629–636, July 1984.

[2] Y. M. Shtarkov, "Universal sequential coding of single messages," *Problems of Inform. Trans.*, vol. 23, pp. 175–186, July 1987.

[3] J. Rissanen, "Fisher information and stochastic complexity," *IEEE Trans. Inform. Theory*, vol. 42, no. 1, pp. 40–47, 1996.

[4] T. M. Cover, "Enumerative source encoding," *IEEE Trans. Inform. Theory*, vol. 19, pp. 73–77, Jan 1973.

[5] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic, 1981.

[6] R. E. Krichevskii and V. K. Trofimov, "The performance of universal encoding," *IEEE Trans. Inform. Theory*, vol. 27, pp. 199–207, Mar 1981.

[7] J. Takeuchi and A. R. Barron, "Asymptotically minimax regret by Bayes mixtures," in *Proc. 1998 International Symposium on Information Theory*, Cambridge, MA, U.S.A., Aug. 1998, p. 318.

[8] J. Takeuchi and T. Kawabata, "Exponential curvature of Markov models," in *Proc. 2007 International Symposium on Information Theory*, Nice, France, June 2007.

[9] N. Merhav and M. J. Weinberger, "On universal simulation of information sources using training data," *IEEE Trans. Inform. Theory*, vol. 50, Jan. 2004.

[10] R. B. Ash, *Information Theory*. Wiley, 1967.

[11] M. J. Weinberger, N. Merhav, and M. Feder, "Optimal sequential probability assignment for individual sequences," *IEEE Trans. Inform. Theory*, vol. 40, pp. 384–396, Mar. 1994.

[12] J. Rissanen, "Complexity of strings in the class of Markov sources," *IEEE Trans. Inform. Theory*, vol. 32, pp. 526–532, July 1986.

[13] M. J. Weinberger, J. Rissanen, and M. Feder, "A universal finite memory source," *IEEE Trans. Inform. Theory*, vol. 41, pp. 643–652, May 1995.

[14] F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens, "The context-tree weighting method: Basic properties," *IEEE Trans. Inform. Theory*, vol. 41, pp. 653–664, May 1995.

[15] Á. Martín, G. Seroussi, and M. J. Weinberger, "Linear time universal coding and time reversal of tree sources via fsm closure," *IEEE Trans. Inform. Theory*, vol. 50, no. 7, pp. 1442–1468, July 2004.

[16] P. L. Buhlmann and A. Wyner, "Variable length Markov chains," *Annals of Statistics*, vol. 27, pp. 480–513, 1998.

[17] Á. Martín, G. Seroussi, and M. J. Weinberger, "Type classes of tree models," in *Proc. 2007 International Symposium on Information Theory*, Nice, France, June 2007, full article in preparation.

[18] S. W. Golomb, "Run-length encodings," *IEEE Trans. Inform. Theory*, vol. 12, pp. 399–401, July 1966.

[19] I. Kontoyiannis, L. Lastras-Montano, and S. Meyn, "Relative entropy and exponential deviation bounds for general Markov chains," in *Proc. 2005 International Symposium on Information Theory*, Adelaide, Australia, Sept. 2005, pp. 1563– 1567.

[20] R. Nohre, "Some topics in descriptive complexity," Ph.D. dissertation, Department of Computer Science, The Technical University of Linkoping, Sweden, 1994.

[21] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* New York: Wiley, 1991.

Stochastic chains with memory of variable length

A. Galves* E. Löcherbach

April 9, 2008

Dedicated to Jorma Rissanen on his 75'th birthday

Abstract

Stochastic chains with memory of variable length constitute an interesting family of stochastic chains of infinite order on a finite alphabet. The idea is that for each past, only a finite suffix of the past, called *context*, is enough to predict the next symbol. These models were first introduced in the information theory literature by Rissanen (1983) as a universal tool to perform data compression. Recently, they have been used to model up scientific data in areas as different as biology, linguistics and music. This paper presents a personal introductory guide to this class of models focusing on the algorithm Context and its rate of convergence.

1 Introduction

Chains with memory of variable length appear in Rissanen's 1983 paper called *A universal data compression system*. His idea was to model a string of symbols as a realization of a stochastic chain where the length of the memory needed to predict the next symbol is not fixed, but is a deterministic function of the string of the past symbols.

Considering a memory of variable length is a practical way to overcome the well known difficulty of the exponentially growing number of parameters which are needed to describe a Markov chain when its order increases. However if one wants to fit accurately complex data using a Markov chain of fixed order, one has to use a very high order. And this means that to estimate the parameters of the model we need huge samples, which makes this approach unsuitable for many practical issues.

It turns out that in many important scientific data, the length of the relevant portion of the past is not fixed, on the contrary it depends on the past. For instance, in molecular biology, the translation of a gene into a protein is initiated by a fixed specific sequence of nucleotide bases called *start codon*. In other words, the start codon designs the end of the relevant portion of the past to be considered in the translation.

The same phenomenon appears in other scientific domains. For instance in linguistics, both in phonology and in syntax, there is the notion of domains in which the grammar operates to define admissible strings of forthcoming symbols. In other terms, the boundary of the linguistic domain defines the relevant part of the past for the processing of the next linguistic units.

*This work is part of PRONEX/FAPESP's project *Stochastic behavior, critical phenomena and rhythmic pattern identification in natural languages* (grant number 03/09930-9), CNPq's project *Stochastic modeling of speech* (grant number 475177/2004-5) and CNRS-FAPESP project *Probabilistic phonology of rhythm*. AG is partially supported by a CNPq fellowship (grant 308656/2005-9).

Rissanen's ingenious idea was to construct a stochastic model that generalizes this notion of relevant domain to any kind of symbolic strings.

To be more precise, Rissanen (1983) called *context* the relevant part of the past. The stochastic model is defined by the set of all contexts and an associated family of transition probabilities.

Models with memory of variable length are not only less expensive than the classical fixed order Markov chains, but also much more clever since they take into account the structural dependencies present in the data. This is precisely what the set of contexts expresses.

Rissanen has introduced models having memory of variable length as a universal system of data compression. His goal was to compress in real time a string of symbols generated by an unknown source. To do this, we have to estimate at each step the length of the context of the string observed until that time step, as well as the associated transition probabilities.

If we knew the contexts, then the estimation of the associated transition probabilities could be done using a classical procedure such as maximum likelihood estimation. Therefore, the main point is to put hands on the context length. In his seminal 1983 paper, Rissanen solved this problem by introducing the algorithm *Context*. This algorithm estimates in a consistent way both the length of the context as well as the associated transition probability.

The class of models with memory of variable length raises interesting questions from the point of view of statistics. Examples are the rate of convergence and the fluctuations of the algorithm Context and other estimators of the model. Another challenging question would be how to produce a robust version of the algorithm Context.

But also from the point of view of probability theory, this class of models is interesting. In effect, if the length of the contexts is not bounded, then chains with memory of variable length are chains of infinite order. Existence, uniqueness, phase-transitions, perfect simulation are deep mathematical questions that should be addressed to in this new and challenging class of models.

Last but not least, models of variable length revealed to be very performing tools in applied statistics, by achieving in an efficient way classification tasks in proteomics, genomics, linguistics, classification of musical styles, and much more.

In what follows we present a personal introductory guide to this class of models with no attempt to give a complete survey of the subject. We will mainly focus on the algorithm Context and present some recent results obtained by our research team.

2 Probabilistic context trees

In what follows A will represent a finite alphabet of size $|A|$. Given two integers $m \leq n$, we will denote by x_m^n the sequence $(x_m, \ldots, x_n)$ of symbols in A. Let A_+^* be the set of all finite sequences, that is

$$A_+^* = \bigcup_{k=1}^{\infty} A^{\{-k,\ldots,-1\}}.$$

We shall write $\underline{\mathrm{A}} = A^{\{\ldots,-n,\ldots,-2,-1\}}$, and denote by $x_{-\infty}^{-1}$ any element of $\underline{\mathrm{A}}$.

Our main object of interest is what we shall call *context length function*.

Definition 2.1 *A context length function l is a function $l : A_+^* \to \{1, 2, \ldots\} \cup \{\infty\}$ satisfying the following two properties.*
(i) For any $k \geq 1$, for any $x_{-k}^{-1} \in A_+^*$, we have

$$l(x_{-k}^{-1}) \in \{1, \ldots, k\} \cup \{+\infty\}.$$

(ii) For any $x_{-\infty}^{-1} \in \underline{A}$, *if* $l(x_{-k}^{-1}) = k$ *for some* $k \geq 1$, *then*

$$\begin{aligned} l(x_{-i}^{-1}) &= \infty, \quad \text{for any } i < k \\ l(x_{-i}^{-1}) &= k, \quad \text{for any } i > k. \end{aligned}$$

Intuitively, given a sequence $x_{-\infty}^{-1}$, the function l tells us, at which position in the past we can stop since we have reached the end of the context. The first condition is a kind of adaptivity condition. It tells us that we can decide whether the end of the context has already been reached at step k just by inspecting the past sequence up to that step. If l equals $+\infty$, we have to look further back in the past. The second condition is a consistency condition. It tells us that once we have reached the bound of the context, we do not have to look further back in the past, and that the context of a longer sequence x_{-i}^{-1}, $i > k$, is also the context of x_{-k}^{-1}. In other terms, once the identification of the context is made at a given step k, this decision will not be changed by any further data present in the past before k.

By abuse of notation, we shall also call l the natural extension of the context length function to $\underline{A}$, given by

$$l(x_{-\infty}^{-1}) = \inf \{k \geq 1 : \, l(x_{-k}^{-1}) < +\infty\},$$

with the convention that $\inf \emptyset = +\infty$.

Definition 2.2 *For any* $x_{-\infty}^{-1} \in \underline{A}$, *we shall call* $x_{-l(x_{-\infty}^{-1})}^{-1}$ *the context associated to* l *of the infinite sequence* $x_{-\infty}^{-1}$.

Definition 2.3 *Let* l *be a given context length function. A stationary stochastic chain* $(X_n)_{n \in \mathbb{Z}}$ *taking values in* A *is a chain having memory of variable length, if for any infinite past* $x_{-\infty}^{-1} \in \underline{A}$ *and any symbol* $a \in A$, *we have*

$$P\left(X_0 = a | X_{-\infty}^{-1} = x_{-\infty}^{-1}\right) = P\left(X_0 = a | X_{-l(x_{-\infty}^{-1})}^{-1} = x_{-l(x_{-\infty}^{-1})}^{-1}\right). \tag{2.1}$$

We shall use the short hand notation

$$p(a|x_{-k}^{-1}) = P\left(X_0 = a | X_{-k}^{-1} = x_{-k}^{-1}\right). \tag{2.2}$$

We are mainly interested in those values of $p(a|x_{-k}^{-1})$ *where* $k = l(x_{-\infty}^{-1})$.

Observe that the set $\{\ell(X_{-\infty}^{-1}) = k\}$ is measurable with respect to the σ-algebra generated by X_{-k}^{-1}. Thus we have

Proposition 2.4 *Let* (X_n) *be a stationary chain as in definition 2.3, having context length function* l. *Put* $\mathbb{F}_k = \sigma\{X_{-k}, \ldots, X_{-1}\}, k \geq 1$. *Then* $l(X_{-\infty}^{-1})$ *is a* $(\mathbb{F}_k)_k-$*stopping time.*

Given a context length function l, we define an associated countable subset $\tau \subset A_+^*$ by

$$\tau = \tau^l = \{x_{-k}^{-1} : k = l(x_{-k}^{-1}), \, k \geq 1\}.$$

To simplify notation, we will denote by $\underline{x}$ and $\underline{y}$ generic elements of τ.

Definition 2.5 *Given a finite sequence* x_{-k}^{-1}, *we shall call suffix of* x_{-k}^{-1} *each string* x_{-j}^{-1} *with* $j \leq k$. *If* $j < k$ *we call* x_{-j}^{-1} *proper suffix of* x_{-k}^{-1}. *Now let* $\mathcal{S} \subset A_+^*$. *We say that* $\mathcal{S}$ *satisfies the suffix property if for no* $x_{-k}^{-1} \in \mathcal{S}$, *there exists a proper suffix* $x_{-j}^{-1} \in \mathcal{S}$ *of* x_{-k}^{-1}.

The following proposition follows immediately from property (ii) of definition 2.1.

Proposition 2.6 *Given a context length function l, the associated set τ^l satisfies the suffix property.*

As a consequence, τ can be identified with the set of leaves of a rooted tree with a countable set of finite labeled branches.

Definition 2.7 *We call probabilistic context tree on A the ordered pair (τ, p), where*

$$p = \{p(.|\underline{x}),\ \underline{x} \in \tau\}$$

is the family of transition probabilities of (2.2). We say that the probabilistic context tree (τ, p) is unbounded if the function l is unbounded.

Definition 2.8 *Let $(X_n)_{n\in\mathbb{Z}}$ be a stationary chain and let (τ, p) be a probabilistic context tree. We shall say that $(X_n)_{n\in\mathbb{Z}}$ is compatible with (τ, p), if (2.2) holds for all $\underline{x} \in \tau$.*

In order to illustrate these mathematical concepts, let us consider the following example.

Example 2.9 *Consider a two-symbol alphabet $A = \{0, 1\}$ and the following context length function*

$$l(x_{-\infty}^{-1}) = \inf\{k : x_{-k} = 1\}.$$

Then the associated tree τ is given by

$$\tau = \{10^k, k \geq 0\},$$

where 10^k represents the sequence $(x_{-k-1}, x_{-k}, \ldots, x_{-1})$ such that $x_{-i} = 0$ for all $1 \leq i \leq k$ and $x_{-k-1} = 1$.

The associated transition probabilities are defined by

$$P(X_0 = 1|X_{-1} = \ldots = X_{-k} = 0, X_{-k-1} = 1) = q_k, k \geq 0,$$

with $0 \leq q_k \leq 1$.

Clearly, the stochastic chain associated to this context length function l is a chain of infinite order. This raises the mathematical question of existence of such a process. It is straightforward to see that the following proposition holds true.

Proposition 2.10 *Suppose that $X_0 = 1$. Put $T_1 = \inf\{k \geq 1 : X_k = 1\}$. A necessary and sufficient condition for $T_1 < +\infty$ almost surely is*

$$\sum_{k\geq 0} q_k = +\infty. \tag{2.3}$$

This means that if (2.3) is satisfied, then – provided the chain starts from 1 – almost surely there will be appearance of an infinite number of the symbol 1. This implies that there exists a non-trivial stationary chain associated to this probabilistic context tree.

Observe that the process X_n is actually a renewal process with the renewal times defined as follows.

$$T_0 = \sup\{n < 0 : X_n = 1\},$$

and for $k \geq 1$,

$$T_k = \inf\{n > T_{k-1} : X_n = 1\} \quad \text{and} \quad T_{-k} := \sup\{n < T_{-(k-1)} : X_n = 1\}.$$

This example shows clearly that the tree of contexts defines a partition of all possible pasts with the exception of the single string composed of all symbols identical to 0. The condition (2.3) shows that it is possible to construct the chain taking values in the set of all sequences having an infinite number of symbols 1 both to the left and to the right of the origin.

However, we could also include this exceptional string to the set of possible contexts by defining an extra parameter q_∞. This is the choice of Csiszár and Talata (2006). If q_∞ is strictly positive, then condition (2.3) implies that after a finite time, there will appearance of the symbol 1, even if we start with an infinity of symbols 0. In other terms, this exceptional string does not have to be considered if we are interested in the stationary regime of the chain.

In case that $q_\infty = 0$ and (2.3) holds, we have the phenomenon of phase transition. One of the phases is composed of only one string having only the symbol 0.

The renewal process is an interesting example of a chain having memory of unbounded variable length. In the case where the probabilistic context tree is bounded, the corresponding chain is in fact a Markov chain whose order is equal to the maximal context length. However, the tree of contexts provides interesting additional information concerning the dependencies in the data and the structure of the chain. This raises the issue how to estimate the context tree out of the data. This was originally solved in Rissanen's 1983 paper using the algorithm Context.

At this point it is important to discuss the following minimality issue. Among all possible context trees fitting the data, we want of course to identify the smallest one. This is the tree corresponding to the smallest context length function. More precisely, if l and l' are context length functions, we shall say that $l \leq l'$ if $l(x_{-\infty}^{-1}) \leq l'(x_{-\infty}^{-1})$ for any string $x_{-\infty}^{-1} \in \underline{\mathrm{A}}$. From now on we shall call *context* of a string $x_{-\infty}^{-1}$ the context associated to the minimal context length function. Estimating this minimal context is precisely the goal of the algorithm Context.

3 The algorithm Context

We now present the algorithm Context introduced by Rissanen (1983). The goal of the algorithm is to estimate adaptively the context of the next symbol X_n given the past symbols X_0^{n-1}. The way the algorithm Context works can be summarized as follows. Given a sample produced by a chain with variable memory, we start with a maximal tree of candidate contexts for the sample. The branches of this first tree are then pruned starting from the leaves towards the root until we obtain a minimal tree of contexts well adapted to the sample. We associate to each context an estimated probability transition defined as the proportion of time the context appears in the sample followed by each one of the symbols in the alphabet. We stop pruning once the gain function exceeds a given threshold.

Let $X_0, X_1, \ldots, X_{n-1}$ be a sample from the finite probabilistic tree (τ, p). For any finite string x_{-j}^{-1} with $j \leq n$, we denote $N_n(x_{-j}^{-1})$ the number of occurrences of the string in the sample

$$N_n(x_{-j}^{-1}) = \sum_{t=0}^{n-j} \mathbf{1}\left\{X_t^{t+j-1} = x_{-j}^{-1}\right\}. \tag{3.4}$$

Rissanen first constructs a maximal candidate context $X_{n-M(n)}^{n-1}$ where $M(n)$ is a random length defined as follows

$$M(n) \;=\; \min\left\{i = 0, 1, \ldots, \lfloor C_1 \log n \rfloor : N_n(X_{n-i}^{n-1}) \;>\; \frac{C_2\, n}{\sqrt{\log n}}\right\}. \tag{3.5}$$

Here C_1 and C_2 are arbitrary positive constants. In the case the set is empty we take $M(n) = 0$.

Rissanen then shortens this maximal candidate context by successively pruning the branches according to a sequence of tests based on the likelihood ratio statistics. This is formally done as follows.

If $\sum_{b\in A} N_n(x_{-k}^{-1}b) > 0$, define the estimator of the transition probability p by

$$\hat{p}_n(a|x_{-k}^{-1}) = \frac{N_n(x_{-k}^{-1}a)}{\sum_{b\in A} N_n(x_{-k}^{-1}b)} \tag{3.6}$$

where $x_{-j}^{-1}a$ denotes the string $(x_{-j}, \ldots, x_{-1}, a)$, obtained by concatenating x_{-j}^{-1} and the symbol a. If $\sum_{b\in A} N_n(x_{-k}^{-1}b) = 0$, define $\hat{p}_n(a|x_{-k}^{-1}) = 1/|A|$.

For $i \geq 1$ we define

$$\Lambda_n(x_{-i}^{-1}) = 2 \sum_{y\in A} \sum_{a\in A} N_n(yx_{-i}^{-1}a) \log \left[\frac{\hat{p}_n(a|x_{-i}^{-1}y)}{\hat{p}_n(a|x_{-i}^{-1})} \right], \tag{3.7}$$

where yx_{-i}^{-1} denotes the string $(y, x_{-i}, \ldots, x_{-1})$, and where

$$\hat{p}_n(a|x_{-i}^{-1}y) = \frac{N_n(yx_{-i}^{-1}a)}{\sum_{b\in A} N_n(yx_{-i}^{-1}b)}.$$

Notice that $\Lambda_n(x_{-i}^{-1})$ is the log-likelihood ratio statistic for testing the consistency of the sample with a probabilistic suffix tree (τ, p) against the alternative that it is consistent with (τ', p') where τ and τ' differ only by one set of sibling nodes branching from x_{-i}^{-1}. $\Lambda_n(x_{-i}^{-1})$ plays the role of a gain function telling us whether it is worth or not taking a next step further back in the past.

Rissanen then defines the length of the estimated current context $\hat{\ell}_n$ as

$$\hat{\ell}_n(X_0^{n-1}) = 1 + \max\left\{ i = 1, \ldots, M(n) - 1 : \Lambda_n(X_{n-i}^{n-1}) > C_2 \log n \right\}, \tag{3.8}$$

where C_2 is any positive constant.

Then, the result in Rissanen (1983) is the following.

Theorem 3.1 *Given a realization $X_0, \ldots, X_{n-1}$ of a probabilistic suffix tree (τ, p) with finite height, then*

$$P\left(\hat{\ell}_n(X_0^{n-1}) \neq \ell(X_0^{n-1})\right) \longrightarrow 0 \tag{3.9}$$

as $n \to \infty$.

Rissanen proves this result in a very short and elegant way. His starting point is the following upper bound.

$$\begin{aligned} &P\left(\hat{\ell}_n(X_0^{n-1}) \neq \ell(X_0^{n-1})\right) \leq \\ &P\left(\hat{\ell}_n(X_0^{n-1}) \neq \ell(X_0^{n-1}) | N_n\left(X_{n-\ell(X_0^{n-1})}^{n-1}\right) > \frac{C_2 n}{\sqrt{\log n}}\right) P\left(N_n\left(X_{n-\ell(X_0^{n-1})}^{n-1}\right) > \frac{C_2 n}{\sqrt{\log n}}\right) \\ &\qquad + P\left(\bigcup_{w\in\tau} \left\{ N_n(w) \leq \frac{C_2 n}{\sqrt{\log n}} \right\}\right). \end{aligned} \tag{3.10}$$

Then he provides the following explicit upper bound for the conditional probability in the right-hand side of (3.10)

$$P\left(\hat{\ell}_n(X_0^{n-1}) \neq \ell(X_0^{n-1}) | N_n\left(X_{n-\ell(X_0^{n-1})}^{n-1}\right) > \frac{C_2 n}{\sqrt{\log n}}\right) \leq C_1 \log n \, e^{-C_2'\sqrt{\log n}}, \tag{3.11}$$

where C_1, C_2 and C_2' are positive constants independent of the maximum of the context length function.

With respect to the second term he only observes that, by ergodicity, for each $x_{-k}^{-1} \in \tau$ we have

$$P\left(N_n\left(x_{-k}^{-1}\right) \leq \frac{C_2 n}{\sqrt{\log n}}\right) \longrightarrow 0 \tag{3.12}$$

as $n \to \infty$. Since τ is finite the convergence in (3.12) implies the desired result.

4 The unbounded case

In his original paper, Rissanen was only interested in the case of bounded context trees. However, from the mathematical point of view, it is interesting to consider also the case of unbounded probabilistic context trees corresponding to chains of infinite order. It can be argued that also from an applied point of view the unbounded case must be considered as noisy observation of Markov chains generically have infinite order memory.

The unbounded case raises immediately the preliminary question of existence and uniqueness of the corresponding chain. This issue can be addressed by adapting to probabilistic context trees the conditions for existence and uniqueness that have already been proved for infinite order chains. This is precisely what is done in the paper by Duarte et al. (2006) who adapt the type A condition presented in Fernández and Galves (2002) in the following way.

To simplify the presentation, let us introduce some extra notation. Recall that $\underline{x}$ and $\underline{y}$ denote generic elements of τ. Given $\underline{x} = x_{-i}^{-1}$ and $\underline{y} = y_{-j}^{-1}$, we shall write $\underline{x} \stackrel{k}{=} \underline{y}$ if and only if $k \leq \min\{i, j\}$ and $x_{-1} = y_{-1}, \ldots, x_{-k} = y_{-k}$.

Definition 4.1 *A probabilistic suffix tree (τ, p) on A is of type A if its transition probabilities p satisfy the following conditions.*

1. **Weakly non-nullness**, *that is*

$$\sum_{a \in A} \inf_{\underline{x} \in \tau} p(a|\underline{x}) > 0\,; \tag{4.13}$$

2. **Continuity**, *that is*

$$\beta(k) \;=\; \max_{a \in A} \sup\{|p(a|\underline{x}) - p(a|\underline{y})|, \underline{y} \in \tau, \underline{x} \in \tau \;\; \textit{with} \;\; \underline{x} \stackrel{k}{=} \underline{y}\} \;\to\; 0 \tag{4.14}$$

as $k \to \infty$. We also define

$$\beta(0) = \max_{a \in A} \sup\{|p(a|\underline{x}) - p(a|\underline{y})|, \underline{y} \in \tau, \underline{x} \in \tau \;\; \textit{with} \;\; x_{-1} \neq y_{-1}\}.$$

The sequence $\{\beta(k)\}_k \in I\!N$ is called the **continuity rate**.

For a probabilistic suffix tree of type A with summable continuity rate, the maximal coupling argument used in Fernández and Galves (2002) implies the uniqueness of the law of the chain consistent with it.

We now present a slightly different version of the algorithm Context using the same gain function Λ_n but in which the length of the maximum context candidate is now deterministic and nor more random. More precisely, we define the length of the biggest candidate context now as

$$k(n) = C_1 \log n \tag{4.15}$$

with a suitable positive constant C_1.

The intuitive reason behind the choice of the upper bound length $C_1 \log n$ is the impossibility of estimating the probability of sequences of length much longer than $\log n$ based on a sample of length n. Recent versions of this fact can be found in Marton and Shields (1994, 1996) and Csiszár (2002).

Now, the definition of $\hat{\ell}_n$ is similar to the one in the original algorithm of Rissanen, that is

$$\hat{\ell}_n(X_0^{n-1}) = 1 + \max\left\{i = 1, \ldots, k(n) - 1 : \Lambda_n(X_{n-i}^{n-1}) \; > \; C_2 \log n\right\} , \tag{4.16}$$

where C_2 is any positive constant.

The reason for taking the length of the maximum context candidate deterministic and no more random is to be able to use the classical results on the convergence of the law of $\Lambda_n(x_{-i}^{-1})$ to a chi-square distribution. However, we are not in a Markov setup since the probabilistic context tree is unbounded, and the chi-square approximation only works for Markov chains of fixed finite order.

To overcome this difficulty, we use the canonical Markov approximation of chains of infinite order presented in Fernández and Galves (2002) that we recall now by adapting the definitions and theorem to the framework of probabilistic context trees. The goal is to approximate a chain compatible with an unbounded probabilistic context tree by a sequence of chains compatible with bounded probabilistic context trees.

Definition 4.2 *For all $k \geq 1$, the canonical Markov approximation of order k of a chain $(X_n)_{n\in\mathbb{Z}}$ is the chain with memory of variable length bounded by k compatible with the probabilistic context tree $(\tau^{[k]}, p^{[k]})$ where*

$$\tau^{[k]} = \{\underline{x} \in \tau; l(\underline{x}) \leq k\} \cup \{x_{-k}^{-1}; \underline{x} \in \tau, l(\underline{x}) \geq k\} \tag{4.17}$$

for all $a \in A$, $\underline{x} \in \tau$, and where

$$p^{[k]}(a|x_{-j}^{-1}) \; := \; P(X_0 = a | X_{-j}^{-1} = x_{-j}^{-1}) \tag{4.18}$$

for all $x_{-j}^{-1} \in \tau^{[k]}$.

Observe that for contexts $\underline{x} \in \tau$ which length does not exceed k, we have $p^{[k]}(a|\underline{x}) = p(a|\underline{x})$. However, for sequences x_{-k}^{-1} which are internal nodes of τ, there is no easy explicit formula expressing $p^{[k]}(\cdot|x_{-k}^{-1})$ in terms of the family $\{p(\cdot|\underline{y}), \underline{y} \in \tau\}$.

The main result of Fernández and Galves (2002) that will be used in the proof of the consistency of the algorithm Context can be stated as follows.

Theorem 4.3 *Let $(X_n)_{n\in\mathbb{Z}}$ be a chain compatible with a type A probabilistic context tree (τ, p) with summable continuity rate, and let $(X_n^{[k]})_{n\in\mathbb{Z}}$ be its canonical Markov approximation of order k. Then there exists a coupling between $(X_n)_{n\in\mathbb{Z}}$ and $(X_n^{[k]})_{n\in\mathbb{Z}}$ and a constant $C > 0$ such that*

$$P\left(X_0 \neq X_0^{[k]}\right) \leq C\beta(k)\,. \tag{4.19}$$

Using this result and the classical chi-square approximation for Markov chains, Duarte et al. (2006) proved the consistency of their version of the algorithm Context in the unbounded case and also provided an upper bound for the rate of convergence. Their result is the following.

Theorem 4.4 *Let $X_0, X_2, \ldots, X_{n-1}$ be a sample from a type A unbounded probabilistic suffix tree (τ, p) with continuity rate $\beta(j) \leq f(j)\exp\{-j\}$, with $f(j) \to 0$ as $j \to \infty$. Then, for any choice of positive constants C_1 and C_2 in (4.15) and (4.16), there exist positive constants C and D such that*

$$P\left(\hat{\ell}_n(X_0^{n-1}) \neq \ell(X_0^{n-1})\right) \leq C_1 \log n(n^{-C_2} + D/n) + Cf(C_1 \log n)\,.$$

The proof can be sketched very easily. Take $k = k(n) = C_1 \log(n)$ and construct a coupled version of the processes $(X_t)_{t\in\mathbb{Z}}$ and $(X_t^{[k(n)]})_{t\in\mathbb{Z}}$. First of all notice that for $k = k(n)$,

$$\begin{aligned} &P\left(\hat{\ell}_n(X_0, \ldots, X_{n-1}) \neq \ell(X_0, \ldots, X_{n-1})\right) \leq \\ &P\left(\hat{\ell}_n(X_0^{[k]}, \ldots X_{n-1}^{[k]}) \neq \ell(X_0^{[k]}, \ldots X_{n-1}^{[k]})\right) + P\left(\bigcup_{i=1}^{n}\{X_i \neq X_i^{[k]}\}\right). \end{aligned} \tag{4.20}$$

Using the inequality (4.19) of Fernández and Galves (2002), the second term in (4.20) can be bounded above as

$$P\left(\bigcup_{i=1}^{n}\{X_i \neq X_i^{[k]}\}\right) \leq n\,C\,\beta(k(n)).$$

The first term in (4.20) can be treated using the classical chi-square approximation for the log-likelihood ratio test for Markov chains of fixed order k.

More precisely, we know that for fixed x_{-i}^{-1}, under the null hypothesis, the statistics $\Lambda_n(x_{-i}^{-1})$, given by (3.7), has asymptotically chi-square distribution with $|A| - 1$ degrees of freedom (see, for example, van der Vaart (1998)). We recall that, for each x_{-i}^{-1} the null hypothesis (H_0^i) is that the true context is x_{-i}^{-1}.

Since we are going to perform a sequence of $k(n)$ sequential tests where $k(n) \to \infty$ as n diverges, we need to control the error in the chi-square approximation. For this, we use a well-known asymptotic expansion for the distribution of $\Lambda_n(x_{-i}^{-1})$ due to Hayakawa (1977) which implies that

$$P\left(\Lambda_n(x_{-i}^{-1}) \leq t | H_0^i\right) = P\left(\chi^2 \leq t\right) + D/n\,, \tag{4.21}$$

where D is a positive constant and χ^2 is random variable with distribution chi-square with $|A| - 1$ degrees of freedom.

Therefore, it is immediate that

$$P\left(\Lambda_n(x_{-i}^{-1}) \;>\; C_2 \log n\right) \leq e^{-C_2 \log n} + D/n\,.$$

By the way we defined $\hat{\ell}_n$ in (4.16), in order to find $\hat{\ell}_n(X_0^{n-1})$ we have to perform at most $k(n)$ tests. We want to give an upper bound for the overall probability of type I error in a sequence of $k(n)$ sequential tests. An upper bound is given by the Bonferroni inequality, which in our case can be written as

$$P\left(\cup_{i=2}^{k(n)}\{\Lambda_n(x_{-i}^{-1}) > C_2 \log n\}|H_0^i\right) \leq \sum_{i=2}^{k(n)} P(\Lambda_n(x_{-i}^{-1}) > C_2 \log n|H_0^i).$$

This last term is bounded above by $C_1 \log n(n^{-C_2} + D/n)$. This concludes the proof.

Theorem 4.4 not only proves the consistency of the algorithm Context, but it also gives an upper bound for the rate of convergence. The estimation of the rate of convergence is crucial because it gives a bound on the minimum size of a sample required to guarantee, with a given probability, that the estimated tree is the good one. This is the issue we address to in the next section.

5 Rate of convergence of the algorithm Context

Note that Rissanen's original theorem 3.1 as well as theorem 4.4 only show that all the contexts identified are true contexts with high probability. In other words, the estimated tree is a subtree of the true tree with high probability. In the case of bounded probabilistic context trees this missing point was handled with in Weinberger et al. (1995). This paper not only proves that the set of all contexts is reached, but also gives a bound for the rate of convergence.

More precisely, let us define the empirical tree

$$\hat{\tau}_n = \left\{ X_{j-\hat{\ell}_j(X_0^{j-1})}^{j-1} : j = n/2, \ldots, n \right\}. \tag{5.22}$$

Actually, this is a slightly simplified version of the empirical tree defined in Weinberger et al. (1995). In particular, we are neglecting all the computational aspects considered there. But from the mathematical point of view, this definition perfectly does the job. Their convergence result is the following.

Theorem 5.1 *Let (τ, p) be a bounded probabilistic context tree and let $X_0, \ldots, X_n$ be compatible with (τ, p). Then we have*

$$\sum_{n \geq 1} P(\hat{\tau}_n \neq \tau) \log n < +\infty.$$

In the unbounded case, this issue was treated without estimation of the rate of convergence in Ferrari and Wyner (2003) and including estimation of the rate of convergence in Galves and Leonardi (2008).

This last paper considers another slightly modified version of the algorithm Context using a different gain function, which has been introduced in Galves et al. (2007). More precisely, let us define for any finite string $x_{-k}^{-1} \in A_+^*$ the gain function

$$\Delta_n(x_{-k}^{-1}) = \max_{a \in A} |\hat{p}_n(a|x_{-k}^{-1}) - \hat{p}_n(a|x_{-(k-1)}^{-1})|.$$

This gain function is well adapted to use exponential inequalities for the empirical transition probabilities in the pruning procedure rather than the chi-square approximation of the log-likelihood ratio as in theorems 3.1 and 4.4.

The theorem is stated in the following framework. Consider a a stationary chain $(X_n)_{n\in\mathbb{Z}}$ compatible with an unbounded probabilistic context tree (τ, p). For this chain, we define the sequence $(\alpha_n)_{n\in\mathbb{N}}$ by

$$\begin{aligned}\alpha_0 &= \sum_{a\in A} \inf_{\underline{x}\in\tau} p(a|\underline{x}),\\ \alpha_n &= \inf_{x_{-n}^{-1}} \sum_{a\in A} \inf_{\underline{y}\in\tau: l(\underline{y})\geq n, \underline{y}^{n}=x_{-n}^{-1}} p(a|\underline{y}).\end{aligned}$$

We assume that the probabilistic context tree (τ, p) satisfies the condition (4.13) of weakly non-nullness, that is $\alpha_0 > 0$. We assume also the following summability condition

$$\alpha = \sum_{n\geq 0}(1-\alpha_n) < +\infty. \tag{5.23}$$

Given a sequence $x_1^j = (x_1, \ldots, x_j) \in A^j$ we denote by

$$p(x_1^j) \;=\; I\!P(X_1^j = x_1^j).$$

Then for an integer $m \geq 1$, we define

$$D_m = \min_{x_{-k}^{-1}\in\tau: k\leq m} \max_{a\in A}\{\, |p(a|x_{-k}^{-1}) - p(a|x_{-(k-1)}^{-1})|\,\}, \tag{5.24}$$

and

$$\epsilon_m = \min\{\, p(x_{-k}^{-1}) \colon k \leq m \text{ and } p(x_{-k}^{-1}) > 0\,\}. \tag{5.25}$$

Intuitively, D_m tells us how distinctive is the difference between transition probabilities associated to the exact contexts and those associated to a string shorter one step in the past. We do not want to impose restrictions on the transition probabilities elsewhere then at the end of the branches of the context tree. This has to do with the pruning procedure which goes from the leaves to the root of the tree.

In the unbounded case, a natural way to state the convergence results is to consider truncated trees. The definition is the following. Given an integer K we will denote by $\tau|_K$ the tree τ *truncated* to level K, that is

$$\tau|_K = \{x_{-k}^{-1} \in \tau \colon k \leq K\} \cup \{x_{-K}^{-1} \text{ such that } x_k^{-1} \in \tau \text{ for some } k \geq K\}.$$

Actually, this is exactly the same tree which was called $\tau^{[K]}$ in (4.17). The notation $\tau|_K$ is more suitable for what follows.

The associated empirical tree of height k is defined in the following way.

Definition 5.2 *Given $\delta > 0$ and $k < n$, the empirical tree is defined as*

$$\hat{\tau}_n^{\delta,k} = \{x_{-r}^{-1}, 1 \leq r \leq k : \Delta_n(x_{-r}^{-1}) > \delta \,\wedge\, \Delta_n(y_{-(r+j)}^{-(r+1)} x_{-r}^{-1}) \leq \delta,\ \forall\, y_{-j}^{-(r+1)}, 1 \leq j \leq k-r\}.$$

In case $r = k$, the string $y_{-(r+j)}^{-(r+1)}$ is empty.

Note that in this definition, the parameter δ expresses the coarseness of the pruning criterion and k is the maximal length of the estimated contexts.

Now, Galves and Leonardi (2008) obtain the following result on the rate of convergence for the truncated context tree.

Theorem 5.3 *Let (τ, p) be a probabilistic context tree satisfying (4.13) and (5.23). Let $X_0, \ldots, X_n$ be a stationary stochastic chain compatible with (τ, p). Then for any integer K, any k satisfying*

$$k > \max_{\underline{x} \in \tau|_K} \min \{\ell(\underline{y}) \colon \underline{y} \in \tau,\ \underline{x} \stackrel{K}{=} \underline{y}\}, \tag{5.26}$$

for any $\delta < D_k$ and for each

$$n > \frac{2(|A|+1)}{\min(\delta, D_k - \delta)\epsilon_k} + k \tag{5.27}$$

we have that

$$P(\hat{\tau}_n^{\delta,k}|_K \neq \tau|_K) \ \leq 4\, e^{\frac{1}{e}}\, |A|^{k+2} \exp[-(n-k)\, \frac{[\min(\frac{\delta}{2}, \frac{D_k - \delta}{2}) - \frac{|A|+1}{(n-k)\epsilon_k}]^2 \epsilon_k^2 C}{4|A|^2(k+1)}],$$

where

$$C = \frac{\alpha_0}{8e(\alpha + \alpha_0)}.$$

In this theorem, the empirical trees have to be of height $k \geq K$ for the following reason. Truncating τ at level K implies that contexts longer than K are cut before reaching their end, and associated transition probabilities might not differ when comparing them at length K and $K-1$. That's why we consider the bigger empirical tree of height k satisfying condition (5.26). This guaranties that for each element $\underline{x}$ of the truncated empirical tree there is at least one real context $\underline{y}$ which has $\underline{x}$ as its suffix.

As a consequence of theorem 5.3, Galves and Leonardi (2008) obtain the following strong consistency result.

Corollary 5.4 *Let (τ, p) be a probabilistic context tree satisfying the conditions of theorem 5.3. Then*

$$\hat{\tau}_n^{\delta,k}|_K = \tau|_K, \tag{5.28}$$

eventually almost surely, as $n \to \infty$.

The main ingredient of the proof of theorem 5.3 is an exponential upper bound for the deviations of the empirical transition probabilities. More precisely, Galves and Leonardi (2008) prove the following result.

Theorem 5.5 *For any finite sequence x_{-k}^{-1} with $p(x_{-k}^{-1}) > 0$, any symbol $a \in A$, any $t > 0$ and any $n > \frac{|A|+1}{tp(x_{-k}^{-1})} + k$ the following inequality holds.*

$$P(|\hat{p}_n(a|x_{-k}^{-1}) - p(a|x_{-k}^{-1})| > t) \leq$$
$$2|A|\, e^{\frac{1}{e}} \exp[-(n-k)\, \frac{[t - \frac{|A|+1}{(n-k)p(x_{-k}^{-1})}]^2 p(x_{-k}^{-1})^2 C}{4|A|^2(k+1)}], \tag{5.29}$$

where

$$C = \frac{\alpha_0}{8e(\alpha + \alpha_0)}. \tag{5.30}$$

The proof of this theorem is inspired by recent exponential upper bounds obtained by Dedecker and Doukhan (2003), Dedecker and Prieur (2005) and Maume-Deschamps (2006). It is based on the following loss-of-memory inequality of Comets et al. (2002).

Theorem 5.6 *Let $(X_n)_{n\in\mathbb{Z}}$ be a stationary stochastic chain compatible with the probabilistic context tree (τ, p) of theorem 5.3. Then, there exist a sequence $\{\rho_l\}_{l\in\mathbb{N}}$ such that for any $i \geq 1$, any $k > i$, any $j \geq 1$ and any finite sequence x_1^j, the following inequality holds*

$$\sup_{x_1^i \in A^i} |P(X_k^{k+j-1} = x_1^j | X_1^i = y_1^i) - p(x_1^j)| \;\leq\; j\,\rho_{k-i-1}\,. \tag{5.31}$$

Moreover, the sequence $\{\rho_l\}_{l\in\mathbb{N}}$ is summable and

$$\sum_{l\in\mathbb{N}} \rho_l \;\leq\; 1 + \frac{2\alpha}{\alpha_0}.$$

Theorem 5.3 generalizes to the unbounded case previous results in Galves et al. (2008) for the case of bounded context trees. Note that the definition of the context tree estimator depends on the parameter δ, the same appearing in the constants of the exponential bound. To assure the consistency of the estimator we have to choose a δ sufficiently small, depending on the true probabilities of the process. The same thing happens to the parameter k. Therefore, this estimator is not universal, meaning that for fixed δ and k it fails to be consistent for all variable memory processes for which conditions (5.26) and (5.27) are not satisfied. We could try to overcome this difficulty by letting $\delta = \delta(n) \to 0$ and $k = k(n) \to +\infty$ as n increases. But doing this, we loose the exponential character of the upper bound. This could be considered as an illustration of the result in Finesso et al. (1996) who proved that in the simpler case of estimating the order of a Markov chain, it is not possible to have a universal estimator with exponential bounds for the probability of overestimation.

6 Some final comments and bibliographic remarks

Chains with memory of variable length were introduced in the information theory literature by Rissanen (1983) as a universal system for data compression. Originally called by Rissanen *tree machine*, *tree source*, *context models*, etc., this class of models recently became popular in the statistics literature under the name of *Variable Length Markov Chains (VLMC)*, coined by Bühlmann and Wyner (1999).

Rissanen (1983) not only introduced the notion of variable memory models but he also introduced the algorithm Context to estimate the probabilistic context tree. From Rissanen (1983) to Galves et al. (2008), passing by Ron et al. (1996) and Bühlmann and Wyner (1999), several variants of the algorithm Context have been presented in the literature. In all the variants the decision to prune a branch is taken by considering a *gain* function.

Rissanen (1983), Bühlmann and Wyner (1999) and Duarte et al. (2006) all defined the gain function in terms of the log likelihood ratio function. Rissanen (1983) proved the weak consistency of the algorithm Context in the case where the contexts have a bounded length. Bühlmann and Wyner (1999) proved the weak consistency of the algorithm also in the finite case without assuming a prior known bound on the maximal length of the memory but using a bound allowed to grow with the size of the sample.

A different gain function was introduced in Galves et al. (2008), considering differences between successive empirical transition probabilities and comparing them with a given threshold δ. An interesting consequence of the use of this different gain function was obtained by Collet et al. (2007). They proved that in the case of a binary alphabet and when taking δ within a suitable interval, it is possible to recover the context tree in the bounded case out from a noisy sample where each symbol can be flipped with small probability independently of the others.

The case of unbounded probabilistic context trees as far as we know was first considered by Ferrari and Wyner (2003) who also proved a weak consistency result for the algorithm Context in this more general setting. The unbounded case was also considered by Csiszár and Talata (2006) who introduced a different approach for the estimation of the probabilistic context tree using the Bayesian Information Criterion (BIC) as well as the Minimum Description Length Principle (MDL). We refer the reader to this last paper for a nice description of other approaches and results in this field, including the context tree maximizing algorithm by Willems et al. (1995). We also refer the reader to Garivier (2006a, b) for recent and elegant results on the BIC and the Context Tree Weighting Method (CTW). Garivier (2006c) is a very good presentation of models having memory of variable length, BIC, MDL, CTW and related issues in the framework of information theory.

With exception of Weinberger et al. (1995), the issue of the rate of convergence of the algorithm estimating the probabilistic context tree was not addressed in the literature until recently. Weinberger et al. (1995) proved in the bounded case that the probability that the estimated tree differs from the finite context tree is summable as a function of the sample size. Assuming weaker hypotheses than Ferrari and Wyner (2003), Duarte et al. (2006) proved in the unbounded case that the probability of error decreases as the inverse of the sample size.

Leonardi (2007) obtained an upper bound for the rate of convergence of penalized likelihood context tree estimators. It showed that the estimated context tree truncated at any fixed height approximates the real truncated tree at a rate that decreases faster than the inverse of an exponential function of the penalizing term. The proof mixes the approaches of Galves et al. (2008) and Csiszár and Talata (2006).

Several interesting papers have recently addressed the question of classification of proteins and DNA sequences using models with memory of variable length, which in bio-informatics are often called prediction suffix trees (PST). Many of these papers have been written from a bio-informatics point of view focusing on the development of new tools rather than being concerned with mathematically rigorous proofs. The interested reader can find a starting point to this literature for instance in the papers by Bejerano et al. (2001), Bejerano and Yona (2001), Eskin et al. (2000), Leonardi (2006) and Miele et al. (2005). The same type of analysis has been used successfully to classification tasks in other domains like musicology (Lartillot et al. 2003), linguistics (Selding et al. 2001), etc.

This presentation did not intend to be exhaustive and the bibliography in many cases only gives a few hints about possible starting points to the literature. However, we think we have presented the state of the art concerning the rate of convergence of context tree estimators.

In the introduction we said that *Rissanen's ingenious idea was to construct a stochastic model that generalizes the notion of relevant domain (in biology or linguistics) to any kind of symbolic strings.* Actually, God only knows what Jorma had in mind when he invented this class of models. The French poet Paul Eluard wrote a book called *Les frères voyants*. This was the name given in the middle-age to people guiding blind persons. So maybe Rissanen acted as a *frère voyant* using his intuition to push mathematics and statistics into a challenging new direction.

References

[1] G. Bejerano, Y. Seldin, H. Margalit and N. Tishby, "Markovian domain fingerprinting: statistical segmentation of protein sequences", *Bioinformatics*, vol. 17, pp. 927–934, 2001.

[2] G. Bejerano and G. Yona, "Variations on probabilistic suffix trees: statistical modeling and prediction of protein families", *Bioinformatics*, vol. 17, Number 1, pp. 23–43, 2001.

[3] P. Bühlmann and A.J. Wyner, "Variable length Markov chains", *Ann. Statist.*, vol. 27, pp. 480–513, 1999.

[4] P. Collet, A. Galves and F. Leonardi, "Random perturbations of stochastic chains with unbounded variable length memory", manuscript, can be downloaded from ArXiv: math/0707.2796., 2007.

[5] F. Comets, R. Fernández and P. Ferrari, "Processes with long memory: Regenerative construction and perfect simulation", *Ann. of Appl. Probab.*, vol. 12, Number 3, pp. 921–943, 2002.

[6] I. Csiszár, "Large-scale typicality of Markov sample paths and consistency of MDL order estimators. Special issue on Shannon theory: perspective, trends, and applications", *IEEE Trans. Inform. Theory*, vol. 48, Number 6, pp. 1616–1628, 2002.

[7] I. Csiszár and Z. Talata, "Context tree estimation for not necessarily finite memory processes, via BIC and MDL", *IEEE Trans. Inform. Theory*, vol. 52, Number 3, pp. 1007–1016, 2006.

[8] J. Dedecker and P. Doukhan, "A new covariance inequality and applications", *Stochastic Process. Appl.*, vol. 106, Number 1, pp. 63–80, 2003.

[9] J. Dedecker and C. Prieur, "New dependence coefficients. Examples and applications to statistics", *Probab. Theory Related Fields*, vol. 132, pp. 203–236, 2005.

[10] D. Duarte, A. Galves and N.L. Garcia, "Markov approximation and consistent estimation of unbounded probabilistic suffix trees", *Bull. Braz. Math. Soc.*, vol. 37, Number 4, pp. 581–592, 2006.

[11] E. Eskin, W.N. Grundy and Y. Singer, "Protein family classification using sparse Markov transducers", manuscript, can be downloaded from http://citeseer.ist.psu.edu/328658.html, 2000.

[12] R. Fernández and A. Galves, "Markov approximations of chains of infinite order", Fifth Brazilian School in Probability (Ubatuba, 2001), *Bull. Braz. Math. Soc. (N.S.)*, vol. 33, Number 3, pp. 295–306, 2002.

[13] F. Ferrari and A. Wyner, "Estimation of general stationary processes by variable length Markov chains", *Scand. J. Statist.*, vol. 30, Number 3, pp. 459–480, 2003.

[14] L. Finesso, C.-C. Liu P. and P. Narayan, "The optimal error exponent for Markov order estimation", *IEEE Trans. Inform. Theory*, vol. 42, Number 5, pp. 1488–1497, 1996.

[15] A. Galves and F. Leonardi, "Exponential inequalities for empirical unbounded context trees", manuscript, to appear in a special issue of *Progress in Probability*, V. Sidoravicius and M. E. Vares, Eds., Birkhäuser, can be downloaded from ArXiv: math/0710.5900, 2007.

[16] A. Galves, V. Maume-Deschamps and B. Schmitt, "Exponential inequalities for VLMC empirical trees", *ESAIM. Probability and Statistics*, to appear 2008.

[17] A. Garivier, "Consistency of the unlimited BIC Context Tree estimator", *IEEE Trans. Inform. Theory*, vol. 52, Number 10, pp. 4630–4635, 2006.

[18] A. Garivier, "Redundancy of the context-tree weighting method on renewal and Markov renewal processes", *IEEE Trans. Inform. Theory*, vol. 52, Number 12, pp. 5579–5586, 2006.

[19] A. Garivier, "Modèles contextuels et alphabets infinis en théorie de l'information", PhD thesis, Université de Paris Sud, can be downloaded from http://www.math.u-psud.fr/ garivier/, Nov. 2006.

[20] T. Hayakawa, "The likelihood ratio criterion and the asymptotic expansion of its distribution", *Ann. Inst. Statist. Math.*, vol. 29, Number 3, pp. 359–378, 1977.

[21] O. Lartillot, S. Dubnov, G. Assayag and G. Bejerano, "A system for computer music generation by learning and improvisation in a particular style", *IEEE Computer J.*, vol. 36, Number 10, pp. 73–80, 2003.

[22] F. Leonardi, "A generalization of the PST algorithm: modeling the sparse nature of protein sequences", *Bioinformatics*, vol. 22, Number 7, pp. 1302–1307, 2006.

[23] F. Leonardi, "Rate of convergence of penalized likelihood context tree estimators", manuscript, can be downloaded from ArXiv: math/0701810v2, 2007.

[24] K. Marton and P.C. Shields, "Entropy and the consistent estimation of joint distributions", *Ann. Probab.*, vol. 22, Number 2, pp. 960–977, 1994.

[25] K. Marton and P.C. Shields, "Correction: "Entropy and the consistent estimation of joint distributions" [Ann. Probab. **22** (1994), no. 2, 960–977; MR1288138 (95g:94004)]", *Ann. Probab.*, vol. 24, Number 1, pp. 541–545, 1996.

[26] V. Maume-Deschamps, *Exponential inequalities and estimation of conditional probabilities*, Lecture Notes in Statist., vol. 187, Heidelberg: Springer, 2006.

[27] V. Miele, P. Y. Bourguignon, D. Robelin, G. Nuel and H. Richard, "seq++: a package for biological sequences analysis with a range of Markov-related models", *BioInformatics*, vol. 21, Number 11, pp. 2783–2874, 2005.

[28] J. Rissanen, "A universal data compression system", *IEEE Trans. Inform. Theory*, vol. 29, Number 5, pp. 656–664, 1983.

[29] D. Ron, Y. Singer and N. Tishby, "The power of amnesia: learning probabilistic automata with variable memory length", *Machine Learning*, vol. 25, pp. 117–149, 1996.

[30] Y. Seldin, G. Bejerano and N. Tishby, "Unsupervised sequence segmentation by a mixture of variable memory Markov models", manuscript, can be downloaded from http://citeseer.ist.psu.edu/449505.html, 2001.

[31] A.W. van der Vaart, *Asymptotic statistics*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge: Cambridge University Press, 1998.

[32] M.J. Weinberger, J. Rissanen and M. Feder, "A universal finite memory source", *IEEE Trans. Inform. Theory*, vol. 41, Number 3, pp. 643–652, 1995.

[33] E. M. J. Willems, Y. M. Shtarkov and T. J. Tjalkens, "The Context-Tree weighting method: Basic properties", *IEEE Trans. Inform. Theory*, vol. 41, pp. 653–664, 1995.

Antonio Galves
Instituto de Matemática e Estatística
Universidade de São Paulo
BP 66281
05315-970 São Paulo, Brasil
e-mail: galves@ime.usp.br

Eva Löcherbach
Université Paris-Est
LAMA – UMR CNRS 8050
61, Avenue du Général de Gaulle
94000 Créteil, France
e-mail: locherbach@univ-paris12.fr

Some Information-Theoretic Computations Related to the Distribution of Prime Numbers

I. Kontoyiannis*

March 27, 2008

Abstract

We illustrate how elementary information-theoretic ideas may be employed to provide proofs for well-known, nontrivial results in number theory. Specifically, we give an elementary and fairly short proof of the following asymptotic result,

$$\sum_{p\leq n} \frac{\log p}{p} \sim \log n, \quad \text{as } n \to \infty,$$

where the sum is over all primes p not exceeding n. We also give finite-n bounds refining the above limit. This result, originally proved by Chebyshev in 1852, is closely related to the celebrated prime number theorem.

1 Introduction

The significant depth of the connection between information theory and statistics appears to have been recognized very soon after the birth of information theory [17] in 1948; a book-length exposition was provided by Kullback [12] already in 1959. In subsequent decades much was accomplished, and in the 1980s the development of this connection culminated in Rissanen's celebrated work [14][15][16], laying the foundations for the notion of stochastic complexity and the Minimum Description Length principle, or MDL.

Here we offer a first glimpse of a different connection, this time between information theory and number theory. In particular, we will show that basic information-theoretic arguments combined with elementary computations can be used to give a new proof for a classical result concerning the distribution of prime numbers. The problem of understanding this "distribution" (including the issue of exactly what is meant by that statement) has, of course, been at the heart of mathematics since antiquity, and it has led, among other things, to the development of the field of analytic number theory; e.g., Apostol's text [1] offers an accessible introduction and [2] gives a more historical perspective.

A major subfield is *probabilistic* number theory, where probabilistic tools are used to derive results in number theory. This approach, pioneered by, among others, Mark Kac and Paul Erdös from the 1930s on, is described, e.g., in Kac's beautiful book [11], Billingsley's review [3], and Tenenbaum's more recent text [18]. The starting point in much of the relevant literature is the

*Department of Informatics, Athens University of Economics and Business, Patission 76, Athens 10434, Greece. Email: yiannis@aueb.gr. Web: http://pages.cs.aueb.gr/users/yiannisk/.

following setup: For a fixed, large integer n, choose a random integer N from $\{1, 2, \ldots, n\}$, and write it in its unique prime factorization,

$$N = \prod_{p \le n} p^{X_p}, \tag{1}$$

where the product runs over all primes p not exceeding n, and X_p is the largest power $k \ge 0$ such that p^k divides N. Through this representation, the uniform distribution on N induces a joint distribution on the $\{X_p\ ;\ p \le n\}$, and the key observation is that, for large n, the random variables $\{X_p\}$ are distributed approximately like independent geometrics. Indeed, since there are exactly $\lfloor n/p^k \rfloor$ multiples of p^k between 1 and n,

$$\Pr\{X_p \ge k\} = \Pr\{N \text{ is a multiple of } p^k\} = \frac{1}{n}\Big\lfloor \frac{n}{p^k} \Big\rfloor \approx \Big(\frac{1}{p}\Big)^k, \quad \text{for large } n, \tag{2}$$

so the distribution of X_p is approximately geometric. Similarly, for the joint distribution of the $\{X_p\}$ we find,

$$\Pr\{X_{p_i} \ge k_{p_i} \text{ for primes } p_1, p_2, \ldots, p_m \le n\} = \frac{1}{n}\Big\lfloor \frac{n}{p_1^{k_1} p_2^{k_2} \cdots p_m^{k_m}} \Big\rfloor \approx \Big(\frac{1}{p_1}\Big)^{k_1}\Big(\frac{1}{p_2}\Big)^{k_2} \cdots \Big(\frac{1}{p_m}\Big)^{k_m},$$

showing that the $\{X_p\}$ are approximately independent.

This elegant approximation is also mathematically powerful, as it makes it possible to translate standard results about collections of independent random variables into important properties that hold for every "typical" integer N. Billingsley in his 1973 Wald Memorial Lectures [3] gives an account of the state-of-the-art of related results up to that point, but he also goes on to make a further, fascinating connection with the *entropy* of the random variables $\{X_p\}$.

Billingsley's argument essentially begins with the observation that, since the representation (1) is unique, the value of N and the values of the exponents $\{X_p\}$ are in a one-to-one correspondence; therefore, the entropy of N is the same as the entropy of the collection $\{X_p\}$,[1]

$$\log n = H(N) = H(X_p\ ;\ p \le n).$$

And since the random variables $\{X_p\}$ are approximately independent geometrics, we should expect that,

$$\log n = H(X_p\ ;\ p \le n) \approx \sum_{p \le n} H(X_p) \approx \sum_{p \le n} \Big[\frac{\log p}{p-1} - \log\Big(1 - \frac{1}{p}\Big)\Big], \tag{3}$$

where in the last equality we simply substituted the well-known expression for the entropy of a geometric random variable (see Section 2 for details on the definition of the entropy and its computation). For large p, the above summands behave like $\frac{\log p}{p}$ to first order, leading to the asymptotic estimate,

$$\sum_{p \le n} \frac{\log p}{p} \approx \log n, \quad \text{for large } n.$$

Our main goal in this paper is to show that this approximation can indeed be made rigorous, mostly through elementary information-theoretic arguments; we will establish:

[1] For definiteness, we take log to denote the natural logarithm to base e throughout, although the choice of the base of the logarithm is largely irrelevant for our considerations.

Theorem 1. As $n \to \infty$,

$$C(n) := \sum_{p \le n} \frac{\log p}{p} \sim \log n, \tag{4}$$

where the sum is over all primes p not exceeding n.[2]

As described in more detail in the following section, the fact that the joint distribution of the $\{X_p\}$ is asymptotically close to the distribution of independent geometrics is not sufficient to turn Billingsley's heuristic into an actual proof – at least, we were not able to make the two "$\approx$" steps in (3) rigorous directly. Instead, we provide a proof in two steps. We modify Billingsley's heuristic to derive a *lower bound* on $C(n)$ in Theorem 2, and in Theorem 3 we use a different argument, again going via the entropy of N, to compute a corresponding *upper bound*. These two combined prove Theorem 1, and they also give finer, finite-n bounds on $C(n)$.

In Section 2 we state our main results and describe the intuition behind their proofs. We also briefly review some other elegant information-theoretic arguments connected with bounds on the number of primes up to n. The appendix contains the remaining proofs.

Before moving on to the results themselves, a few words about the history of Theorem 1 are in order. The relationship (4) was first proved by Chebyshev [7][6] in 1852, where he also produced finite-n bounds on $C(n)$, with explicit constants. Chebyshev's motivation was to prove the celebrated prime number theorem (PNT), stating that $\pi(n)$, the number of primes not exceeding n , grows like,

$$\pi(n) \sim \frac{n}{\log n}, \quad \text{as } n \to \infty.$$

This was conjectured by Gauss around 1792, and it was only proved in 1896; Chebyshev was not able to produce a complete proof, but he used (4) and his finer bounds on $C(n)$ to show that $\pi(n)$ is of order $\frac{n}{\log n}$. Although we will not pursue this direction here, it is actually not hard to see that the asymptotic behavior of $C(n)$ is intimately connected with that of $\pi(n)$. For example, a simple exercise in summation by parts shows that $\pi(n)$ can be expressed directly in terms of $C(n)$:

$$\pi(n) = \frac{n+1}{\log(n+1)} C(n) - \sum_{k=2}^{n} \Big(\frac{k+1}{\log(k+1)} - \frac{k}{\log k} \Big) C(k), \quad \text{for all } n \ge 3. \tag{5}$$

For the sake of completeness, this is proved in the appendix.

The PNT was finally proved in 1896 by Hadamard and by de la Vallée-Pousin. Their proofs were not elementary – both relied on the use of Hadamard's theory of integral functions applied to the Riemann zeta function $\zeta(s)$; see [2] for some details. In fact, for quite some time it was believed that no elementary proof would ever be found, and G.H. Hardy in a famous lecture to the Mathematical Society of Copenhagen in 1921 [4] went as far as to suggest that "*if anyone produces an elementary proof of the PNT ... he will show that ... it is time for the books to be cast aside and for the theory to be rewritten.*" It is, therefore, not surprising that Selberg and Erdös' announcement in 1948 that they had produced such an elementary proof caused a great sensation in the mathematical world; see [9] for a survey. In our context, it is interesting to note that Chebyshev's result is again used explicitly in one of the steps of this elementary proof.

Finally we remark that, although the simple arguments in this work fall short of giving estimates precise enough for an elementary information-theoretic proof of the PNT, it may not be entirely unreasonable to hope that such a proof may exist.

[2] As usual, the notation "$a_n \sim b_n$ as $n \to \infty$" means that $\lim_{n\to\infty} a_n/b_n = 1$.

2 Primes and Bits: Heuristics and Results

2.1 Preliminaries

For a fixed (typically large) $n \geq 2$, our starting point is the setting described in the introduction. Take N to be a uniformly distributed integer in $\{1, 2, \ldots, n\}$ and write it in its unique prime factorization as in (1),

$$N = \prod_{p \leq n} p^{X_p} = p_1^{X_1} \cdot p_2^{X_2} \cdot \cdots \cdot p_{\pi(n)}^{X_{\pi(n)}},$$

where $\pi(n)$ denotes the number of primes $p_1, p_2, \ldots, p_{\pi(n)}$ up to n, and X_p is the largest integer power $k \geq 0$ such that p^k divides N. As noted in (2) above, the distribution of X_p can be described by,

$$\Pr\{X_p \geq k\} = \frac{1}{n} \Big\lfloor \frac{n}{p^k} \Big\rfloor. \quad \text{for all } k \geq 1, \tag{6}$$

This representation also gives simple upper and lower bounds on its mean $E(X_p)$,

$$\mu_p \;:=\; E(X_p) \;=\; \sum_{k \geq 1} \Pr\{X_p \geq k\} \leq \sum_{k \geq 1} \Big(\frac{1}{p}\Big)^k = \frac{1/p}{1 - 1/p} = \frac{1}{p-1}, \tag{7}$$

$$\text{and} \quad \mu_p \;\geq\; \Pr\{X_p \geq 1\} \geq \frac{1}{p} - \frac{1}{n}. \tag{8}$$

Recall the important observation that the distribution of each X_p is close to a geometric. To be precise, a random variable Y with values in $\{0, 1, 2, \ldots\}$ is said to have a geometric distribution with mean $\mu > 0$, denoted $Y \sim \text{Geom}(\mu)$, if $\Pr\{Y = k\} = \mu^k/(1+\mu)^{k+1}$, for all $k \geq 0$. Then Y of course has mean $E(Y) = \mu$ and its entropy is,

$$h(\mu) := H(\text{Geom}(\mu)) = -\sum_{k \geq 0} \Pr\{Y = k\} \log \Pr\{Y = k\} = (\mu + 1)\log(\mu + 1) - \mu \log \mu. \tag{9}$$

See, e.g., [8] for the standard properties of the entropy.

2.2 Billingsley's Heuristic and Lower Bounds on $C(n)$

First we show how Billingsley's heuristic can be modified to yield a lower bound on $C(n)$. Arguing as in the introduction,

$$\log n \overset{(a)}{=} H(N) \overset{(b)}{=} H(X_p\,;\; p \leq n) \overset{(c)}{\leq} \sum_{p \leq n} H(X_p) \overset{(d)}{\leq} \sum_{p \leq n} H(\text{Geom}(\mu_p)) \overset{(e)}{=} \sum_{p \leq n} h(\mu_p), \tag{10}$$

where (a) is simply the entropy of the uniform distribution, (b) comes from the fact that N and the $\{X_p\}$ are in a one-to-one correspondence, (c) is the well-known subadditivity of the entropy, (d) is because the geometric has maximal entropy among all distributions on the non-negative integers with a fixed mean, and (e) is the definition of $h(\mu)$ in (9). Noting that $h(\mu)$ is nondecreasing in μ and recalling the upper bound on μ_p in (7) gives,

$$\log n \leq \sum_{p \leq n} h(\mu_p) \leq \sum_{p \leq n} h(1/(p-1)) = \sum_{p \leq n} \Big[\frac{p}{p-1} \log\Big(\frac{p}{p-1}\Big) - \frac{1}{p-1} \log\Big(\frac{1}{p-1}\Big)\Big]. \tag{11}$$

Rearranging the terms in the sum proves:

Theorem 2. For all $n \geq 2$,

$$T(n) := \sum_{p \leq n} \Big\lceil \frac{\log p}{p-1} - \log\Big(1 - \frac{1}{p}\Big)\Big\rceil \geq \log n.$$

Since the summands above behave like $\frac{\log p}{p}$ for large p, it is not difficult to deduce the following lower bounds on $C(n) = \sum_{p \leq n} \frac{\log p}{p}$:

Corollary 1. [Lower Bounds on $C(n)$]

$$(i) \qquad \liminf_{n \to \infty} \frac{C(n)}{\log n} \geq 1;$$

$$(ii) \qquad C(n) \geq \frac{86}{125} \log n - 2.35, \qquad \text{for all } n \geq 16.$$

Corollary 1 is proved in the appendix. Part (i) proves half of Theorem 1, and (ii) is a simple evaluation of the more general bound derived in equation (16) in the proof: For any $N_0 \geq 2$, we have,

$$C(n) \geq \Big(1 - \frac{1}{N_0}\Big)\Big(1 - \frac{1}{1 + \log N_0}\Big) \log n + C(N_0) - T(N_0), \quad \text{for all } n \geq N_0.$$

2.3 A Simple Upper Bound on $C(n)$

Unfortunately, it is not clear how to reverse the inequalities in equations (10) and (11) to get a corresponding upper bound on $C(n)$ – especially inequality (c) in (10). Instead we use a different argument, one which is less satisfying from an information-theoretic point of view, for two reasons. First, although again we do go via the entropy of N, it is not necessary to do so; see equation (13) below. And second, we need to use an auxiliary result, namely, the following rough estimate on the sum, $\vartheta(n) := \sum_{p \leq n} \log p$:

$$\vartheta(n) := \sum_{p \leq n} \log p \leq (2 \log 2) n, \quad \text{for all } n \geq 2. \tag{12}$$

For completeness, it is proved at the end of this section.

To obtain an upper bound on $C(n)$, we note that the entropy of N, $H(N) = \log n$, can be expressed in an alternative form: Let Q denote the probability mass function of N, so that $Q(k) = 1/n$ for all $1 \leq k \leq n$. Since $N \leq n = 1/Q(N)$ always, we have,

$$H(N) = E[-\log Q(N)] \geq E[\log N] = E\Big[\log \prod_{p \leq n} p^{X_p}\Big] = \sum_{p \leq n} E(X_p) \log p. \tag{13}$$

Therefore, recalling (8) and using the bound (12),

$$\log n \geq \sum_{p \leq n} \Big(\frac{1}{p} - \frac{1}{n}\Big) \log p = \sum_{p \leq n} \frac{\log p}{p} - \frac{\vartheta(n)}{n} \geq \sum_{p \leq n} \frac{\log p}{p} - 2 \log 2,$$

thus proving:

Theorem 3. [Upper Bound] For all $n \geq 2$,

$$\sum_{p \leq n} \frac{\log p}{p} \leq \log n + 2 \log 2.$$

Theorem 3 together with Corollary 1 prove Theorem 1. Of course the use of the entropy could have been avoided entirely: Instead of using that $H(N) = \log n$ in (13), we could simply use that $n \geq N$ by definition, so $\log n \geq E[\log N]$, and proceed as before.

Finally (paraphrasing from [10, p. 341]) we give an elegant argument of Erdös that employs a cute, elementary trick to prove the inequality on $\vartheta(n)$ in (12). First observe that we can restrict attention to odd n, since $\vartheta(2n) = \vartheta(2n-1)$, for all $n \geq 2$ (as there are no even primes other than 2). Let $n \geq 2$ arbitrary; then every prime $n+1 < p \leq 2n+1$ divides the binomial coefficient,

$$B := \binom{2n+1}{n} = \frac{(2n+1)!}{n!(n+1)!},$$

since it divides the numerator but not the denominator, and hence the product of all these primes also divides B. In particular, their product must be no greater than B, i.e.,

$$\prod_{n+1<p\leq 2n+1} p \leq B = \frac{1}{2}\binom{2n+1}{n} + \frac{1}{2}\binom{2n+1}{n+1} \leq \frac{1}{2}(1+1)^{2n+1} = 2^{2n},$$

or, taking logarithms,

$$\vartheta(2n+1) - \vartheta(n+1) = \sum_{n+1<p\leq 2n+1} \log p = \log\Big[\prod_{n+1<p\leq 2n+1} p\Big] \leq (2\log 2)n.$$

Iterating this bound inductively gives the required result.

2.4 Other Information-Theoretic Bounds on the Primes

Billingsley in his 1973 Wald Memorial Lectures [3] appears to have been the first to connect the entropy with properties of the asymptotic distribution of the primes. Although there are no results in that work based on information-theoretic arguments, he does suggest the heuristic upon which part of our proof of Theorem 2 was based, and he also goes in the opposite direction: He uses probabilistic techniques and results about the primes to compute the entropy of several relevant collections of random variables.

Chaitin in 1979 [5] gave a proof of the fact that there are infinitely many primes, using algorithmic information theory. Essentially the same argument proves a slightly stronger result, namely that, $\pi(n) \geq \frac{\log n}{\log\log n+1}$, for all $n \geq 3$. Chaitin's proof can easily be translated into our setting as follows. Recall the representation (1) of a uniformly distributed integer N in $\{1, 2, \ldots, n\}$. Since p^{X_p} divides N, we must have $p^{X_p} \leq n$, so that each X_p lies in the range,

$$0 \leq X_p \leq \Big\lfloor \frac{\log n}{\log p} \Big\rfloor \leq \frac{\log n}{\log p},$$

and hence, $H(X_p) \leq \log\big(\frac{\log n}{\log p} + 1\big)$. Therefore, arguing as before,

$$\log n = H(N) = H(X_p\,;\, p \leq n) \leq \sum_{p\leq n} H(X_p) \leq \sum_{p\leq n} \log\Big(\frac{\log n}{\log 2} + 1\Big) \leq \pi(n)(\log\log n + 1),$$

where the last inequality holds for all $n \geq 3$.

It is interesting that the same argument applied to a different representation for N yields a marginally better bound: Suppose we write,

$$N = M^2 \prod_{p\leq n} p^{Y_p},$$

where $M \geq 1$ is the largest integer such that M^2 divides N, and each of the Y_p are either zero or one. Then $H(Y_p) \leq \log 2$ for all p, and the fact that $M^2 \leq n$ implies that $H(M) \leq \log\lfloor\sqrt{n}\rfloor$. Therefore,

$$\log n = H(N) = H(M, Y_{p_1}, Y_{p_2}, \ldots, Y_{p_{\pi(n)}}) \leq H(M) + \sum_{p \leq n} H(Y_p) \leq \frac{1}{2}\log n + \pi(n)\log 2,$$

which implies that $\pi(n) \geq \frac{\log n}{2\log 2}$, for all $n \geq 2$.

Finally we mention that in Li and Vitányi's text [13], an elegant argument is given for a more accurate lower bound on $\pi(n)$. Using ideas and results from algorithmic information theory, they show that, $\pi(n) = \Omega\big(\frac{n}{(\log n)^2}\big)$. But the proof (which they attribute to unpublished work by P. Berman (1987) and J. Tromp (1990)) is somewhat involved, and uses tools very different to those developed here.

Appendix

Proof of the Summation-by-parts Formula (5). Note that, since $\pi(k)-\pi(k-1)$ is zero unless k is prime, $C(n)$ can be expressed as a sum over all integers $k \leq n$,

$$C(n) = \sum_{2 \leq k \leq n} [\pi(k) - \pi(k-1)]\frac{\log k}{k}. \tag{14}$$

Each of the following steps is obvious, giving,

$$\begin{aligned}
\pi(n) &= \sum_{k=2}^{n} [\pi(k) - \pi(k-1)] \\
&= \sum_{k=2}^{n} [\pi(k) - \pi(k-1)]\frac{\log k}{k}\frac{k}{\log k} \\
&\overset{(a)}{=} \sum_{k=2}^{n} \Big[C(k) - C(k-1)\Big]\frac{k}{\log k} \\
&= \sum_{k=2}^{n} C(k)\frac{k}{\log k} - \sum_{k=2}^{n} C(k-1)\frac{k}{\log k} \\
&= \sum_{k=2}^{n} C(k)\frac{k}{\log k} - \sum_{k=1}^{n-1} C(k)\frac{k+1}{\log(k+1)} \\
&= \frac{n+1}{\log(n+1)}\,C(n) - \sum_{k=2}^{n}\Big(\frac{k+1}{\log(k+1)} - \frac{k}{\log k}\Big)C(k) \;-\; \frac{2}{\log 2}\,C(1),
\end{aligned}$$

where (a) follows from (14). This proves (5), since $C(1) = 0$, by definition. □

Proof of Corollary 1. Choose and fix any $N_0 \geq 2$ and let $n \geq N_0$ arbitrary. Then,

$$\log n \leq T(n) = T(N_0) + \sum_{N_0 < p \leq n} \Big[\frac{\log p}{p-1} - \log\Big(1 - \frac{1}{p}\Big)\Big] \leq T(N_0) + \sum_{N_0 < p \leq n} \Big[\frac{1}{1-\frac{1}{p}}\frac{\log p}{p} + \frac{N_0}{N_0 - 1}\frac{1}{p}\Big],$$

where the last inequality follows from the inequality $-\log(1-x) \le x/(1-\delta)$, for all $0 \le x \le \delta < 1$, with $\delta = 1/N_0$. Therefore,

$$\begin{aligned} \log n &\le T(N_0) + \sum_{N_0 < p \le n} \Big[\frac{1}{1-\frac{1}{N_0}} \frac{\log p}{p} + \frac{N_0}{N_0-1} \frac{\log p}{\log N_0} \frac{1}{p} \Big] \\ &= T(N_0) + \Big(\frac{N_0}{N_0-1} \Big) \Big(1 + \frac{1}{\log N_0} \Big) \Big(C(n) - C(N_0) \Big). \end{aligned} \tag{15}$$

Dividing by $\log n$ and letting $n \to \infty$ yields,

$$\liminf_{n\to\infty} \frac{C(n)}{\log n} \ge \frac{(N_0-1)\log N_0}{N_0(1+\log N_0)},$$

and since N_0 was arbitrary, letting now $N_0 \to \infty$ implies (i).

For all $n \ge N_0$, (15) implies,

$$C(n) \ge \Big(1 - \frac{1}{N_0} \Big) \Big(1 - \frac{1}{1+\log N_0} \Big) \log n + C(N_0) - T(N_0), \tag{16}$$

and evaluating this at $N_0 = 16$ gives (ii). □

Acknowledgments

Thanks to Peter Harremoës for spotting a small error in an earlier version of this paper.

References

[1] T.M. Apostol. *Introduction to Analytic Number Theory.* Springer-Verlag, New York, 1976.

[2] P.T. Bateman and H.G. Diamond. A hundred years of prime numbers. *Amer. Math. Monthly*, 103(9):729–741, 1996.

[3] P. Billingsley. The probability theory of additive arithmetic functions. *Ann. Probab.*, 2:749–791, 1974.

[4] H. Bohr. Address of Professor Harald Bohr. In *Proceedings of the International Congress of Mathematicians (Cambridge, 1950) vol. 1*, pages 127–134. Amer. Math. Soc., Providence, RI, 1952.

[5] G.J. Chaitin. Toward a mathematical definition of "life". In *Maximum entropy formalism (Conf., Mass. Inst. Tech., Cambridge, Mass., 1978)*, pages 477–498. MIT Press, Cambridge, Mass., 1979.

[6] P.L. Chebychev. Mémoire sur les nombres premiers. *J. de Math. Pures Appl.*, 17:366–390, 1852.

[7] P.L. Chebychev. Sur la totalité des nombres premiers inférieurs à une limite donnée. *J. de Math. Pures Appl.*, 17:341–365, 1852.

[8] T.M. Cover and J.A. Thomas. *Elements of Information Theory.* J. Wiley, New York, 1991.

[9] H.G. Diamond. Elementary methods in the study of the distribution of prime numbers. *Bull. Amer. Math. Soc. (N.S.)*, 7(3):553–589, 1982.

[10] G.H. Hardy and E.M. Wright. *An Introduction to the Theory of Numbers*. The Clarendon Press Oxford University Press, New York, fifth edition, 1979.

[11] M. Kac. *Statistical Independence in Probability, Analysis and Number Theory*. Published by the Mathematical Association of America. Distributed by John Wiley and Sons, Inc., New York, 1959.

[12] S. Kullback. *Information Theory and Statistics*. Dover Publications Inc., Mineola, NY, 1997. Reprint of the second (1968) edition.

[13] M. Li and P. Vitányi. *An Introduction to Kolmogorov Complexity and its Applications*. Springer-Verlag, New York, second edition, 1997.

[14] J. Rissanen. A universal prior for integers and estimation by minimum description length. *Ann. Statist.*, 11(2):416–431, 1983.

[15] J. Rissanen. Stochastic complexity. *J. Roy. Statist. Soc. Ser. B*, 49(3):223–239, 253–265, 1987. With discussion.

[16] J. Rissanen. *Stochastic Complexity in Statistical Inquiry*. World Scientific, Singapore, 1989.

[17] C.E. Shannon. A mathematical theory of communication. *Bell System Tech. J.*, 27:379–423, 623–656, 1948.

[18] G. Tenenbaum. *Introduction to Analytic and Probabilistic Number Theory*, volume 46 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1995.

MDL Model Averaging for Linear Regression

ERKKI P. LISKI[1] and ANTTI LISKI[2]
University of Tampere[1] and Tampere University of Technology[2]
Finland

Abstract

Estimators formed after model selection really are like mixtures of many potential estimators. Sometimes it is advantageous to smooth estimators across several models, rather than rely only on the model that is suggested by a single selection criterion. The main theme of this paper is the problem of selecting the weights for averaging across estimates obtained from a set of models. Some existing model average (MA) methods are based on exponential AIC or BIC weights (e.g. Burnham and Anderson 2002). Bayesian model averaging is a related technique (see e.g. Hoeting et al. 1999). Recently Leung and Barron (2006) and Hansen (2007) have developed methods for combining estimators from various models. This paper considers selecting the model weights by using Rissanen's MDL criterion and compares the potential performance of alternative MA estimators in simulation experiments.

2000 *Mathematics Subject Classification.* 62B10, 62J05, 62F99.

Key words or phrases. Model selection, NML, AIC, BIC, Mallows' C_p.

1 Introduction

In statistical practice one typically has multiple plausible models available. Model selection is most often regarded as a way to select just the best model, and then inference is conditioned on that model. In regression a common practice is to decide which variables to include in the model, and to use these variables to fit the response. A large number of criteria has been developed over the past few decades to select the best model.

It is known that model selection procedures can be unstable, as a small perturbation in the data may lead to significant changes in model choice. If the inference done with an estimate on the chosen model does not take into account model uncertainty, it often means underreporting of variability. Model averaging (MA) is an alternative to model selection. There is a large Bayesian literature on MA, for literature reviews see e.g. Draper (1995) and Hoeting et al. (1999). Given a set of models, we may find several plausible models according to some model selection criterion. In this case, it has been suggested estimation strategies that utilize more than just a single model. This entails a weighted average estimator for many alternative models. Buckland et al. (1997) suggested exponential AIC and BIC weights (see also Burnham and Anderson 2002). Hjort and Claeskens (2003) developed a general large-sample likelihood apparatus for MA estimators.

The Minimum Description Length (MDL) principle provides a generic solution to the model selection problem. By viewing models as a means of providing statistical descriptions of observed data, the comparison between competing models is based on the stochastic complexity (SC) of each description. The Normalized Maximum Likelihood (NML) form of the SC (Rissanen 1996) contains a component that may be interpreted as the parametric complexity of the model class. Once the SC for the data, relative to a class of suggested models, is calculated, it serves as

a criterion for selecting the optimal model with the smallest SC. This is the MDL principle (Rissanen 1978, 1986, 1996, 2000, 2007) for model choice.

In this paper we consider the NML density as an implementation of the MDL principle for model selection in the linear regression context, where attention is restricted to Gaussian linear models. Then we propose a model average estimator with weights selected by the MDL criterion. It turns out that, under squared error loss, the resulting mixture estimator usually performs better than the corresponding selection based estimator.

2 The Model

We have n pairs of observations $(y_1, \boldsymbol{x}_1), \ldots, (y_n, \boldsymbol{x}_n)$, where y_i is real valued and $\boldsymbol{x}_i$ is a $k_M \times 1$ vector, $1 \leq i \leq n$. Assume that the data follow a classical nonparametric regression model

$$y_i = \mu(\boldsymbol{x}_i) + \sigma\varepsilon_i, \qquad 1 = 1, \ldots, n, \tag{1}$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are independent random variables such that for each $1 \leq i \leq n$

$$E(\varepsilon_i|\boldsymbol{x}_i) = 0 \quad \text{and} \quad E(\varepsilon_i^2|\boldsymbol{x}_i) = 1$$

and the positive constant σ defines the scale of the additive error $\sigma\varepsilon_i$.

Model (1) is called a nonparametric regression model when μ belongs to some general (infinite dimensional) function class. Here we assume that μ is in the space of square integrable functions L_2 whose elements admit representations as infinite dimensional linear models for which

$$\mu(\boldsymbol{x}) = \sum_{j=1}^{\infty} \beta_j \varphi_j(\boldsymbol{x}) \tag{2}$$

for some set of known functions $\{\varphi_1, \varphi_2, \ldots\}$ and real valued coefficients $\beta_1, \beta_2, \ldots$. We assume that (2) converges in mean square, i.e.

$$E[\mu(\boldsymbol{x}) - \mu_m(\boldsymbol{x})]^2 \to 0 \quad \text{as} \quad m \to \infty,$$

where

$$\mu_m(\boldsymbol{x}) = \sum_{j=1}^{m} \beta_j \varphi_j(\boldsymbol{x}).$$

The practical significance of (2) is that any $\mu \in L_2$ may be well approximated by $\mu_m(\boldsymbol{x})$ with a finite number of m terms. In the sequel we denote generally $x_{ij} = \varphi_j(\boldsymbol{x}_i)$. Note that the above approach is a standard technique in nonparametric regression (see e.g. Efromovich 1999 and Eubank 1999). This approach is also similar to series estimators in econometrics (see e.g. Newey 1997).

Now the model (1) can be written as a linear model

$$y_i = \sum_{j \in \mathcal{M}_m} x_{ij}\beta_j + b_{im} + \sigma\varepsilon_i, \qquad i = 1, 2, \ldots, n, \tag{3}$$

where $\mathcal{M}_m = \{1, 2, \ldots, k_m\}$ with $k_m \leq n$,

$$b_{im} = \sum_{j=k_m+1}^{\infty} \beta_j x_{ij}$$

is the approximation error and the random errors $\varepsilon_1, \ldots, \varepsilon_n$ are like in (1). Here the quantity k_m plays the role of a smoothing parameter. Sometimes $\mu_m(\boldsymbol{x})$ is called a truncated series approximation of μ and k_m a truncation point.

To obtain an estimate of μ one may employ an approximating linear model by omitting b_i which effect is considered negligible. In matrix notation an approximating model $\mathcal{M}_m$ takes the form

$$\boldsymbol{y} = \boldsymbol{\mu}_m + \sigma\boldsymbol{\varepsilon}, \tag{4}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)'$, $\boldsymbol{\mu}_m = \boldsymbol{X}_m\boldsymbol{\beta}_m$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$ and $\boldsymbol{\beta}_m = (\beta_1, \ldots, \beta_{k_m})'$ is the $k_m \times 1$ vector of unknown regression coefficients. Here $\boldsymbol{X}_m$ is the $n \times k_m$ matrix with ij element x_{ij}. We shall consider a set of approximating models $\{\mathcal{M}_1, \ldots, \mathcal{M}_M\}$ such that $\mathcal{M}_1 \subseteq \mathcal{M}_2 \subseteq \cdots \subseteq \mathcal{M}_M \subseteq \{1, 2, \ldots, n\}$, where $\mathcal{M}_m$ refers to the model (4). We suppose that M is an integer for which the matrix $\boldsymbol{X}_{k_M}$ is of full column rank. Thus $k_1 \leq k_2 \leq \cdots \leq k_M$, and consequently all $\boldsymbol{X}_m$ with $1 \leq m \leq M$ are of full column rank.

Then the least-squares estimate of $\boldsymbol{\beta}_m$ is

$$\hat{\boldsymbol{\beta}}_m(\boldsymbol{y}) = (\boldsymbol{X}_m'\boldsymbol{X}_m)^{-1}\boldsymbol{X}_m'\boldsymbol{y},$$

and the corresponding estimate of $\boldsymbol{\mu}_m$ is

$$\hat{\boldsymbol{\mu}}_m = \boldsymbol{H}_m\boldsymbol{y}, \tag{5}$$

where $\boldsymbol{H}_m$ denotes the projection matrix $\boldsymbol{X}_m(\boldsymbol{X}_m'\boldsymbol{X}_m)^{-1}\boldsymbol{X}_m'$. Often the regression literature refers to the matrix $\boldsymbol{H}_m$ as the hat matrix. Denote $\boldsymbol{b}_m = (b_{1m}, \ldots, b_{nm})'$ and note that $\boldsymbol{\mu} = \boldsymbol{\mu}_m + \boldsymbol{b}_m$. Thus $(\boldsymbol{I} - \boldsymbol{H}_m)\boldsymbol{\mu} = (\boldsymbol{I} - \boldsymbol{H}_m)\boldsymbol{b}_m$, and consequently $\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_m = (\boldsymbol{I} - \boldsymbol{H}_m)\boldsymbol{b}_m - \sigma\boldsymbol{H}_m\boldsymbol{\varepsilon}$. Therefore the model error $r_m = \|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_m\|^2$ is

$$r_m = \boldsymbol{b}_m'(\boldsymbol{I} - \boldsymbol{H}_m)\boldsymbol{b}_m + \sigma^2\boldsymbol{\varepsilon}'\boldsymbol{H}_m\boldsymbol{\varepsilon} - 2\boldsymbol{b}_m'(\boldsymbol{I} - \boldsymbol{H}_m)\boldsymbol{H}_m\boldsymbol{\varepsilon}. \tag{6}$$

Taking the conditional expectation of r_m we obtain

$$E(r_m|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \boldsymbol{b}_m'(\boldsymbol{I} - \boldsymbol{H}_m)\boldsymbol{b}_m + \sigma^2 k_m,$$

since by assumptions of the model (1) the conditional expectation of the last term in the expression (6) is zero and $E(\sigma^2\boldsymbol{\varepsilon}'\boldsymbol{H}_m\boldsymbol{\varepsilon}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \sigma^2 k_m$.

Example. Assume that (2) is an orthogonal series representation for $\mu(x),\ x \in [0, 1]$, where

$$\beta_j = \int_0^1 \varphi_j(x)\mu(x)\,\mathrm{d}x, \qquad j = 1, 2, \ldots$$

and $\{1, \varphi_2, \varphi_3, \ldots\}$ is an orthonormal basis for $\mu \in L_2[0, 1]$. An example of such a basis is

$$\varphi_j(x) = \sqrt{2}\cos((j-1)\pi x), \qquad j = 1, 2, \ldots . \tag{7}$$

Other popular examples of orthogonal basis functions are orthogonal polynomials and wavelets. In this example we also assume that the basis functions are orthonormal with respect to the uniform design $x_j = (j - 1/2)/n$, $j = 1, \ldots, n$ like the cosine basis (7):

$$\sum_{i=1}^n \varphi_j(x_i)\varphi_k(x_i) = \begin{cases} 0, & j \neq k \\ n, & j = k \end{cases} \tag{8}$$

for all $j, k \in \{1, 2, \dots\}$ and $\varphi_1 \equiv 1$.

Using the orthogonality properties (8) it is easy to show that the least-squares estimate of β_j is

$$\hat{\beta}_j = \frac{1}{n} \sum_{i=1}^{n} \varphi_j(x_i) Y_i, \qquad j = 1, \dots, n.$$

If we assume the model like (4) and independent errors, then the coefficients $\hat{\beta}_1, \dots, \hat{\beta}_n$ are mutually independent and asymptotically

$$\hat{\beta}_j \sim \mathrm{N}\Big(\frac{1}{n} \sum_{i=1}^{n} \mu(x_i)\varphi_j(x_i), \frac{\sigma^2}{n} \Big), \qquad j = 1, \dots, n.$$

□

3 The MDL Model Selection and Averaging

In the MDL model selection we assume the approximating model (4) with normally distributed random errors, i.e. $\boldsymbol{\varepsilon} \sim \mathrm{N}_n(\mathbf{0}, \boldsymbol{I})$, where $\boldsymbol{I}$ is the $n \times n$ identity matrix. The response data $\boldsymbol{y}$ are modelled with the normal density functions

$$f(\boldsymbol{y}; \boldsymbol{\beta}_m, \sigma_m^2) = \frac{1}{(2\pi\sigma_m^2)^{n/2}} \exp\Big(-\frac{1}{2\sigma_m^2} \|\boldsymbol{y} - \boldsymbol{X}_m \boldsymbol{\beta}_m\|^2 \Big), \tag{9}$$

where $\|\cdot\|$ denotes the Euclidean norm of a vector and $1 \le m \le M$. Under these assumptions $\hat{\boldsymbol{\beta}}_m$ is the maximum likelihood (ML) estimate of $\boldsymbol{\beta}_m$ and

$$\hat{\sigma}_m^2 = \|\boldsymbol{y} - \hat{\boldsymbol{\mu}}_m\|^2 / n$$

the ML estimate of σ^2.

Consider the normalized maximum likelihood (NML) function (Rissanen 1996 and 2007, Barron, Rissanen and Yu 1998)

$$\hat{f}(\boldsymbol{y}; m) = \frac{f(\boldsymbol{y}; \hat{\boldsymbol{\theta}}_m(\boldsymbol{y}))}{C(m)}, \tag{10}$$

where $\hat{\boldsymbol{\theta}}_m = (\hat{\boldsymbol{\beta}}_m', \hat{\sigma}_m^2)'$ and

$$C(m) = \int f(\boldsymbol{y}; \hat{\boldsymbol{\theta}}_m(\boldsymbol{y}))\, \mathrm{d}\boldsymbol{y} \tag{11}$$

is the normalizing constant. Thus $\hat{f}(\boldsymbol{y}; m)$ is a density function, provided that $C(m)$ is bounded. Rissanen (1996) considers the NML function in the context of coding and modelling theory and takes

$$-\log \hat{f}(\boldsymbol{y}; m) = -\log f(\boldsymbol{y}; \hat{\boldsymbol{\theta}}_m(\boldsymbol{y})) + \log C(m) \tag{12}$$

as the "shortest code length" for the data $\boldsymbol{y}$ that can be obtained with the model $\mathcal{M}_m$ and calls it *the stochastic complexity* of $\boldsymbol{y}$, given $\mathcal{M}_m$. The last term in the equation (12) is called *the parametric complexity.*

Here we consider the model class

$$\mathcal{M}_m = \{\, f(\boldsymbol{y}; \boldsymbol{\theta}_m) : m \in \{1, \dots, M\} \,\} \tag{13}$$

defined by the normal densities (9). The aim of variable selection is to find the optimal value of index m. According to *the MDL* (Minimum Description Length) *principle* we seek to find the index value $m = \hat{m}$ that minimizes the stochastic complexity:

$$-\log \hat{f}(\boldsymbol{y}; \hat{m}) = \min_m \{-\log f(\boldsymbol{y}; \hat{\boldsymbol{\theta}}_m(\boldsymbol{y})) + \log C(m)\}.$$

Since $\hat{m}$ maximizes (10), we may call it the NML estimate of m within the model class $\mathcal{M}_m$.

For the normal distribution (9), however, the normalizing constant $C(m)$ is not bounded and hence the NML function is not defined. One approach to this problem is to constrain the data space properly (Rissanen 2000). For the constrained data space the stochastic complexity $C(m)$ is bounded, but it will depend on certain hyperparameters (Rissanen 2007 p. 116). The negative logarithm of $\hat{f}(\boldsymbol{y}; m)$ multiplied by 2 is given by

$$-2\log \hat{f}(\boldsymbol{y}; m) = n \log \hat{\sigma}_m^2 + k_m \log \frac{\|\hat{\boldsymbol{\mu}}_m\|^2}{\hat{\sigma}_m^2} - 2\log \Gamma\left(\frac{n-k_m}{2}\right) - 2\log \Gamma\left(\frac{k_m}{2}\right) + L(m) + c,$$

where the constant c is common to all models and therefore it can be ignored in model selection. The code length $L(m)$ for m is small and will be omitted. If we denote $\mathrm{MDL}_m = -2\log \hat{f}(\boldsymbol{y}; m)$ and omit $L(m) + c$, the NML model selection criterion takes the form (Hansen and Yu 2001, Liski 2006)

$$\mathrm{MDL}_m = n \log S_m^2 + k_m \log F_m + \log[k_m(n-k_m)],$$

where $S_m^2 = \|\boldsymbol{y} - \hat{\boldsymbol{\mu}}_m\|^2/(n-k_m)$ and $F_m = \|\hat{\boldsymbol{\mu}}_m\|^2/(k_m S_m^2)$.

Consider a mixture density

$$\sum_{m=1}^{M} w_m \hat{f}(\boldsymbol{y}; m) \quad \text{with} \quad \sum_{m=1}^{M} w_m = 1,$$

where $\hat{f}(\boldsymbol{y}; m)$ are NML densities and w_m nonnegative weights $1 \le m \le M$. If we select the model $m = \hat{m}$ and encode the data using the selected model $\hat{m}$, then the code length for the data is $\log[1/w_{\hat{m}} \hat{f}(\boldsymbol{y}; \hat{m})]$. On the other hand, the mixture model yields the code length $\log[1/\sum_m w_m \hat{f}(\boldsymbol{y}; m)]$ which is always shorter if $w_{\hat{m}} \neq 1$. Therefore, it seems advantageous to encode with a mixture (cf. also Liang and Barron 2005). However, the problem of finding the weight vector still remains.

Given the data $\boldsymbol{y}$, $\hat{f}(\boldsymbol{y}; m)$ can be interpreted as the likelihood of the model $\mathcal{M}_m$, $m = 1, 2, \ldots, M$. This leads to the NML distribution

$$\hat{p}(m; \boldsymbol{y}) = \frac{\hat{f}(\boldsymbol{y}; m)}{\sum_{i=1}^{M} \hat{f}(\boldsymbol{y}; i)} = \frac{\exp(-\mathrm{MDL}_m/2)}{\sum_{i=1}^{M} \exp(-\mathrm{MDL}_i/2)} \tag{14}$$

for models (13). Thus the MDL distribution (14) may be used to define the empirical selected weight vector

$$\hat{\boldsymbol{w}} = (\hat{p}(1; \boldsymbol{y}), \ldots, \hat{p}(M; \boldsymbol{y}))' \tag{15}$$

(cf. Rissanen 2007, Subsection 5.2.2).

4 Alternative Model Average Estimators

It is well-known that a model selection procedure can be unstable, as small changes in the data may lead to significant changes in model choice. The inference done with a single estimate $\hat{\boldsymbol{\beta}}_m$

based on the chosen model $\mathcal{M}_m$ does not take into account model uncertainty, and therefore may be too optimistic. To deal with uncertainty in model selection we study model average estimation. Let $\hat{\boldsymbol{\mu}}_w$ denote an MA estimator of $\boldsymbol{\mu}$. It is a convex combination of estimators (5) such that

$$\hat{\boldsymbol{\mu}}_w = \sum_{m=1}^{M} w_m \hat{\boldsymbol{\mu}}_m = \sum_{m=1}^{M} w_m \boldsymbol{H}_m \boldsymbol{y} = \Big(\sum_{m=1}^{M} w_m \boldsymbol{H}_m \Big) \boldsymbol{y} = \boldsymbol{H}_w \boldsymbol{y}, \tag{16}$$

where $\boldsymbol{H}_w$ denotes the implied hat matrix $\sum_{m=1}^{M} w_m \boldsymbol{H}_m$. Note that although every hat matrix $\boldsymbol{H}_m$ is idempotent, the implied hat matrix $\boldsymbol{H}_w$ is generally not. Selecting the model weights by the NML distribution (15) yields an operational model average estimator.

Bayesian model averaging is widely used in the literature and so we refer to these works by, among others, Draper (1995) and for literature reviews see Hoeting, et al. (1999). An alternative can be based on the analogue of Bayesian model probabilities for frequentist statistics. Such a weigh scheme has been implied in a series of papers by Akaike (see e.g. Akaike 1978 and 1979) and expounded further by Buckland, et al. (1997) and Burnhan and Anderson (2002). Akaike's suggestion derives from the Akaike information criterion (AIC). The Akaike weights are defined as

$$w_m \propto \exp(-\operatorname{AIC}_m /2)$$

normalized to have unit sum. In the present context of ML estimation $\operatorname{AIC}_m = n \log \hat{\sigma}_m^2 + 2k_m$. For a Bayesian the weights

$$w_m \propto \exp(-\operatorname{BIC}_m /2)$$

can serve as a rough approximation to the posterior probabilities for models $\mathcal{M}_m$, where $\operatorname{BIC}_m = n \log \hat{\sigma}_m^2 + k_m \log n$ is the Bayesian information criterion (Schwarz 1978) for $\mathcal{M}_m$.

Hansen (2007) proposed the Mallows' criterion

$$C(\boldsymbol{w}) = \|\boldsymbol{y} - \hat{\boldsymbol{\mu}}_w\|^2 + 2\sigma^2 \boldsymbol{k}' \boldsymbol{w}, \tag{17}$$

where $\boldsymbol{k} = (k_1, \ldots, k_M)'$. The empirical Mallows' weight vector $\hat{\boldsymbol{w}}$ is selected so that the criterion (17) attains its minimum. In practice σ^2 should be replaced with some consistent estimator. In the simulation experiments our choice is $\sigma^2 = \hat{\sigma}_M^2$. There is no closed form solution to minimizing of (17) and the weight vector must be found numerically.

Leung and Barron (2006) considered MA estimators under the Gaussian model (9) when σ^2 is assumed to be known. They defined the weights to be

$$w_m \propto \exp(-\alpha \frac{\hat{r}_m}{2\sigma^2}), \qquad \alpha > 0, \tag{18}$$

where

$$\hat{r}_m = \|\boldsymbol{y} - \hat{\boldsymbol{\mu}}_m\|^2 + \sigma^2 (2k_m - n) \tag{19}$$

is an unbiased estimate of the model error

$$r_m = \|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_m\|^2 \tag{20}$$

in the sense that $E(\hat{r}_m) = E(r_m)$ (Akaike 1970 and 1973, Mallows 1973, Stein 1973 and 1981). The tuning parameter α adjusts the degree of concentration of the weights on the models with small model error estimates. They derived simple and accurate bounds on (20) and its estimate (19). The weights (18) are not directly operational, however, since σ^2 and α should be estimated. Note that also $C(\boldsymbol{w}) - n\sigma^2$ is an unbiased estimate of the model error in the sense that

$$E[C(\boldsymbol{w}) - n\sigma^2] = E(r_w),$$

where $r_w = \|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_w\|^2$ (Hansen 2007).

5 Simulations

In this section we report on simulation investigations of the performance potential of the MDL, AIC, BIC and Mallows' MA estimators. The setting is the regression

$$y_i = \sum_{j=1}^{K} x_{ij}\beta_j + \varepsilon_i, \tag{21}$$

where $K = 100$ and $k_M < K$. We fix in (21) $x_{i1} = 1$, and the remaining elements x_{ij} are mutually independent and follow the normal distribution $\mathrm{N}(0, 1)$. The errors ε_i are independent of x_{ij} and are $\mathrm{N}(0, 1)$. We impose the structure of gradual decay on β and varied

$$\beta_j = c\frac{A(a)}{j^{a+1/2}}$$

with different values a and c. Here $A(a)$ is such a normalizing constant that $\mathrm{Var}(\sum_{j=1}^{K} x_{ij}\beta_j) = c^2$. The parameter a controls the speed of decay and the coefficient c determines the value of $R^2 = c^2/(1 + c^2)$, a measure of explained variation. In the reported experiments the sample size is set to $n = 50$ and the maximum model order k_M varied between 10, 20, 30, 40, 45.

Let MDL($\boldsymbol{w}$), AIC($\boldsymbol{w}$), BIC($\boldsymbol{w}$), MalC($\boldsymbol{w}$) denote the MDL, AIC, BIC and Mallows' MA estimators, respectively. To evaluate estimators we compute the model error (20). We then summarise the overall performance by computing the average model error (AME) over 10 000 iterations in each of our various set-ups. We normalize the AME by that of the estimator $\hat{\boldsymbol{\beta}}_M$, so that unity indicates equivalence with $\hat{\boldsymbol{\beta}}_M$ in the AME sense. Then the AME curves as a function of R^2 are displayed.

In the first experiment (Figure 1) we illustrate the effect of a (the speed of decay) on the performance of estimators. The parameter a varied between 0.5, 1.0, 1.5 and 2 when $M = 40$ and $n = 50$. The results from the first experiment show that the performance of MDL($\boldsymbol{w}$) and BIC($\boldsymbol{w}$) are close to each others. Overall, the AIC($\boldsymbol{w}$) estimator has clearly higher AME relative to its competitors. For large values of a (1.5 and 2) the MalC($\boldsymbol{w}$) has slightly higher AME than MDL($\boldsymbol{w}$) and BIC($\boldsymbol{w}$), but for the values 0.5 and 1 there are some crossings of the AME curves. In all cases the AME curves are increasing functions of R^2.

The second experiment (Figure 2) depicts the dependence of the AME on the maximal model order k_M that varied between 10, 20, 30 and 45. The main message is clear: the AME curves of the MA estimators MDL($\boldsymbol{w}$), BIC($\boldsymbol{w}$) and MalC($\boldsymbol{w}$) are pretty close to each others and their performance relative to the AIC($\boldsymbol{w}$) improves when k_M increases. Results of further simulation experiments (not reported here) confirmed the finding, that the performance of MDL($\boldsymbol{w}$) and BIC($\boldsymbol{w}$) relative to the AIC($\boldsymbol{w}$) improves when M/n increases.

Note that although the AIC($\boldsymbol{w}$) does not do too well in our experiments, its relative performance improves when M/n is small. Hansen (2007) reported simulation results showing that the AIC($\boldsymbol{w}$) and MalC($\boldsymbol{w}$) are superior to the BIC($\boldsymbol{w}$) when M/n is small. In his experiments both M and n varied (K is sufficiently large).

Finally, in Figure 3 we illustrate by simulation how MDL($\boldsymbol{w}$) is superior to the MDL model selection estimator in the AME sense. The MDL($\boldsymbol{w}$) consistently outperforms the MDL model selection estimator.

References

Akaike, H. (1970). Statistical Predictor Identification. *Annals of the Institute of Statistical Mathematics*, 22, 203–217.

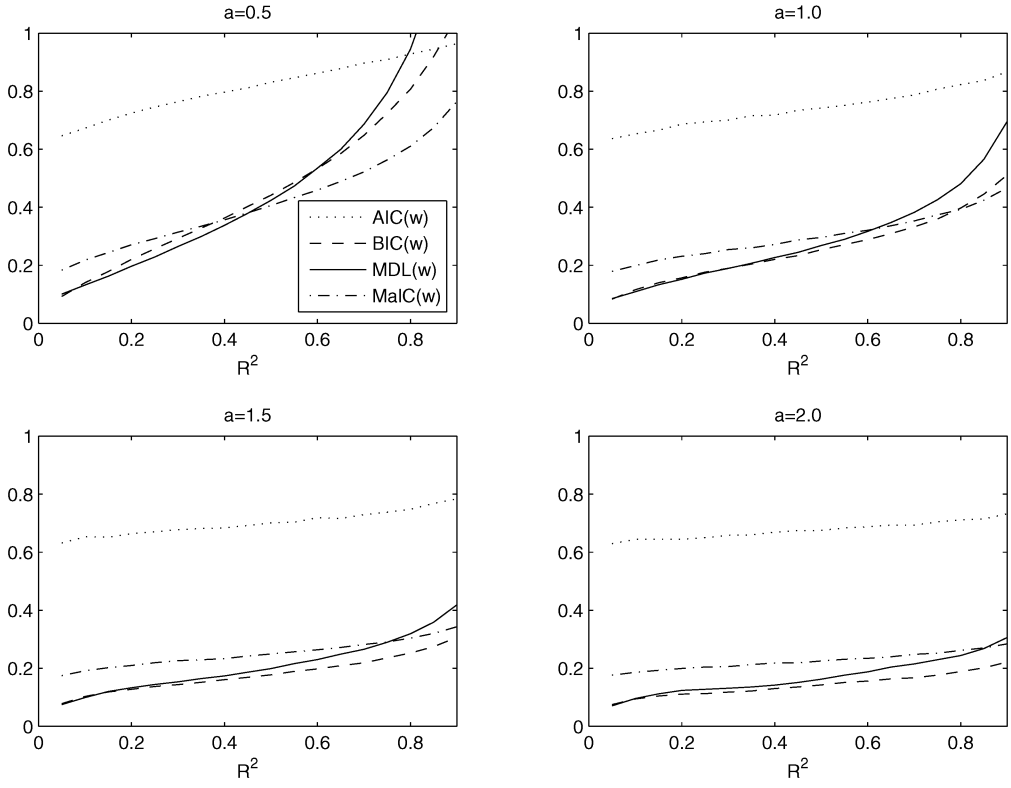


Figure 1: The AME curves of MDL($\boldsymbol{w}$), BIC($\boldsymbol{w}$), AIC($\boldsymbol{w}$) and MalC($\boldsymbol{w}$) as a function of R^2 for $a = 0.5$, 1.0, 1.5 and 2.0 when $n = 50$ and $M = 40$.

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. Pages 267–281 *in* B. N. Petrov, and F. Csaki, (eds.) *Second International Symposium on Information Theory.* Akademiai Kiado, Budapest.

Akaike, H. (1978). A Bayesian Analysis of the Minimum AIC Procedure. *Annals of the Institute of Statistical Mathematics*, 30, 9–14.

Akaike, H. (1979). A Bayesian Extension of the Minimum AIC Procedure to Autoregressive Model Fitting. *Biometrika*, 66, 237–242.

Barron, A. R., Rissanen, J. and Yu, B. (1998). The MDL principle in modeling and coding. *Special Issue of Information Theory to Commemorate 50 Years of Information Theory*, 44, 2743–2760.

Buckland, S. T. Burnham, K. P. and Augustin, N. H. (1999). Model Selection: An Integral Part of Inference. *Biometrics*, 53, 603–618.

Burnham, K. P. and Anderson D. R. (2002). *Model Selection and Multi-model Inference.* New York, Springer-Verlag.

Draper, D. (1995). Assessment and Propagation of Model Uncertainty. *Journal of the Royal Staistical Society, B* 57, 45–70.

Efromovich, S. (1999). *Nonparametric Curve Estimation.* New York, Springer-Verlag.

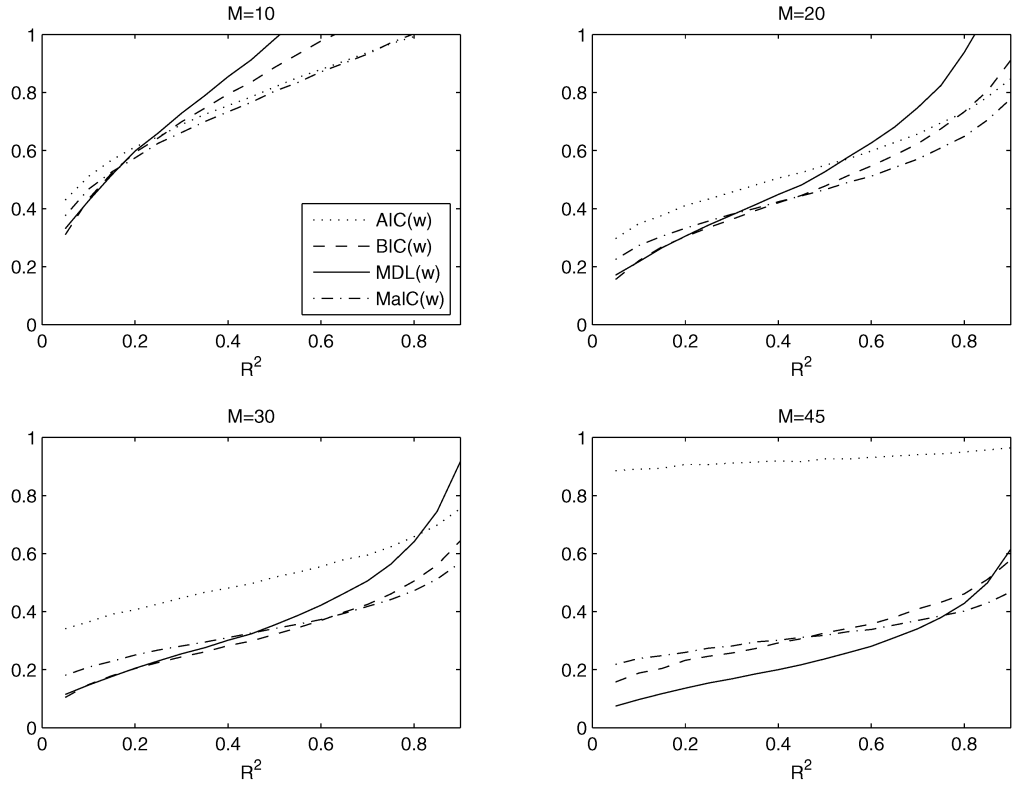


Figure 2: The AME curves of MDL($\boldsymbol{w}$), BIC($\boldsymbol{w}$), AIC($\boldsymbol{w}$) and MalC($\boldsymbol{w}$) as a function of R^2 for $M = 10$, 20, 30 and 45 when $n = 50$ and $a = 1$.

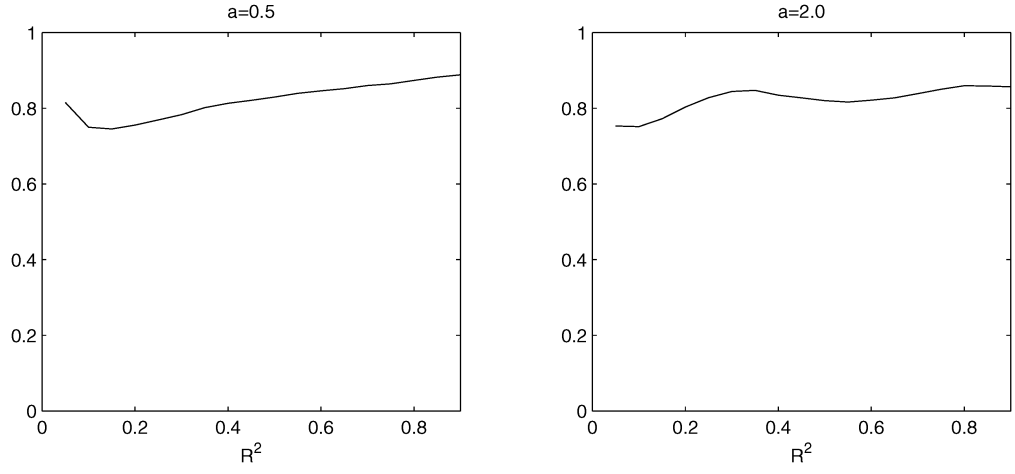


Figure 3: The AME of the MDL($\boldsymbol{w}$) normalized by that of the MDL model selection estimator when $n = 50$ and $M = 40$.

Eubank, R. L. (1999). *Nonparametric Regression and Spline Smoothing*. New York, Springer-Verlag.

Hansen, B. E. (2007). Least Squares Model Averaging. *Econometrica* 75 (4), 1175–1189.

Hansen, A. J. and Yu, B. (2001). Model Selection and the Principle of Minimum Description Length. *Journal of the American Statistical Association*, 96, 746–774.

Hjort, N. L. and Claeskens, G. (2003). Frequentist Model Average Estimators. *Journal of the American Statistical Association*, 98, 879–899.

Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999). Bayesian Model Averaging: A Tutorial. *Statiscal Science*, 14, 382–417.

Leung, G. and Barron, A. R. (2006). Information Theory and Mixing Least-Squares Regressions. *IEEE Transactions on Information Theory*, IT-52, No. 8, 3396–3410.

Liang, F. and Barron, A. R. (2005) Exact Minimax Predictive Density Estimation and MDL. In: Grünwald, P. D., Myung, I. J. and Pitt, M. A. (Eds.). *Advances in Minimum Description Length: Theory and Applications*. Cambridge, MA: MIT Press.

Liski, E. P. (2006). Normalized ML and the MDL Principle for Variable Selection in Linear Regression In: Liski, E. P., Isotalo, J., Niemelä, J., Puntanen, S., and Styan, G. P. H. (Eds.). *Festschrift for Tarmo Pukkila on His 60th Birthday*, 159–172. Tampere, Department of Mathematics, Statistics and Philosophy.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, 15, 661–675.

Newey, W. K. (1997). Convergence Rates and Asymptotic Normality for Series Estimators. *Journal of Econometrics*, 79, 147–168.

Rissanen, J. (1978). Modeling by Shortest Data Description. *Automatica*, 14, No. 1, 465–471.

Rissanen, J. (1986). Stochastic Complexity and Modeling. *Annals of Statistica*, 14, 1080–1100.

Rissanen, J. (1996). Fisher Information and Stochastic Complexity. *IEEE Transactions on Information Theory*, IT-42, No. 1, 40–47.

Rissanen, J. (2000). MDL Denoising. *IEEE Trans. on Information Theory*, IT-46, No. 1, 2537–2543.

Rissanen, J. (2007). *Information and Complexity and in Statistical Modeling*. New York, Springer-Verlag.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461–464.

Stein, C. (1973). Estimation of the Mean of a Multivariate Normal Distribution. In: *Proceedings of the Prague Symposium in Asymptotic Statistics, 1973*, 345–381.

Stein, C. (1981). Estimation of the Mean of a Multivariate Normal Distribution. *Annals of Statistics*, 9, 1135–1151.

What is Information?

Jerzy Konorski*
Faculty of Electronics, Telecomm. & Informatics
Gdansk University of Technology
80-952 Gdansk
Poland
jekon@eti.pg.gda.pl

Wojciech Szpankowski†
Department of Computer Science
Purdue University
W. Lafayette, IN 47907
U.S.A.
spa@cs.purdue.edu

Dedicated to our colleague and friend Jorma Rissanen, the philosopher of information theory

Abstract

The notion of *information* has so far been quantified mostly in statistical terms, giving rise to Shannon's information theory and the principles of digital data transmission. Studies of systems involving complex, intelligent, and autonomous agents, not uncommon in contemporary science, call for a new look at the measures of information that place importance on context, semantics, structures, and rationality. In this essay we propose a framework for measuring information inspired by the *event-driven approach*. We then illustrate our definition with several examples ranging from distributed computer systems to biology and economics.

1 Introduction

In this essay we muse on the notion of *information*, hoping to capture some of its essential aspects and provoke a discussion. We point out the need for a new definition of information that might be applied in contemporary science and engineering ranging from biology to chemistry, economics, and physics. We shall proceed inductively, giving examples from which hopefully a formal framework will arise.

Advances in information technology, the abundance of information systems and services, the much-trumpeted advent of information society, or even the Information Age (recently embodied in the communities of Web 2.0), almost obscure the fact that the common buzzword – the *i*-word – remains undefined in its generality, though considerable collective effort was harnessed into its understanding (cf. [6, 8, 20, 21, 31, 36, 38]). Shannon wrote in [32]: "The word "information" has been given many different meanings ... it is likely that at least a number of these will prove sufficiently useful in certain applications and deserve further study and permanent recognition."

Shannon's successful theory of information defines *statistical* information that quantifies to what extent a recipient of data can reduce statistical uncertainty associated with its source by observing the output of a source-recipient channel. Shannon also argued in his 1948 paper: "These semantic aspects of communication are irrelevant to the engineering problem." The channel error rate, on the other hand, does matter: for example, with a 50% binary error rate, the amount of statistical information sent through a binary symmetric channel is zero. But it

*The work of this author is sponsored by the AFOSR Grant FA8655-08-1-3018 and the Ministry of Science and Higher Education, Poland, Grant PBZ-MNiSW-02/II/2007.

†This work was supported in part by the NSF Grants CCR-0208709, CCF-0513636, and DMS-0503742, NIH Grant R01 GM068959-01, and the AFOSR Grant FA8655-08-1-3018.

seems that the intuitive understanding of information cannot be formalized without bringing into the picture the *timing* of data (consider a train departure notice served a recipient after the stated departure time), *spatial* aspect of information (imagine the same notice arriving at a different location), the *objective* its recipient wants to achieve (consider the same notice served a recipient not going anywhere), and the knowledge of the recipient's internal rules of conduct, or *protocol* for short (consider a recipient at the output of a channel with a high bit error rate, whose protocol dictates that the channel be regarded as perfect, hence received data be used *bona fide*).

The *context* of data cannot be abstracted from, either. Even at a high error rate some information may be recovered from the context e.g., a math textbook transmitted over such a channel might still be recognized as such. This point becomes particularly valid in the realm of biosystems – most biological information depends on where it is retrieved e.g., its location within a cell, a piece of DNA or protein. This important aspect is not yet well understood or analyzed in information theory. Biology is above all about context, and so a periodic pattern, while containing less statistical information than a random sequence, may contain a lot more *biological* information. In fact, in a recent paper [11] the authors argue that a random string and an exactly duplicated string add nothing or almost nothing to a biological information content. On the other hand, any context-dependent information measure must take into account the relationship between a given string and other related strings.

So what is information? In this essay, following C. F. von Weiszsäcker, we first argue that that there is *no absolute meaning* of information. Then, using an event-driven approach, we propose a definition that encompasses two of Weiszsäcker's premises, namely that "Information is only that which produces information" (relativity) and "Information is only that which is understood" (rationality) [36]. We then present some examples illustrating new aspects of information within the framework that we adopt here. We conclude with remarks suggesting some future work and leading to more questions. As a matter of fact, we hope to put forward some educated questions as to the issues and tools that lie before researchers interested in information, rather than come up with definite answers.

A preliminary version of this essay was prepared for the October 2005 workshop *Information Beyond Shannon* at Orlando FL. We thank the participants of the workshop for lively and constructive comments, some of which have found their way into the present version.

2 Event-Driven Approach

An intuitive relationship between data (any sequence of interpretable symbols) and information is that data may or may not carry information. One may observe that *a piece of data carries information if it helps its recipient achieve some objective.* In fact, this observation, stated more or less explicitly, was the point of departure of early textbooks on information technology [28]. There has been little formal apparatus, however, to quantitatively account for all its facets. To generalize and add precision we observe that *a piece of data carries information if it can impact a recipient's objective, under a given protocol and within a given context.*

Thus information has a flavor of *relativity* and *rationality*: it derives from the recipient's knowledge (gathered from the context), capability (implied by its protocol), and the pursued

objective. Underlying the latter are also temporal and spatial aspects, for the usefulness of data may depend on the timing and location of its generation and reception.

We offer more examples to illustrate the role of protocol. Clearly, a speaker of Chinese (a more knowledgeable recipient) can make out a lot more of a textbook on VLSI circuit design written in that language than a non-speaker (a less knowledgeable recipient). However, the latter can by default regard some strings of symbols that do not look like an ethnic language as a blueprint of a VLSI circuit; hence, the protocol can make up for the lack of knowledge (if applied only to the drawings in the textbook) or bring about catastrophic results (if applied to the Chinese characters of the text body). Furthermore, a duplicate notice of a train departure time does not contribute to the objective of catching that train and therefore is of no informational value (the recipient already knows it), unless the recipient's protocol stipulates that at least one confirmation of the train departure time be received. Finally, in a secret sharing scheme, decryption keys separated in time and space seem to carry zero information until they are brought together into one location at the same time. Indeed, information carried by data is not only related to its context, but also to a recipient's protocol, the rule dictating how to handle received data.

Having said this, we still need a quantitative definition, an analogue of Shannon's statistical information, retaining the flavors of relativity and rationality, and with a potential to reflect temporal and spatial aspects. Can we attempt formal definitions of the *amount* of information and maximum amount of information carried by a channel – capacity – without a lengthy specification of the semantics of data? One possibility is to adopt an event-driven approach which we sketch below.

An event-driven approach offers a few advantages. First, it is well-established among the engineering community thanks to the work of C. A. R. Hoare and others in the field of operating systems and distributed algorithms. Second, it is discrete and timeless in nature, yet allows for dynamic characterization of systems evolving in continuous time. Finally, it is able to formalize such intuitions as causality and consistency of local views without specifying the semantics of the involved events. At the same time, it generalizes the data-information relationship: now it is events that may or may not carry information; in particular, an event may correspond to reception of a piece of data, a clock tick etc. The event-driven approach-inspired formalization goes along the following lines:

- A universe is populated by systems (living organisms, institutions, communities, software agents, Internet domains etc.) pursuing specified *objectives*.
- A system's current state is expressible through a number of system variables (e.g., memory content, parameter configuration, operational status of constituent subsystems); an observable change of state marks an *event* (e.g., clock tick, execution of a specific operation, reception of a piece of data from another system).
- A partial order on the set of events may be defined as the order in which the events occur at a given system (with simultaneous events not precluded); the set of events preceding an event is called the *context* of the event.
- Events may have attributes e.g., time of occurrence and semantics, as defined by the

system's *protocol* i.e., specification of how the system handles the events in order to pursue its objectives.

We would like to regard information as another (measurable) attribute of an event reflecting our previous discussion. To this end, define an objective functional that maps a system's protocol P and a context C (a sequence of events related to the communication between the source and recipient systems) into any space with ordered points; further we only consider the one-dimensional Euclidean space i.e., real axis. The idea is that P along with C determine $\mathit{objective}(P, C)$, the extent to which the recipient system's objective has been achieved. For simplicity assume that P remains fixed throughout the system's lifetime. In particular, monotonicity of $\mathit{objective}(P, C)$ in C is desirable, for it implies that successive events help achieve the objective. That is, we would like $\mathit{objective}(P, C + E) \geq \mathit{objective}(P, C)$ for any event E and context C, where $C + E$ is the new context extended by event E. Before defining a possible measure of information we discuss more examples to support our approach.

Example 1. [*Decimal Representation*] Assume that a system's objective is to learn the number π and P has the system compute successive decimal digits approximating π from below. Each computed digit is then regarded as an event and $\mathit{objective}(P, C)$ is a real-valued function monotonically increasing and asymptotically stabilizing in C. As an illustration, imagine we are drawing circles of circumferences $3, 3.1, 3.14, 3.141$ etc., and measure the respective diameters i.e., $.9549, .9868, .9995, .9998$, which asymptote to the ideal 1.

Example 2. [*Shannon Information*] In Shannon's information theory [31] *objective* is defined as *statistical ignorance of the recipient* or *statistical uncertainty of the recipient.* It is measured by the number of binary decisions to recognize the event E, that is, $-\log P(E)$, where $P(E)$[1] is the probability as computed by the recipient. For various generalizations the reader is referred to [18, 21]. Observe also that spatial and temporal aspects of information were mostly left out in Shannon's theory.

Example 3. [*Distributed Information*] In an (N, N)-threshold secret sharing scheme [29], N subkeys of the decryption key roam among geographically dispersed systems. By the protocol P, the event corresponding to the reception of another subkey from a fellow system does not give access to the secret unless receptions of *all* the other subkeys are already in C. Likewise, an observed pixel of a digital image may increase a viewer's ability to understand the image depending on how many neighboring pixels have already been observed (this example illustrates that the event-driven approach also covers spatial, rather than temporal, contexts – in general, there is no difficulty evaluating the objective functional as long as events are processed sequentially). In passing we may wonder what is the difference between distributed and local information; is one bit *here* equivalent of one bit *there*?

Example 4. [*Temporal Information*] The impulses exchanged along nerves or processed within neural cells of a living organism critically depend on timing e.g., a stimulus generated by a pain

[1] We shall write $P(E)$ for the probability of an event E since from the context one easily distinguishes it from the protocol P.

receptor is useless if it arrives too late to administer a defensive gesture. Spatio-temporal coding is widely acknowledged to be the most important information processing feature of networks of neurons [17]. This remarkable coding scheme forces groups of neurons, involved in the same learning or memory retrieval task, to communicate and process information through *timing* and *location*. The spatial aspect of this form of coding arises due to functional differentiation of neurons. Usually, neurons involved in processing of related tasks or designed to respond to similar cues are clustered in the same region of the brain. Examples include the well known receptor maps in the olfactory bulb, the cochleotopic (frequency) regions in the primary auditory cortex (where different regions of neurons respond to different frequencies in the stimulus), and the topographic feature maps in the visual area of mammalian brains (where neurons discriminate against different orientations of the visual stimulus).

Similarly, clock ticks are relevant when judging the usefulness of successive speech or video frames sent over a packet network. Since they share network resources with unpredictable data traffic, the frames arrive at the destination irregularly, as quantified by delay jitter. Premature and overdue arrivals (events with too few or too many clock ticks in the context) are equally unwelcome, though are handled in a different way: the former have to be buffered before delivery and the latter are typically discarded. In general, incurred delay (e.g., in biological and computer networks) is a nontrivial issue not yet successfully addressed by information theory [14].

Example 5. [*Wireless Networks*] In a wireless ad hoc network, each mobile terminal (MT) can physically communicate only within its transmission range. To maintain network-wide connectivity and so achieve the objective of each MT (i.e., a high throughput of data packets), P prescribes setup and maintenance of relay paths between remote MTs. These are temporary in nature due to the terminal mobility. Thus there are both path discovery and path disruption events; consequently and somewhat counterintuitively, $objective(P, C)$ may not increase in C. Recent research [12, 13, 15] indicates that for $objective(P, C)$ to increase in C, a quite unorthodox P is needed that restricts paths to two-hop and trades buffer space for bandwidth, a thought at the core of the so-called time capacity paradox.

Example 6. [*Herding, Web 2.0, DNA*] The conclusion of the previous example suggests that $objective(P, C)$ increases in C provided that P is somehow "rational." Unfortunately, studies of the so called *herding effects* disprove that intuition too: an individual contemplating an action behaves rationally by observing and following the majority of other individuals (as shown by Bayesian analysis). After a short while, however, further observations provide no more insight into the benefits of the action [4]. Perhaps, then, one can only assert that $objective(P, C)$ is nondecreasing in C provided that P is rational? There are examples that run counter even that intuition. Imagine a user session with a Web search engine in which too much data, or the presence of conflicting data, paralyze the user's ability to act; from another perspective, a growing number of users contributing their ideas to a digital Web 2.0 community may at some point prevent a required broad consensus. Equally daunting is the well-known fact that the sheer amount of data contained in a biological database (e.g., human genome) may blur patterns leading to the identification of relevant human traits. In fact, in a massive data set, such as a biological database or results of an Internet search, the situation is not unlike a radio channel crossed by interfering signal paths: what is noise for one receiver (query) may well be

useful information for another. The problem of discovering and quantifying the amount of useful information thus acquires a new meaning.

Example 7. [*Cooperative and Noncooperative Settings*] Consider now a system where the objective functionals defined at different subsystems are in conflict (e.g., the problem of Byzantine generals, DoS or selfish attacks on communication protocols such as IEEE 802.11 [19]). The simplest example are two data sources contending for a multiple access channel (e.g., ALOHA system). Various forms of P may then calibrate the sources' behavior from cooperative (where *objective*(P, C) increases in the total number of data transmission events in C i.e., in the overall channel utilization) to noncooperative (where *objective*(P, C) increases in the number of own data transmission events) to malicious (where *objective*(P, C) decreases in the number of the other source's data transmission events).

Example 8. [*Rissanen's Stochastic Complexity and MDL*] Included in *objective*(P, C) may be the cost of the very recognition and interpretation of C. Imagine a recipient knowing that the source uses an optimal code for its stream of data, but having to learn on the fly the stochastic mechanism according to which the source generates data. As time passes, the model reveals itself to the recipient who can then hypothesize about data sent. In 1978 Rissanen [23, 24, 25, 27] introduced the *Minimum Description Length* (MDL) principle, an incarnation of Occam's Razor stating that the best hypothesis is the one that gives the shortest description of data. Realizing that Kolmogorov complexity is uncomputable, MDL selects a code for which the *total* description length of code and data is minimal. Rissanen stresses that we should "make no assumptions" about a *true* data generating process. In practice, we must restrict the class of process models.

More precisely, let $\mathcal{M}_k = \{Q_\theta : \theta \in \Theta\}$ be a set of finitely parameterized distributions of dimension k. One could argue, and some did, that the best (shortest) description of a string $x = (x_1, \ldots, x_n)$ should be $-\log Q_\theta(x)$, as suggested by the Kraft correspondence for prefix codes. As pointed out by Rissanen and others, this is not correct since one must also describe the distribution Q_θ itself. But this can be accomplished by a *universal data compression* algorithm. Rissanen proposed two possible solutions, namely *two-part codes* and the *normalized maximum likelihood* (NML) code that we briefly describe below.

In the two-part coding, one first describes a distribution Q_θ and then describes the string x using Q_θ. Let $\mathcal{C}$ be a code that maps Θ to $\{0,1\}^*$. Then the *stochastic complexity* $S(x)$ is

$$S(x) = \min_{\theta \in \Theta} \left[\text{length}(\mathcal{C}(\theta)) - \log Q_\theta(x) \right],$$

and the MDL principle states that one should choose θ^* that achieves the above minimum.

In the normalized maximum likelihood (NML) code, first the parameter $\hat{\theta}$ is chosen to minimize $-\log Q_\theta(x)$ (as in the classical maximum likelihood estimate), and then the "ideal" code-length $-\log Q_{\hat{\theta}}(x)$ is used as a yardstick against which code performance is measured. This leads to the so called *minimax problem* that finds the best code for the worst distribution and the worst data. It is well known [3, 9, 26] that the regret function defined as

$$r_n^*(\mathcal{M}) = \min_Q \max_x \left[\log \frac{Q_{\hat{\theta}}(x)}{Q_\theta(x)} \right]$$

achieves its optimal value $\log \sum_x Q_{\hat{\theta}}(x)$ for the *normalized maximum distribution*

$$Q_{NML}(x) = \frac{Q_{\hat{\theta}}(x)}{\sum_x Q_{\hat{\theta}}(x)}.$$

The optimal code-length is then $-\log Q_{NML}(x)$. Rissanen in [26] proved, among others, that the minimax regret for $\mathcal{M}_k$ is

$$r_n^*(\mathcal{M}_k) = \frac{k}{2} \ln \frac{n}{2\pi} + \ln \int_\theta \sqrt{|I(\theta)|} d\theta + o(1)$$

where $I(\theta)$ is the Fisher information. Further generalization can be found in [3, 9, 16, 33]. In passing, one still may ask why to restrict analysis to prefix codes? Is there a fundamental lower bound for general codes (cf. [1, 34, 37])?

3 Information and Capacity

We are now in a position to set out a framework for defining the amount of information consistent with the intuition based on our examples and discussion.

Definition 1 *The amount of information carried by event E in context C as perceived at a system with protocol P is*

$$\text{info}_{P,C}(E) = \text{weight}[\mathit{objective}(P, C + E), \mathit{objective}(P, C)], \tag{1}$$

where "weight" measures the change between two (objective) points according to the order defined on the space of values of the objective functional.

Thus an event only carries nonzero information if it changes $\mathit{objective}(P, C)$, a statement consistent with the intuitive flavors of relativity and rationality. The dependence on P and C reflects the obvious observation that one and the same event can produce different information at different recipients, locations, and times. Also note that in view of Example 6, negative information is not unthinkable. In fact, this might lead to an interesting distinction: *nonconfoundable* systems, contrasted with *confoundable* ones, are those whose protocol P precludes negative information regardless of C. One can imagine a smart Web user always able to remove conflicting data from the context and proceed monotonically towards an objective. Whether and for what types of data sources and objective functionals such P exist is an open problem. Finally, it is natural to surmise that both P and C are subject to various constraints implied, respectively, by the systems' architecture and the nature of the event sources. In the spirit of Shannon, one may define the channel capacity between the event source and the recipient as a maximum-type measure on a collection of amounts of information carried by successive events, within the regions of feasible P and C (subject to the said constraints). For a given $C = (E_1, E_2, \ldots)$ and $E_i \in C$, let $C_i := (E_1, \ldots, E_{i-1})$ be the prefix of C consisting of events preceding E_i.

Definition 2 *The capacity of the channel between the event source and recipient is*

$$\text{capacity} = \max_{P \text{ feasible}} \max_{C \text{ feasible}} F\left(\{\text{info}_{P,C_i}(E_i),\ i \geq 1\}\right). \tag{2}$$

for some function $F(\cdot)$.

Depending on the specific case, the function F can be conveniently defined as the sum of all elements of its set argument, the maximum element, etc. If the total amount of information and the feasible C are infinite, it may be convenient to define F as the limiting average information per event:

$$F\left(\{\text{info}_{P,C_i}(E_i),\ i \geq 1\}\right) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \text{info}_{P,C_i}(E_i) \tag{3}$$

provided the limit exists. With so structured a definition it is possible to confine interest to the inner maximum if for some reasons P is regarded as the only feasible.

We now return to some of the previous examples in order to give a quantitative illustration of Definitions 1 and 2.

Example 1. [continuation] In Example 1, the objective in a given context can be measured as the deviation of the corresponding diameter from the ideal 1, so that the amount of information carried by successively computed digits of π is the difference between successive deviations. Hence, the event "3" carries $(1-0)-(1-.9549) = .9549$, "1" carries $(1-.9549)-(1-.9868) = .0319$, "4" carries $(1-.9995)-(1-.9868) = .0127$, the other "1" carries $(1-.9998)-(1-.9995) = .0003$ units of information etc. If F is as in (3), then the capacity of such a channel is zero: an infinite number of events carry a finite total information.

Example 2. [continuation: *Shannon Information and Temporal Capacity*] Does the event-driven approach include Shannon information as a special case? As suggested by the previous discussion, the objective in Shannon information can be viewed as the statistical uncertainty. Consider a memoryless channel and a memoryless source transmitting symbols chosen from a finite set according to some probability distribution. The amount of information carried by an event $E = (x, y)$, where x and y are respectively the transmitted and received symbol, can be measured by the difference between the recipient's degree of uncertainty as to x before and after reception of y i.e.,

$$\text{info}_{P,C}(E) = -\log P(x) - [-\log P(x|y)].$$

Note that because of our memoryless setting, there is no explicit dependence on C. If the channel is noiseless (error-free), then $P(x|y) = 1$ iff $x = y$, thus $\text{info}_{P,C}(E) = -\log P(x)$. Taking F in our definition of capacity as in (3), we find for a context $C = (E_1, \ldots, E_n)$

$$F\left(\{\text{info}_{P,C_i}(E_i),\ i \geq 1\}\right) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \text{info}_{P,C_i}(E_i) = -\sum_{x} P(x) \log P(x) = H(X).$$

Here, X is a random variable describing the source. The right-hand side of the above relationship we recognize as Shannon's entropy of the source. In a noisy channel, the limiting average information per event becomes

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \text{info}_{P,C_i}(E_i) = \sum_{(x,y)} P(x,y)(-\log P(x) - [-\log P(x|y)]) = I(X;Y),$$

where Y is a random variable describing the output of the channel. This we recognize as Shannon's mutual information. It is easy to see now that, with the protocol P fixed and the

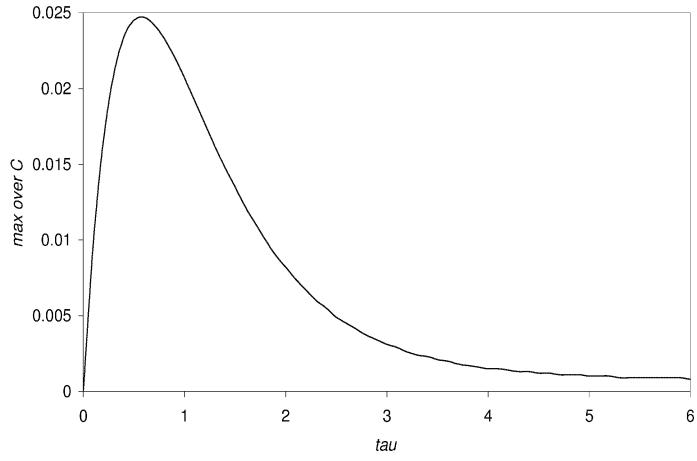


Figure 1: Temporal capacity as a function of τ.

maximization only taken over C, the channel capacity in the sense of Definition 2 coincides with Shannon's capacity

$$\max_{C \text{ feasible}} \frac{1}{n} \sum_{i=1}^{n} \text{info}_{P,C_i}(E_i) \sim \max_{P(X)} I(X;Y),$$

where the right-hand side maximum is taken over all possible distributions of X. This is so because in our memoryless setting, any feasible context must have been produced by some $P(X)$.

Recall that Shannon's celebrated channel coding theorem states that as long as the transmission rate does not exceed the channel capacity, information can be sent with as small a frequency of errors as desired provided unlimited time and resources are available to encode and decode the message. Thus, temporal (or spatial) aspects of information are not considered. However, they can easily be addressed in this setting, and the relevance of optimizing the protocol can be demonstrated.

Consider a memoryless binary symmetric channel with "temporal errors": the longer a binary symbol takes to reach the recipient, the lower the probability of a successful transmission. Each transmitted symbol is received in error with probability $\Phi(\varepsilon, t)$, where ε is the "instant error" rate and t is the incurred channel delay. A plausible function Φ should increase from 0 to 1 for $\varepsilon \in (0,1)$, and increase from ε to 1 as t varies between 0 and ∞. Assume further that the recipient's protocol P enables determination of t when a symbol is received, and if $t \geq \tau$ prescribes erasure of the received symbol. Thus $X \in \{0,1\}$ and $Y \in \{0,1,erasure\}$. Let the source be memoryless with $P(X=1) = p$ and the channel delay be represented by a random variable D with a known probability distribution function $F(t) = P(D < t)$. We only need to slightly modify the amount of information carried by an event $E = (x,y)$, namely

$$\text{info}_{P,C}(x,y) = \begin{cases} \log P(x|y,\ D < \tau) - \log P(x) & \text{if } y = 0,1 \\ 0 & \text{if } y = \text{erasure.} \end{cases}$$

Then the limiting average information per event again coincides with the mutual information

$I(X;Y)$. To calculate the latter let us introduce the conditional probability

$$\phi := P(Y=1|X=0, D<\tau) = P(Y=0|X=1, D<\tau) = \frac{\int_0^\tau \Phi(\varepsilon,t)dF(t)}{F(\tau)},$$

which plays the role of the "temporal error" rate. Standard calculation yields

$$I(X;Y) = [H_b((1-\phi)(1-p)+\phi p) - H_b(\phi)]F(\tau), \tag{4}$$

where $H_b(u) = -u\log u - (1-u)\log(1-u)$ is the binary entropy function for $u \in [0,1]$. By a similar argument as above, the maximization of (4) over p corresponds to maximization over feasible C in (2). The maximum is attained at $p = 1/2$ and yields

$$[1 - H_b(\phi)]F(\tau),$$

the maximum mutual information for a given τ, that is, Shannon channel capacity. In Figure 1 we plot this quantity against τ assuming $F(t) = 1-e^{-t}$ and $\Phi(\varepsilon,t) = 1-(1-\varepsilon)^{t+1}$ for $\varepsilon = 0.3$. (Note that the average channel delay is the time unit.) We see that in the case of a stringent delay bound the capacity of the channel is adversely affected by frequent erasures; when the delay bound becomes ineffective, frequent temporal errors dominate infrequent erasures to produce a somewhat counterintuitive drop in mutual information. We now recall that τ represents the recipient's protocol P; hence if we maximize over τ, which corresponds to the outer maximum in (2), we get a clear estimate of the channel capacity.

Example 3. [continuation] Let N subkeys move at random and independently of one another among $A \times A$ stations regularly spaced within a square area. For simplicity let the movements of

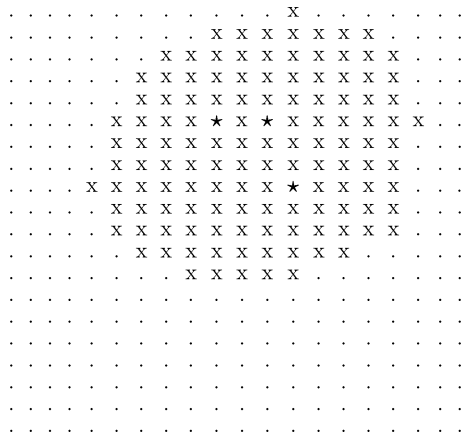


Figure 2: Access to the secret ($N = 3, A = 20, d = 8$).

the subkeys be synchronized to unit time slots. In each slot a station can improve its objective by having temporary access to the secret, which happens if it is within Euclidean distance d from each of the subkeys. This is illustrated in Fig. 2, where the current subkey locations are marked "*" and stations with access to the secret are marked "x."

Assume that the larger N, the more valuable the secret, which results in each "x" station improving its objective proportionally to N. If all the stations act as one system, then an event E defines N new locations of the subkeys. Here, N is a parameter of the set of feasible C, d is a parameter of P, and

$$\text{info}_{P,C}(E) = N \times \{\#\text{ of stations having access to secret}\}.$$

Note that since the movement of the subkeys is memoryless, there is no explicit dependence on C. The limiting average information per event per station, which thus equals N times the probability of access per event, is plotted in Figure 3 (obtained by a Monte Carlo simulation). The maximum of each curve corresponds to the channel capacity as expressed by the inner maximum in (2) i.e., with respect to C, given P.

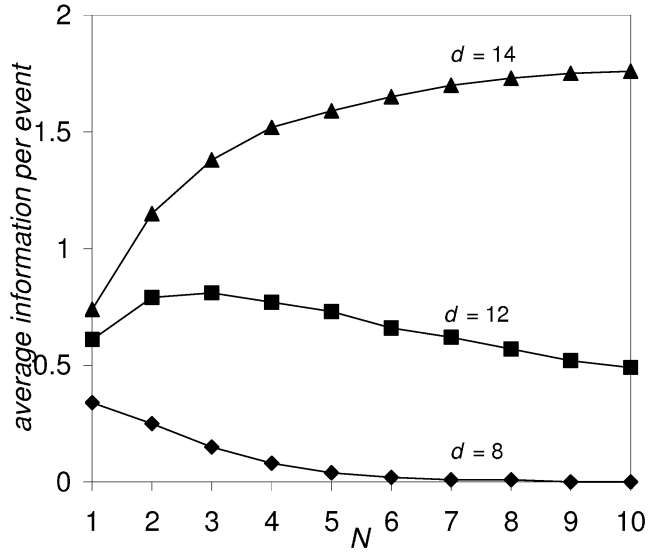


Figure 3: Normalized average information per event for secret sharing ($A = 20$).

Example 7. [continuation: *Noncooperative Settings; Value of Information*] We should point out that calculating the capacity in the above framework seems to be particularly difficult in a distributed system featuring multiple autonomous agents. For example, in economics one often considers the value of information [21] which measures (perhaps in dollars) the difference between the payoffs of an informed action and an uninformed action. Consider a simple entry deterrence game [10]. Suppose an Internet service provider (ISP) has a major business client (Incumbent) who can use either Standard or Premium service. Another business (Entrant) is considering entry i.e., becoming the ISP's client with only Standard service available. Both Incumbent and Entrant choose their strategies (Standard/Premium and enter/not enter) simultaneously and without prior coordination. Thus a one-shot noncooperative game arises with payoffs given in Table 1. Here, K is the surcharge Incumbent pays for Premium service ($0 \leq K \leq 3$). While the other payoff components are rather arbitrary, the relationships between them are important:

Table 1: Entry deterrence payoffs (arbitrary units): Incumbent's (*left*) and Entrant's (*right*)

	enter	not enter
Premium	$3-K,\quad -1$	$5-K,\quad 0$
Standard	$2,\quad 1$	$3,\quad 0$

- Entrant's payoff is neutral if she does not enter, otherwise it is negative if Incumbent chooses Premium (Entrant pays entrance fee, but receives a less-than-fair share of ISP's resources), and is positive if Incumbent chooses Standard (Entrant receives a fair share of ISP's resources),
- Incumbent is better off if Entrant does not enter (there is no competition for ISP's resources), and given Entrant's choice, Incumbent's well-being depends on K e.g., $K > 2$ ($K < 1$) makes Standard (Premium) a dominating strategy.

It is easy to see that for $K > 2$ the only Nash equilibrium (NE) is (Standard, enter), while for $K < 1$ the only NE is (Premium, not enter). For $1 \leq K \leq 2$ there exists a unique NE in mixed strategies:

$$\left(\frac{1}{2} * \text{Premium} + \frac{1}{2} * \text{Standard}, (1-(K-1)) * \text{enter} + (K-1) * \text{not\_enter}\right),$$

where $p * s + q * s'$ denotes a mixed strategy "play s with probability p and s' with probability q."

A more realistic model assumes that (a) Entrant has only an estimate K' of K, (b) Incumbent knows both K and K', moreover, is in a position to communicate K to Entrant if she thinks it worthwhile. The question is whether and when Incumbent will indeed communicate K and how much information passes between Incumbent and Entrant. Let E_0^{Ent} and E_0^{Inc} denote the events of acquiring the knowledge of K' by Entrant, and of K and K' by Incumbent. The objective is the expected payoff and the protocols of both players prescribe NE strategies. For Entrant, the NE strategy is:

$$s(E_0^{Ent}) = \begin{cases} \text{not enter}, & K' < 1 \\ [1-(K'-1)] * \text{ enter} + (K'-1) * \text{ not\_enter} & 1 \leq K' \leq 2 \\ \text{enter} & K' > 2. \end{cases}$$

Incumbent, who knows the above strategy, chooses hers so as to maximize the expected payoff. The result is obvious except when $1 \leq K \leq 2$ and $1 \leq K' \leq 2$. Incumbent's expected payoff conditioned on choosing Premium is then

$$[1-(K'-1)] \cdot (3-K) + (K'-1) \cdot (5-K)$$

and conditioned on choosing Standard is

$$[1-(K'-1)] \cdot 2 + (K'-1) \cdot 3.$$

Incumbent chooses Premium if the former payoff is greater than the latter, i.e., if $K < K'$, and Standard if $K > K'$. (If $K = K'$, Incumbent plays 1/2 * Premium + 1/2 * Standard.) Hence,

Table 2: Expected payoffs at NE: Incumbent's (*left*) and Entrant's (*right*)

	$K' < 1$	$1 \le K' \le 2$	$K' > 2$
$K > 2$	3, 0	$[1-(K'-1)]\cdot 2+(K-1)\cdot 3$, $[1-(K'-1)]\cdot 1$	2, 1
$1 \le K \le 2$	$5-K$, 0	**as above** (if $K > K'$) $[1-(K'-1)]\cdot 2+(K'-1)\cdot 3$, 0 (if $K = K'$) **as below** (if $K < K'$)	2, 1
$K < 1$	$5-K$, 0	$[1-(K'-1)]\cdot(3-K)+(K'-1)\cdot(5-K)$, $[1-(K'-1)]\cdot(-1)$	$3-K$, -1

$$s(E_0^{Inc}) = \begin{cases} \text{Premium} & K < 1 \text{ or } (1 \le K \le 2 \text{ and } K' < 1) \\ & \text{[Entrant does not enter]} \\ & \text{or } (1 \le K \le 2 \text{ and } 1 \le K' \le 2 \text{ and } K < K') \\ & \text{[Entrant plays mixed strategy]}, \\ \text{1/2 * Premium + 1/2 * Standard} & 1 \le K \le 2 \text{ and } 1 \le K' \le 2 \text{ and } K = K', \\ \text{Standard} & K > 2 \text{ or } (1 \le K \le 2 \text{ and } K' > 2) \\ & \text{[Entrant enters]} \\ & \text{or } (1 \le K \le 2 \text{ and } 1 \le K' \le 2 \text{ and } K > K') \\ & \text{[Entrant plays mixed strategy]}. \end{cases}$$

For the payoffs in Table 1, the possible Incumbent's and Entrant's expected payoffs at NE are given in Table 2, which both players can compute using game theory basics, but only Incumbent knows which row gives actual payoffs.

Imagine now that just before the game, Incumbent has a chance to communicate K and thus correct Entrant's wrong estimate K' (denote the corresponding event E_1^{Ent}). This she will not consider worthwhile if $K > 2$ and $K' \le 2$ for it would encourage Entrant's entry, thereby decreasing Incumbent's expected payoff (from 3, or a value between 2 and 3, to 2). Similarly for $1 \le K \le 2$ and $K' < 1$. If $1 \le K \le 2$ and $K' > 2$, the communication of K would lead to the mixed strategy NE; this will increase Incumbent's payoff (from 2 to a value between 2 and 3), but at the same time decrease Entrant's payoff (from 1 to 0). If Entrant is noncooperative, she will ignore E_1^{Ent} regarding it as incredible (presumably part of Incumbent's entry deterrence strategy). Knowing that, Incumbent will simply communicate nothing. In all the above cases, the channel between Incumbent and Entrant is as good as closed (unable to carry information).

Only when $K < 1$ and $K' \ge 1$ will the communication of K become worthwhile from Incumbent's viewpoint and credible to Entrant, for Entrant's expected payoff then would rise (from -1, or a value between -1 and 0, to 0). According to (1), the amount of information received by Entrant in this case is:

$$\begin{aligned}\text{info}(E_0, E_1) &= \text{payoff}(E_0^{Ent}, E_1^{Ent}) - \text{payoff}(E_0^{Ent}) \\ &= \begin{cases} 0-(-1)=1, & K' > 2 \\ 0-(-1)\cdot[1-(K'-1)] = 1-(K'-1), & 1 \le K' \le 2. \end{cases}\end{aligned}$$

If the game is played repeatedly, e.g., on a session basis, and each time K and K' are drawn independently from a uniform probability density function on $[0, 3]$, then their joint probability density is $1/9$. The average amount of information received by Entrant (i.e., the average increase in Entrant's objective) per game is:

$$\text{avg\_info} = \frac{\int_0^1 \left[\int_1^2 [1-(K'-1)]dK' + \int_2^3 1dK'\right] dK}{9} = \frac{3}{2\cdot 9}$$

What if neither Incumbent nor Entrant were noncooperative and so K were communicated and E_1^{Ent} were accepted regardless of the expected payoffs? Then

$$\begin{aligned}9\cdot\text{avg\_info} &= \int_0^1 \left[\int_1^2 [1-(K'-1]dK' + \int_2^3 1dK'\right] dk + \int_1^2 \left[\int_0^1 0dK' + \int_2^3 (-1)dK'\right] dK \\ &\quad + \int_2^3 \Big] \int_0^1 1dK' + \int_1^2 (K'-1)dK'\Big] dK = 2,\end{aligned}$$

thus avg_info= 2/9. In summary, the 25% difference between the latter two figures reflects the reduction of channel capacity merely due to noncooperative nature of the involved protocols

4 Final Remarks

Our definition (1) is somewhat similar in spirit to that of the *value of information* discussed in Luenberger [21]. In the presence of a single source of uncertainty about the state of the world among the many possible states, Luenberger considers a decision-maker maximizing the average payoff and calculates the net benefit of receiving an imperfect signal about the true state of the world. Clearly, the net benefit is zero if the signal does not reduce the uncertainty. In such a Bayesian setting, negative values of information are impossible. Although we propose a broader framework, with context and protocol explicitly accounted for, we still need a generalization of imperfect signals (or imperfect events in our wording); ours is a faultless communication system, where events do not get corrupted or misinterpreted. While partly justified by contemporary high-quality transmission and processing infrastructure, this is a serious restriction.

A no less fundamental issue is related to the very notion of information. The foregoing discussion focused upon *communicable* information, which is why events played so central a role: a system remaining in one and the same state cannot change its perception of the achieved objective. However, another strong intuition of information holds it to be embedded in the structure of an object and thus independent of any rational activity – this we may refer to as *structural information*. F. Brooks articulates in [5] : "Shannon and Weaver performed an inestimable service by giving us a definition of information and a metric for information as communicated from place to place. We have no theory however that gives us a metric for

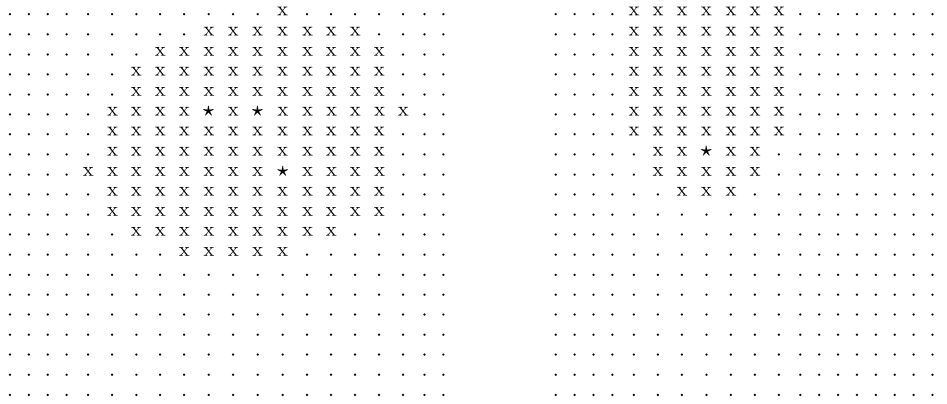


Figure 4: Access to the secret with two different subkey locations ($N = 3, A = 20, d = 8$)

the information embodied in structure ... this is the most fundamental gap in the theoretical underpinning of information and computer science. ... A young information theory scholar willing to spend years on a deeply fundamental problem need look no further." Along with spatial and temporal aspects of information, this is, in our opinion, the most urgent challenge facing our community.

Yet another understanding arises from a conjecture of an organizing principle, a hidden mechanism behind a given object, and the amount of structural information may be related to the remaining uncertainty as to the nature or parameters of the hidden mechanism. In this way a sequence with clear patterns of symbols may be attributed more structural information than a piece of gibberish after all. This is particularly true about biological information as discussed above and in [11]. To illustrate our point, consider again our secret sharing scheme, as in Example 3, and suppose we only know the current "x" stations. Two sets of such stations, corresponding to two different subkey locations, are depicted in Fig. 4. They may be regarded as two states of our system, or two objects of some informational value, the hidden mechanism being the movement of the subkeys. Where can the subkeys be? They can be no further than d from any "x" station, which leaves a number of possible subkey locations marked "?" in Fig. 5. (Particular N-tuples of locations can then be eliminated at the cost of more computation.) We might conclude that the left state (object) contains more structural information than the right one.

In summary, we propose to fundamentally enhance six decades of work in information theory by incorporating the following elements that were, to large extent, not adequately addressed in the past and therefore threaten to raise severe impediments to diverse applications:

Structure: We still lack measures and meters to appraise the amount of organization and information embodied in artifacts and natural objects.

Delay: In typical interacting systems, timeliness of signals is essential to function. Often timely delivery of partial information carries higher value than delayed delivery of complete information. For example, in a signaling cascade associated with a specific cell function, delay or loss of signals

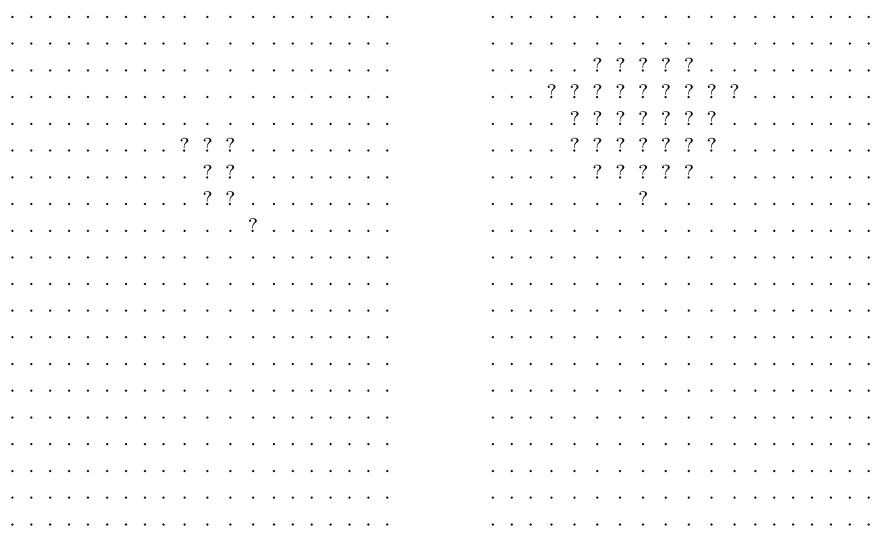


Figure 5: Possible subkey locations

can be lethal.

Space: In interacting systems, spatial localization often limits information exchange – with obvious disadvantages as well as benefits. These benefits typically result from reduction in interference (common examples range from wireless systems to immune response).

Information and control: In addition to delay-bandwidth tradeoffs discussed above, systems often allow modifications to underlying design patterns (e.g., network topology, power distribution and routing in networks). Simply stated, information is exchanged in space and time for decision making, thus timeliness of information delivery along with reliability and complexity constitute basic objectives.

Semantics. In many scientific contexts, one is interested in signals, without knowing precisely what these signals represent (e.g., DNA sequences, spike trains between neurons, whale songs), but little more than that can be assumed a priori. Is there a general way to account for the actual "meaning" of signals in a given context?

Dynamic information. In a complex network, information is not just communicated but also processed and even generated along the way. How can such considerations of dynamic sources be incorporated into an information-theoretic model?

Learnable information. One may argue (and some have) that in all scientific endeavors, the only task is to extract information from data. How much information can actually be extracted from a given data repository? In Shannon theory, one starts from a (possibly unknown) model for the data-generating mechanism and calculates its entropy, but in practice the starting point is only the data. Is there a general theory that provides natural model classes for the data at hand? What is the cost of learning the model, and how does it compare to the cost of actually describing the data?

Limited Resources: In many scenarios, information is limited by available resources (e.g., computing devices, living cell). How much information can be extracted and processed with limited resources?
Quantum Information: Microscopic systems do not seem to obey Shannon's postulates of information. In the quantum world and on the level of living cells, traditional information often fails to accurately describe reality [6].
Value of Information: The impact of rational and noncooperative behavior upon information as well as the value of information, should be studied in more generality.

References

[1] N. Alon and A. Orlitsky, A Lower Bound on the Expected Length of One-to-One Codes, *IEEE Trans. Information Theory*, 40, 1670-1672, 1994.

[2] H. V, von Baeyer, *Information: The New Language of Science*, Harvard University Press, 2004.

[3] A. Barron, J. Rissanen, and B. Yu, The Minimum Description Length Principle in Coding and Modeling, *IEEE Trans. Information Theory*, 44, 2743-2760, 1998.

[4] S. Bikhchandani, D. Hirshleifer, I. Welch, Learning from Others: Conformity, Fads, and Informational Cascades, *J. Econ.Perspectives*, 12 151-170, 1998.

[5] F. Brooks, Three Great Challenges for half-century-old computer science, *J. the ACM*, 50, 25-26, 2003.

[6] C. Brukner, A. Zeilinger, Conceptual Inadequacy of the Shannon Information in Quantum Measurements. *Phys. Rev.* A 63, 2001.

[7] C. Cherry, *On Human Communication*, The MIT Press, Cambridge, Massachusetts, 1978.

[8] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, Second Edition, John Wiley & Sons, New York, 2006.

[9] M. Drmota and W. Szpankowski, Precise Minimax Redundancy and Regret, *IEEE Trans. Information Theory*, 50, 2686-2707, 2004.

[10] D. Fudenberg and J. Tirole, *Game Theory*, Cambridge, MA: MIT Press, 1991.

[11] D. Galas, M. Nykter, G. Carter, N. Price and I. Shmulevich, Set-based Complexity and Biological Information, preprint 2007.

[12] M. Grossglauser and D. Tse, Mobility Increases the Capacity of ad-hoc Wireless Networks, *IEEE/ACM Trans. Networking*, 48, 477-486, 2002.

[13] P. Gupta and P.R. Kumar, Capacity of Wireless Networks, *IEEE Trans. Information Theory*, 46, 388-404, 2000.

[14] B. Hajek and A. Ephremides, Information Theory and Communication Networks: An Unconsummated Union, *IEEE Trans. Information Theory*, 44, 2416-2434, 1998.

[15] P. Jacquet, Space-time Information Propagation in Mobile ad hoc Wireless Networks, `http://ee-wcl.tamu.edu/itw2004/program/jacquet_inv.pdf/`, *ITW 2004*, San Antonio, 2004

[16] P. Jacquet and W. Szpankowski, Markov Types and Minimax Redundancy for Markov Sources *IEEE Trans. Information Theory*, 50, 1393-1402, 2004.

[17] Eric Kandel, James Schwartz, and Thomas Jessell. *Principles of Neural Science*. Appleton and Lange, 2000.

[18] G. Klir, *Uncertainty and Information: Foundations of Generalized Information Theory*, John Wiley & Sons, New York, 2006.

[19] J. Konorski, A Game-Theoretic Study of CSMA/CA under a Backoff Attack *IEEE/ACM Trans. Networking*, 14, 1167-1178, 2006.

[20] B.O. Kuppers, *Information and the Origin of Life*. The MIT Press, Cambridge, Massachusetts, 1990.

[21] D.G. Luenberger, *Information Science*, Princeton Univ. Press, 2006.

[22] D. Marinescu and M. Marinescu, Quantum Information: a Glimpse at the Strange and Intriguing Future of Information, *The Computer Journal*, 2007.

[23] J. Rissanen, A Universal Data Compression System, *IEEE Trans. Information Theory*, 29, 656–664, 1983.

[24] J. Rissanen, Complexity of Strings in the Class of Markov Sources, *IEEE Trans. Information Theory*, 30, 526–532, 1986.

[25] J. Rissanen, Universal Coding, Information, Prediction, and Estimation, *IEEE Trans. Information Theory*, 30, 629–636, 1984.

[26] J. Rissanen, Fisher Information and Stochastic Complexity, *IEEE Trans. Information Theory*, 42, 40–47, 1996.

[27] J. Rissanen, *Stochastic Complexity in Statistical Inquiry*, World Scientific, Singapore, 1998.

[28] J. Seidler, *The Science of Information*, WNT, Warszawa, 1982 (in Polish).

[29] A. Shamir, How to Share a Secret, *Communications of the ACM*, 22, 612-613, 1979.

[30] T. Siegfried, *The Bit and the Pendulum: From Quantum Computing to M Theory – The New Physics of Information*, John Wiley & Sons, New York, 2001.

[31] C. Shannon, A Mathematical Theory of Communication, *Bell System Technical Journal*, 27, 379-423 and 623-656, 1948.

[32] C. Shannon. The Lattice Theory of Information. *IEEE Transaction on Information Theory*, 1:105–107, 1953.

[33] W. Szpankowski, On Asymptotics of Certain Recurrences Arising in Universal Coding, *Problems of Information Transmission*, 34, No.2, 142-146, 1998.

[34] W. Szpankowski, A One-to-One Code and Its Anti-redundancy, *2005 International Symposium on Information Theory*, 1526-1528, Adelaide, 2005.

[35] S. Verdù, On Channel Capacity per Unit Cost, *IEEE Trans. Information Theory* 36, 1019-1030, Sep. 1990.

[36] C.F. von Weiszsäcker and E. von Weiszsäcker, Wideraufname der begrifflichen Frage: Was ist Information?, Nova Acta Leopoldina, 206, 1972.

[37] A. D. Wyner, An Upper Bound on the Entropy Series, *Inform. Control*, 20, 176-181, 1972.

[38] P. Young, *The Nature of Information*, Praeger, New York, 1987.

Genome compression using normalized maximum likelihood models for constrained Markov sources

Ioan Tabus and Gergely Korodi

Institute of Signal Processing
Tampere University of Technology
P.O Box 553, FIN-33101 Tampere, Finland
ioan.tabus@tut.fi, gkorodi@rim.com

Abstract

The paper presents exact and implementable solutions to the problem of universal coding of approximate repeats by using the normalized maximum likelihood model for the class of Markov sources of first order, incorporating constraints which are standard in the context of fast searching similarities over full genomes. A coding scheme combining universal codes for memoryless sources and for sources with memory is then presented. The results when compressing the full human genome show that the combined scheme is able to provide slight improvements over the existing state of the art. As a side result, interesting pairs of sequences may be found, which are highly similar by the new NML model for Markov sources, but have a lower similarity score when evaluated with the NML for memoryless sources.

1 Introduction

The DNA compression problem was studied thoroughly in the last two decades resulting in a wealth of contributions applying various data compression principles [3,4,6,7,9,10,12,16,18].

The most successful schemes for lossless genome compression are exploiting the existence of an important number of approximate repetitions in DNA sequences. There are also other sources of redundancy in DNA and for this reason most DNA compression methods integrate several algorithms, including always a clear coder (representation of the four bases, A,C,G,T using 2 bits/base) and a symbol-wise adaptive coder of low order (e.g., one or two). However, when checking the contribution of each of the encoders in the ensemble, the algorithm for compressing the approximate repeats is the most important and difficult to design. We recently have developed DNA compression schemes where the compressor for approximate repeats was based on a NML universal model for a class of memoryless models in [9],[18] and later for a class of Markov models of first order in [10], to obtain the state of the art for the human genome compression. Section 2 reviews the overall encoding scheme which combines several encoders, each targeting a typical form of redundancy found in DNA sequences.

In this paper we continue the study of the NML model for Markov sources, introducing further refinements. In Section 3 we discuss the NML model for markovian sources of arbitrary orders. Subsequently in Section 4 we only concentrate on first order Markov models and introduce the additional constraint that all considered approximate matches have at least a given number of consecutive exactly matching positions. This constraint enables a fast search time, and is implicitly considered in many modern similarity searching tools. Since the size of genomes is large, typically several giga-bases long, exhaustive search for similar sequence pairs is impossible and the compression results will be highly affected by the compromise between speed of search and sensitivity of finding the most relevant matches.

In order to allow a fast search, all algorithms for finding similar sequences which are in daily use in bioinformatics implement a seed based search, where a number of contiguous matches is assumed (or more recently an arbitrary but fixed pattern of matching is assumed [11]). Marking in advance the blocks having a fixed number q of contiguous matches is very fast (q is relatively small, e.g., $q = 11$), and then, starting from these seeds, longer approximate matches are formed and evaluated. We include this contiguous matching constraint in the definition of our model class and then proceed with developing the NML model for this class.

2 Overview of a generic DNA compression scheme combining several encoders

The compression scheme operates blockwise along the DNA sequence, parsing the sequence into non-overlapping blocks of equal size. The current block to be encoded will be compressed with the winner of several competing methods: #1) a symbolwise adaptive Markov model of first order (not to be confused with encoder #3, the latter using the NML model for Markovian sources for encoding a "matching pattern"); #2 the clear representation using two bits per base; and #3) coding by reference to the best approximate matching block in the past, using the NML for encoding the matching pattern of the current block relative to one of the previously encoded blocks. For each competing method the required codelength is evaluated or quickly fetched from the lookup tables, which store the codelength corresponding to the current sufficient statistics of the block, and only the winner method is subsequently used in the more elaborate process of building the bitstream, using arithmetic coding in conjunction with the implementable coding distributions.

Actually combining the three coders and selecting the candidate blocksizes is an intricate problem and earlier we derived suitable solutions with moderate complexity while achieving excellent overall compression [9], in the spirit of MDL principle. The selections involved at the overall scheme level can be naturally tackled by using the adaptively tracked statistics of the usage of each coder and its individual parameters. We defer to [9] for details of this architecture and continue here with a formal description of the more difficult and important module, that for encoding based on reference to a past block. To not clutter with details, we are omitting the separate description of direct-palindrome matching, the palindromes being obviously treated by reversing the string forming the current block and changing its bases $A \leftrightarrow T$ and $C \leftrightarrow G$, as it is well established with all existing DNA compressors.

We consider the DNA sequence to be encoded $y^n = y_1, y_2, \ldots, y_n$, where $y_i \in \{A, C, G, T\}$ and a sequence of the same length $z^n = z_1, z_2, \ldots, z_n$ which will be used as a regressor sequence, located at an arbitrary position (to be determined) in the past. By using a regressor sequence one hopes to better encode y^n, if the two sequences y^n and z^n possess a high degree of similarity, or equivalently a low distance, in a sense to be defined. The natural, but quite simplistic, Hamming distance $d_H(z^n, y^n)$ counts the number of mismatches between the corresponding symbols z_i and y_i, $i = 1, \ldots, n$, and one can define a match sequence $x^n = x_1, \ldots, x_n$ with $x_i = 0$ if $z_i = y_i$ and $x_i = 1$ if $z_i \neq y_i$. One can generalize the Hamming distance in several ways, by defining many of the meaningful sequence distance measures, where one accounts for insertions, deletions, or instead of binarized matching decisions considers matrices of substitution probabilities, resulting in a wealth of possible distances, which sometimes after a proper normalization are referred to in bioinformatics as "odd-scores". Many of the biologically defined distances or similarity measures were shown to be useful in revealing interesting biological facts, e.g., PAM scores tuned for evaluating evolutionary distances between different species, but were not found yet to be useful in obtaining a good compression/description of full genomes.

Here we pursue however a principled way to account for the possible Markovian dependence of the matching sequence x^n, which also induces a novel similarity score. In order to recover y^n the decoder needs to receive the following: a pointer to the selected optimal regressor z^n (transmitted in a predictive way described in detail in [9]); the matching sequence

x^n, which is transmitted using a universal code; and the mismatching symbols. In the case of a mismatch, when $x_j = 1$, the symbol y_j is encoded by using arithmetic coding for the distribution $P(y|z_j, x_j = 1)$, and based on experimental evidence we chose a uniform distribution for the three possible symbols, i.e., $P(y|z_j, x_j = 1) = 1/3$, where y can be uniformly any of the symbols A, C, G, T, except the symbol z_j, which is excluded since the decoder already knows that $x_j = 1$ is for a mismatch. Our universal model of $P(y^n)$ will thus be based on a universal model for $P(x^n)$, where by the assumed independence $P(y^n) = P(x^n)/3^m$ with m the number of mismatches between y^n and z^n, i.e., $m = d_H(z^n, y^n)$ is the number of ones in the sequence x^n. The universal coding of the sequence x^n will be achieved by using the NML model for various classes of Markov models, as described next.

3 NML model for Markov sources of order k

We assume that the sequence x^n was generated by one model in the class of Markov sources of order k, $\mathcal{M}_k = \{P(x_t|x_{t-1}, \ldots, x_{t-k}) : x_{t-i} \in \mathcal{A}\}$. For notation simplicity we denote the states in the form $j = [x_{t-k} \ldots x_{t-1}]$ and since here the alphabet $\mathcal{A}$ is binary we will identify j with the integer having the binary representation $[x_{t-k} \ldots x_{t-1}]$. The parameters of the model will be denoted $\theta_{ji} = P(x_t|x_{t-1}, \ldots, x_{t-k})$, where $i = x_t \in \mathcal{A}, j = [x_{t-k} \ldots x_{t-1}] \in \mathcal{A}^k$.

In the string $x^n = x_1 \ldots x_n$ we observe the catenation $ji \in \mathcal{A}^{k+1}$ a number of n_{ji} times, with $j \in \mathcal{A}^k, i \in \mathcal{A}$. We denote $n_{j\cdot} = \sum_{i \in \mathcal{A}} n_{ji}$ and $n_{\cdot j} = \sum_{i \in \mathcal{A}} n_{ij}$. The set of all counts is organized as a $m^k \times m$ matrix $\boldsymbol{n}$ having n_{ji} as the (j, i)'th element. The necessary constraints on the counts obtained from any sequence which starts in state $r \in \mathcal{A}^k$ and ends in state $s \in \mathcal{A}^k$ can be seen to be

$$n_{j\cdot} - n_{\cdot j} = \delta_{rj} - \delta_{sj}, \quad \forall j \in \mathcal{A}^k; \tag{1}$$

$$\sum_{j \in \mathcal{A}^k} n_{j\cdot} = \sum_{j \in \mathcal{A}^k} n_{\cdot j} = n - k, \tag{2}$$

where $\delta_{rj} = 1$ if $r = j$ and $\delta_{rj} = 0$ if $r \neq j$.

If one knows the starting state r and all counts $\boldsymbol{n}$ he can find easily the final state s as the value of j for which $\delta_{rj} - n_{j\cdot} + n_{\cdot j}$ is one. Although the final state s is fully determined by the initial state r and the set of counts $\boldsymbol{n}$ we prefer to specify it explicitly in the triplet $(r, s, \boldsymbol{n})$ together with the initial state and the sufficient statistics to characterize any string x^n. Maximum likelihood of the parameters of the Markov chain is a function solely of the matrix of counts $\boldsymbol{n}$ [1]

$$\hat{\theta}_{ji} = \frac{n_{ji}}{n_{j\cdot}}. \tag{3}$$

The number of strings starting from the initial state r, ending in the final state s, and having the matrix of counts $\boldsymbol{n}$ is given by [19]

$$N(r, s, \boldsymbol{n}) = T(r, s, \boldsymbol{n}) \frac{\prod_{j \in \mathcal{A}^k} n_{j\cdot}!}{\prod_{i \in \mathcal{A}} \prod_{j \in \mathcal{A}^k} n_{ji}!}, \tag{4}$$

where $T(r, s, \boldsymbol{n})$ is the sr-th cofactor of the $m^k \times m^k$ matrix M defined as follows: initialize M with all zero elements and then for each pair $(j, i) \in \mathcal{A}^k \times \mathcal{A}$ define $q = j_2 \ldots j_k i \in \mathcal{A}^k$ and set $M(j, q) = \delta_{jq} - \frac{n_{ji}}{n_{j\cdot}}$.

When some of the states $j \in \mathcal{A}^k$ are not observed in the string x^n, the formula (4) has to be applied taking into account only a process containing the observed states. The set of valid triplets $(r, s, \boldsymbol{n})$, i.e., those observed over all sequences x^n when starting from state r, is denoted $\Omega_{r,n}$.

Therefore the NML model can be written

$$\hat{P}(x^n|r(x^n) = r) = \frac{\boldsymbol{n}(x^n)^{\boldsymbol{n}(x^n)}}{\sum_{(r,s,\boldsymbol{n}) \in \Omega_{r,n}} N(r, s, \boldsymbol{n}) \boldsymbol{n}^{\boldsymbol{n}}} \tag{5}$$

where $\boldsymbol{n}^{\boldsymbol{n}}$ is a shorthand notation for $\prod_{i\in\mathcal{A},j\in\mathcal{A}^m}\left(\frac{n_{ji}}{n_{j\cdot}}\right)^{n_{ji}}$. We note that the normalization constant in the denominator of (5) was evaluated in an asymptotic manner in a number of previous works: e.g., [5][8][14][17], but our concern is on exact evaluations, as required in an implementable coder.

Computing the set $\Omega_{r,n}$ and all cardinalities of the Markov types $N(r,s,\boldsymbol{n})$ can be done by directly checking if $(r,s,\boldsymbol{n})$ fulfills the conditions (1), (2), and supplementarily, checking if $T(r,s,\boldsymbol{n})$ is nonzero. A more appealing alternative is to build $\Omega_{r,n}$ and compute the cardinalities $N(r,s,\boldsymbol{n})$ recursively in n. For that we will take into account for each $(r,s,\boldsymbol{n})$ the possible extensions to a triplet $(r,s',\boldsymbol{n}')\in\Omega_{r,n+1}$ which are obtained when appending the symbol $i\in\mathcal{A}$ to any x^n having sufficient statistics $(r,s,\boldsymbol{n})$. The transformation will leave most of the elements of $\boldsymbol{n}$ unchanged, the only updating operations needed are easily seen to be

$$s' = s_2\dots s_k i \tag{6}$$

$$n'_{si} = n_{si}+1 \tag{7}$$

where the final state was written as a sequence of binary symbols, $s=s_1\dots s_k$. The cardinality $N(r,s',\boldsymbol{n}')$ will result immediately by adding all cardinalities $N(r,s,\boldsymbol{n})$ of those $(r,s,\boldsymbol{n})$ that are transformed to $(r,s',\boldsymbol{n}')$. The initial conditions for the recursions can be easily set at $n=k+1$. As an example, if $|\mathcal{A}|=2,k=2$, the 4×2 matrix of counts is $\boldsymbol{n}=[n_{000},n_{000};n_{010},n_{011};n_{100},n_{101};n_{110},n_{111}]$ and for all eight distinct binary strings x^3 we get a different matrix of counts, each such matrices having just one nonzero element, e.g. one of the eight nonzero cardinalities $N(r,s,\boldsymbol{n})$ is $N(01,10,[0,0;1,0;0,0;0,0])=1$.

We have constructed the set $\Omega_{r,n}$ and computed all cardinalities of the Markov types $N(r,s,\boldsymbol{n})$ in our experiments with $m=2,k=2$, in two different ways and checked the identity of results. First we propagated recursively in n the elements of the set $\Omega_{r,n}$ and their associated cardinalities $N(r,s',\boldsymbol{n}')$, by the transformations $(r,s,\boldsymbol{n})\to(r,s',\boldsymbol{n}')\in\Omega_{r,n+1}$ as shown in (6),(7). We also implemented an alternative method where out of all possible $(r,s,\boldsymbol{n})$ only those satisfying the three conditions: (1), (2), and $T(r,s,\boldsymbol{n})\neq 0$, are included in the set $\Omega_{r,n}$, and then (4) is used for the computation of cardinalities $N(r,s',\boldsymbol{n}')$. Both methods resulted in identical sets $\Omega_{r,n}$ and provided identical cardinalities $N(r,s,\boldsymbol{n})$ for all values of n up to the maximum tested one, $n=36$, in the case of $m=2,k=2$. The exact computation of the cardinalities $N(r,s,\boldsymbol{n})$ was performed in a practical time up to values of $n=36$, with the recursive method being several times faster. For the considered DNA application, the length of the block was taken $n=32$, for easy comparisons with the previously reported results in [10], where $m=2,k=1$, and the length $n=32$ was found to be optimal, out of tested values up to $n=128$.

4 Exact computation of the NML model for Markov sources of first order

4.1 NML model for unconstrained Markov sources of first order

In the following we concentrate on the description of a particular Markov model, which proves to be very efficient for the relatively short sequences ($n=32$) used in the block based encoding of full genomes. Let us define the parameters in the Markov model $\theta_{\ell j}=P(x_i=j|x_{i-1}=\ell)$ with $j,\ell\in\{0,1\}$ and group all parameters in the vector $\theta=(\theta_{00},\theta_{10},\theta_{01},\theta_{11})$. From the definition of the conditional probabilities, we require

$$\theta_{00}+\theta_{01}=1;\quad \theta_{10}+\theta_{11}=1 \tag{8}$$

We note that in our block-wise encoding for all but the first block the decoder already knows the value of $x_0=\delta_{y_0,z_0}$, from the preceding block matching performed. The probability of

the sequence x^n will thus be

$$P(x^n|x_0,\theta) = \prod_{i=1}^{n} P(x_i|x_{i-1}) = \theta_{00}^{n_{00}}\theta_{10}^{n_{10}}\theta_{01}^{n_{01}}\theta_{11}^{n_{11}} \tag{9}$$

where $n_{\ell k}$ is the number of times the symbol $x_i = k$ was observed after symbol $x_{i-1} = \ell$ for $i = 1,\ldots,n$ and $k,\ell \in \{0,1\}$. Letting n_k be the number of times $x_i = k$ for $i = 1,\ldots,n$ and $k \in \{0,1\}$ we have the simple connecting relationships $n = n_0 + n_1$,

$$\begin{aligned} n_0 &= n_{\cdot 0} = n_{00} + n_{10} \\ n_1 &= n_{\cdot 1} = n_{01} + n_{11}. \end{aligned} \tag{10}$$

Maximizing the log-probability

$$\log P(x^n|x_0,\theta) = n_{00}\log\theta_{00} + n_{01}\log(1-\theta_{00}) + n_{10}\log(1-\theta_{11}) + n_{11}\log\theta_{11}, \tag{11}$$

subject to the constraints (8) leads to the ML parameters

$$\hat\theta_{00} = n_{00}/(n_{00}+n_{01}) \tag{12}$$

$$\hat\theta_{01} = n_{01}/(n_{00}+n_{01}) \tag{13}$$

$$\hat\theta_{10} = n_{10}/(n_{10}+n_{11}) \tag{14}$$

$$\hat\theta_{11} = n_{11}/(n_{10}+n_{11}). \tag{15}$$

We observe in passing that n_{10} might differ of n_{01} by at most 1, and consequently $n_0 = n_{\cdot 0} = n_{00} + n_{10}$ may be different of $n_{0\cdot} = n_{00} + n_{01}$. This situation in general can be avoided by taking the counts in the string x^n in a circular manner, but we want to keep the exact evaluations, as required by the assumed model (9). The matrix containing the sufficient statistics $\boldsymbol{n} = [n_{00}, n_{01}; n_{10}, n_{11}]$ uniquely determines $\hat\theta$. Now we continue in the usual way to normalize the maximized likelihood [14]

$$P(x^n|x_0,\hat\theta) = \hat\theta_{00}^{n_{00}}\hat\theta_{10}^{n_{10}}\hat\theta_{01}^{n_{01}}\hat\theta_{11}^{n_{11}} \tag{16}$$

in order to get the corresponding NML model

$$P(x^n|x_0) = \frac{P(x^n|x_0,\hat\theta)}{\sum_{x^n\in\{0,1\}^n} P(x^n|x_0,\hat\theta)} \tag{17}$$

$$= \frac{\hat\theta_{00}^{n_{00}}\hat\theta_{10}^{n_{10}}\hat\theta_{01}^{n_{01}}\hat\theta_{11}^{n_{11}}}{\sum_{(x_0,x_n,\boldsymbol{n})\in\Omega_{x_0,n}} N(x_0,x_n,\boldsymbol{n})\hat\theta_{00}^{n_{00}}\hat\theta_{10}^{n_{10}}\hat\theta_{01}^{n_{01}}\hat\theta_{11}^{n_{11}}} \tag{18}$$

where we denote by $N(x_0,x_n,\boldsymbol{n})$ the number of strings having the same sufficient statistics vector $\boldsymbol{n}$, and by $\Omega_{x_0,n}$ the set of all possible sufficient statistics vectors. As before, x_n, which corresponds to the final state s, can be determined from x_0 and $\boldsymbol{n}$, but in line with Section 3, we have specified it for clarity.

We present the evaluation of $N(x_0,x_n,\boldsymbol{n})$ by resorting to a simple decomposition of the original string into two strings, which will also be useful later, for the simple implementation of the coding.

We split the string x^n into two strings $c^{n_0} = c_1\ldots c_{n_0}$ and $d^{n_1} = d_1\ldots d_{n_1}$. The string c^{n_0} lists in order all symbols situated before a 0 in the string x^n and the string d^{n_1} lists the symbols situated before a 1 in x^n. In the string c^{n_0} we have n_{00} zeros and n_{10} ones, while in the string d^{n_1} we have n_{01} zeros and n_{11} ones.

Once we know n_{01}, n_{10} we can find by $x_n = n_{01} - n_{10} + x_0$ the last bit in the string. Knowing the last bit in the string we can go backwards and find $x_{n-1}, x_{n-2},\ldots$ by checking the sequences c^{n_0} and d^{n_1}.

As a simple example take $x_0 = 0$ and $x^5 = 11001$, for which $n_0 = 2, n_1 = 3, n_{00} = 1, n_{01} = 2, n_{10} = 1, n_{11} = 1$, while the related strings are $c^2 = 10$ and $d^3 = 010$. We show now how to reconstruct the string x^5 by knowing $n = 5, n_0 = 2, n_{00} = 1, n_{11} = 1$ and $c^2 = 10$ and $d^3 = 010$. We can find at once $n_{10} = n_0 - n_{00} = 1$ and $n_{01} = n_1 - n_{11} = 2$. Then we get $x_n = n_{01} - n_{10} + x_0 = 2 - 1 + 0 = 1$. The symbol preceding $x_n = 1$ can be found as the last symbol in the sequence d^3, and thus $x_{n-1} = 0$. The symbol preceding $x_{n-1} = 0$ can be found as the last symbol in c^2 thus we have $x_{n-2} = 0$. We can continue the same way till we complete finding all symbols of x^5.

As a final issue, one of the sequences c^{n_0} and d^{n_1} is redundant. Take the case $x_0 = 0$. We know that in d^{n_1} the first symbol is a 0 (because to get to the first 1 in the sequence we need to start from a 0). In a similar way, when $x_0 = 1$, the first symbol in c^{n_0} is necessarily a 1. Thus we need to store only the last $n_0 - 1$ symbols from the sequence c^{n_0}, when $x_0 = 1$, while when $x_0 = 0$ we need to store only the last $n_1 - 1$ symbols of the sequence d^{n_1}. Let us denote e^{m_0} the non-redundant sequence which is identical to c^{n_0} when $x_0 = 0$ while for $x_0 = 1$ we have $e^{m_0} = c_2, \ldots, c_{n_0}$. Similarly we denote by f^{m_1} the sequence d^{n_1} when $x_0 = 1$ and the sequence $d_2, \ldots, d_{n_1}$ when $x_0 = 0$. Other redundancies do not exist, and we can find the number of sequences starting with x_0 and having the same sufficient statistics $\boldsymbol{n} = (n, n_0, n_{00}, n_{11})$ as

$$N(x_0, x_n, \boldsymbol{n}) \quad = \quad \begin{cases} \binom{n_0}{n_{00}} \binom{n_1 - 1}{n_{11}} & if \quad x_0 = 0 \\ \binom{n_0 - 1}{n_{00}} \binom{n_1}{n_{11}} & if \quad x_0 = 1. \end{cases} \tag{19}$$

This number can be seen to be identical to the number found in [2]. Moreover, our decomposition into the strings e^{m_0} and f^{m_1} is also helpful for solving the case when the string x^n is constrained to have q contiguous zeros.

4.2 Constraining the number of contiguous zeros in the string x^n to at least q

Let us denote $\alpha(n, m, q)$ the number of strings with n symbols having a total number of m zeros from which at least q are contiguous. This number can be computed recursively, by the following recursion [9]:

$$\alpha(n, m, q) = \begin{cases} 0, & \text{if } (n < m) \vee (m < q) \\ 1, & \text{if } (n = m) \wedge (n \geq q) \\ \alpha(n-1, m, q) + \alpha(n-1, m-1, q) & \\ \quad + \binom{n-q-1}{m-q} - \alpha(n-q-1, m-q, q), & \text{else.} \end{cases} \tag{20}$$

Now consider an arbitrary string x^n having the sufficient statistics $\boldsymbol{n}$ and its decomposition in two strings c^{n_0} and d^{n_1}. A contiguous run of q zeros in x^n will be translated into a run of $q - 1$ zeros in c^{n_0}. So the number of all strings with sufficient statistics $\boldsymbol{n}$ starting from x_0 is

$$N(x_0, x_n, \boldsymbol{n}) \quad = \quad \begin{cases} \alpha(n_0, n_{00}, q-1) \binom{n_1 - 1}{n_{11}} & if \quad x_0 = 0 \\ \alpha(n_0 - 1, n_{00}, q-1) \binom{n_1}{n_{11}} & if \quad x_0 = 1. \end{cases} \tag{21}$$

4.3 A DNA encoder based on the order-1 NML model

In this section we describe a practical implementation of the order-1 NML algorithm, and its application for DNA sequences. This algorithm uses the model (18) and the constraint on the string counts of same sufficient statistics according to (21). For greater adaptivity, the algorithm also uses several other models, namely order-0 NML and order-1 context coding,

as described in [9] and [10]. In brief, the algorithm decomposes the input file into a sequence of fixed size macroblocks. Each macroblock is tested independently by various NML orders (0 or 1) and block sizes (such as 24, 32, 48 and 96). Among these, the order and block size that yield the shortest encoded macroblock length are signaled in the output, and the corresponding model is used to compress that macroblock. This way the NML order and block size are fixed in each macroblock, but may change in the next one. If no satisfactory regressor has been found, the algorithm reverts to encoding the current block by an order-1 context coder. To make the process even more adaptive, each context coder has a parameter that specifies how frequently the coder should downscale its model [9], making the impact of older statistics less significant. Context coders of different parameters are then combined with the various NML models. Finally, we also add clear encoding as an alternative to prevent significant size expansion.

4.3.1 Initialization

The algorithm requires the initialization of some values prior to the encoding or decoding process. These values need to be stored for fast and frequent access later. First we have to compute and store the numbers $N(x_0, x_n, \boldsymbol{n}) = N(x_0, x_n, \boldsymbol{n})$ for all the possible combinations of x_0 and $\boldsymbol{n}$ (we will use from this point on for simplicity the notation $N(x_0, x_n, \boldsymbol{n})$, since x_n is uniquely determined by x_0 and $\boldsymbol{n}$). The formulae (20) and (21) provide a very fast means of generating these values, so this computation is carried out at run-time, and the results are stored only in memory.

The other necessary values are the probabilities of the strings with same sufficient statistics. Let us introduce the symbol C for the denominator in (18), that is,

$$C = \sum_{(x_0,\boldsymbol{n})\in\Omega_{x_0,n}} N(x_0, x_n, \boldsymbol{n})\hat{\theta}_{00}^{n_{00}}\hat{\theta}_{10}^{n_{10}}\hat{\theta}_{01}^{n_{01}}\hat{\theta}_{11}^{n_{11}}. \tag{22}$$

Then by (18) we can write the probability of the sufficient statistics as

$$P(\boldsymbol{n}|x_0) = \frac{N(x_0, x_n, \boldsymbol{n})\hat{\theta}_{00}^{n_{00}}\hat{\theta}_{10}^{n_{10}}\hat{\theta}_{01}^{n_{01}}\hat{\theta}_{11}^{n_{11}}}{C}. \tag{23}$$

This formula gives a straightforward method for computing the values $P(\boldsymbol{n}|x_0)$, but this computation is time-consuming, hence it is best done just once at compilation phase, and the results are stored on disk for run-time access. Furthermore, since (23) involves floating-point calculations, arithmetic precision errors may occur, but their effect is insignificant so long as the probabilities sum up to 1. In practice, the values $P(\boldsymbol{n}|x_0)$ are converted to integers after multiplying by a large number M, and then the sum of these integers (which may differ from M due to rounding errors) is used as a normalization factor to get back to the probabilities. The value M should be selected satisfying two constraints: first, $M \cdot P(\boldsymbol{n}|x_0) \geq 1$ for all sufficient statistics, and second, that M should be within the range of Arithmetic Coding [13] which is used to encode the corresponding probabilities.

As an illustration of the model costs, in Arithmetic Coding our test implementation uses 64-bit integer operations, which enables an interval resolution of 2^{29}. For the block size of $n = 32$ and the contiguous seed length $q = 11$ we have $\min\{P(\boldsymbol{n}|x_0)\} = 2.7984592 \cdot 10^{-5}$, and we set $M = 10^8$. The pre-computed binary model for (23) takes only 6104 bytes.

4.3.2 Search for the best regressor

For the similarity metric used for the comparison between candidate regressors, we have to take into account that non-matching symbols must be corrected, and this increases the overall cost. Hence for DNA sequences with alphabet $S = \{A, C, G, T\}$ our probability model becomes

$$P(y^n|x_0) = \frac{1}{3^{n_1}}P(x^n|x_0) \tag{24}$$

$$= \frac{1}{3^{n_1}} \frac{1}{N(x_0, x_n, \boldsymbol{n})} P(\boldsymbol{n}|x_0). \tag{25}$$

$$\tag{26}$$

Therefore the cost of encoding a block with a known regressor z^n is

$$-\log P(y^n|x_0; z^n) = -\log P(\boldsymbol{n}|x_0) + \log N(x_0, x_n, \boldsymbol{n}) + n_1 \log 3. \tag{27}$$

Equation (27) provides an efficient and easily computable similarity metric for any regressor and the current block y^n. The candidate regressors are selected from a dictionary assigned to y^n: $D_{y^n} \subseteq S^n$, which is constructed from the contents of the current sliding window, including both normal and complemented palindrome segments, but subject to the constraint that a sufficiently long contiguous run of matching symbols exists between each element of D_{y^n} and y^n [9]. In the search phase the current sufficient statistics is determined by a symbol-by-symbol comparison between $w^n \in D_{y^n}$ and y^n, after which the cost function is given by (27) using the tabulated values of $N(x_0, x_n, \boldsymbol{n})$ and $P(\boldsymbol{n}|x_0)$. The result of the search is going to be the block which minimizes (27), that is,

$$z^n = \arg \min_{w^n \in D_{y^n}} -\log P(y^n|x_0; w^n). \tag{28}$$

4.3.3 Encoding algorithm for the matching mask x^n and segment y^n

Following Equation (27) the coding of a block is done by the following steps:

1. Obtain x_0 from the previous position, based on the current regressor.
2. Encode the regressor location in $\log W$ bits (W is the current window size), and the match type in 1 bit (normal or palindrome).
3. Compute the sufficient statistics vector $\boldsymbol{n} = (n_{00}, n_{10}, n_{01}, n_{11})$ and encode it by using the tabulated probabilities $P(\boldsymbol{n}|x_0)$ from (23).
4. Encode the string conditional on the sufficient statistics

$$P(x^n|\boldsymbol{n}, x_0) = \frac{1}{N(x_0, x_n, \boldsymbol{n})}. \tag{29}$$

5. Correct the non-matching bases in $\log_2 3$ bits each for all $x_i = 1$.

To overcome the computational complexities raised by Step 4 (for $n = 32$ the values $N(x_0, x_n, \boldsymbol{n})$ are typically much larger than the practical Arithmetic Coding limit 2^{29}), this step is split into two tasks: decompose the string x^n into c^{n_0} and d^{n_1}, then first encode by

$$P(c^{n_0}|\boldsymbol{n}, x_0) = \begin{cases} \frac{1}{\alpha(n_0, n_{00}, q-1)} & if \quad x_0 = 0 \\ \frac{1}{\alpha(n_0-1, n_{00}, q-1)} & if \quad x_0 = 1 \end{cases} \tag{30}$$

then by

$$P(d^{n_1}|\boldsymbol{n}, x_0) = \begin{cases} \frac{1}{\binom{n_1 - 1}{n_{11}}} & if \quad x_0 = 0 \\ \frac{1}{\binom{n_1}{n_{11}}} & if \quad x_0 = 1 \end{cases} \tag{31}$$

Encoding by (31) can be done in a straightforward manner [18], iterating through the positions and counting the numbers of 0's and 1's seen so far, from which we form our probability estimation of the next symbol, and use that for encoding. The procedure of encoding by (30) is achieved by another iterative algorithm first described in [9].

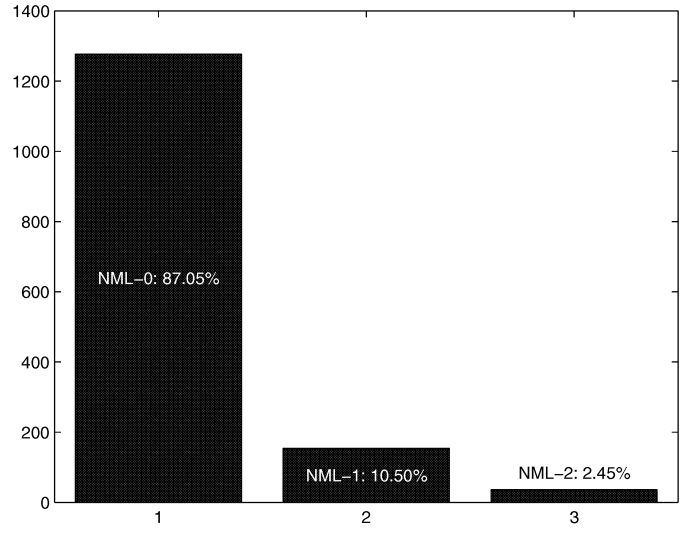


Figure 1: Comparison of NML orders 0, 1 and 2 with block size $n = 32$ for the human growth hormone HUMGHCSA. The bars count the number of blocks for which the corresponding order resulted in strictly smaller code length, than the other two.

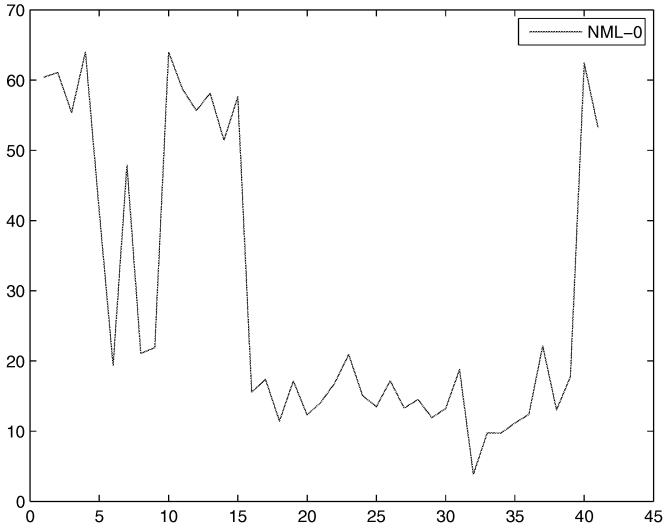
5 Discussion and results

In this section we give for the algorithms presented earlier an empirical evaluation on DNA sequences.

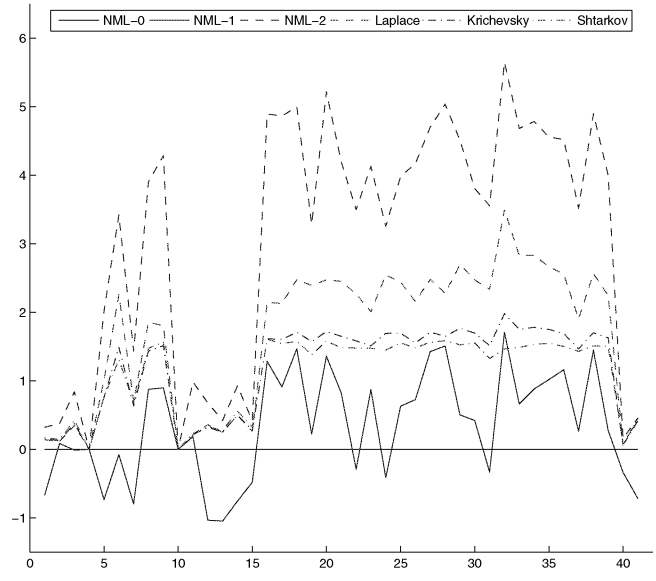
5.1 Higher order NML models

The rationale behind using higher order NML models for DNA sequences stems from the fact that neighboring symbols in the matching bitmask usually have some correlation between them, and this correlation cannot be detected by the order-0 model. Perhaps the most typical case is a partial match inside a block, in which case spontaneously alternating sequence of 0's and 1's suddenly turn into a run of 0's, or vice versa, indicating the presence of a significant exact match. To exploit this redundancy in the best way, one would attempt to test models with orders higher than one. However, several factors impede the efficient use of higher orders. On the practical side, the cardinality of the set of possible sufficient statistics is growing exponentially with the order, which makes difficult both the computation of the normalization constant and the implementation of coding with the NML model. Another factor is that based on tests we have carried out, many relevant matching bit masks for DNA sequences are best processed by order-0 NML, whereas higher order models tend to assign low costs for matches that rarely or never occur in DNA, inflating the costs for the rest. In practice we have found that while a careful combination of order-1 with order-0 NML performs slightly better than order-0 alone, adding order-2 NML (and probably any higher orders) never results in noticeable improvements. Figure 1 illustrates this fact on the human growth hormone HUMGHCSA, showing the number of blocks and their percentages when each order (0, 1 or 2) performed better than the other two.

Figure 2 shows the compression efficiency along the sequence HUMGHCSA of NML for



(a)



(b)

Figure 2: (a) The performance of NML-0 on the sequence HUMGHCSA for block size 32, averaged over 50 blocks. (b) The average codelength of NML with orders 1 and 2, as well as SNML variants, minus the average codelength of NML order-0 on HUMGHCSA.

Markov models of various orders. For better visibility, Figure 2 (a) shows the performance of NML order-0 alone, then Figure 2 (b) gives the difference of other models compared to order-0. We note that NML order-1 is the only model that occasionally improves over order-0 on this file.

5.2 Sequential NML models

Recently, Rissanen and Roos [15] introduced the conditional NML or sequential NML (SNML), which is a sequential universal model intended also for sequences generated by Markov sources. An important practical advantage of SNML over NML is that it can be used for encoding arbitrarily long strings due to its recursive application, while for the block based NML model described in this paper the computation of $\sum_{(r,s,\boldsymbol{n})\in\Omega_{r,n}} N(r,s,\boldsymbol{n})\boldsymbol{n}^{\boldsymbol{n}}$ and handling of its partial terms encounter practical difficulties due to memory storage requirements and effects of finite precision computations, when the size of n is large.

The Sequential NML model has three variants, two of which, for the class of Bernoulli models, coincide with the Laplace and Shtarkov estimators, respectively. These estimators are defined by

$$P_{\mathrm{Lap}}(0|x^n) = \frac{n_0+1}{n+2} \tag{32}$$

and

$$P_{\mathrm{Sht}}(0|x^n) = \frac{(n_0+1)e(n_0)}{(n_0+1)e(n_0)+(n_1+1)e(n_1)}, \tag{33}$$

where $e(n) = (1+1/n)^n$. In our application for DNA compression, we can use these estimators to encode the matching bit masks for each block, then to correct the non-matching bases as before, resulting in a fast and simple algorithm.

For the search of the best regressor we point out that (32) assigns a code length that is a function of only n_0, so we can use the Hamming distance as the similarity metric, like in the case of NML order-0. The estimator (33) depends on the actual bit pattern x^n, however, an exhaustive search carried out on all sequences with $n \leq 32$ revealed that for these short blocks the regressor minimizing the cost of (33) always minimizes the Hamming distance as well, so again our comparison routine is based on the match score with no loss in performance.

Figure 2 (b) shows the performance of these estimators compared to NML order-0, on the sequence HUMGHCSA. We found the results obtained on other sequences to be mostly consistent to this one. The improvement of NML order-0 over SNML indicates that the probabilities of the sufficient statistics of NML order-0,

$$P(n_0) = \frac{\binom{n}{n_0}\left(\frac{n_0}{n}\right)^{n_0}\left(\frac{n-n_0}{n}\right)^{n-n_0}}{\sum_{m\geq N}\binom{n}{m}\left(\frac{m}{n}\right)^{m}\left(\frac{n-m}{n}\right)^{n-m}}, \tag{34}$$

approximate the empirical probabilities gained from the sequences better than those used implicitly by the Laplace and Shtarkov estimators.

5.3 Compressing the human genome

We illustrate the compression performance of the order-1 NML model (18) using the constraint (21), when integrated in the GeNML DNA compression framework [9], on the April 14, 2003 release of the human genome [20]. Table 1 shows the encoded rate for all human chromosomes in two flavors: in the left panel the whole sequences were compressed, which also include a substantial number of non-specified bases (denoted by "N" in the FASTA format). In the right panel the non-specified bases are omitted, leaving only the nucleobases A, C, G and T. This second test is important due to the fact that the N symbols usually come in very long runs, and thus their presence is not indicative to the common type of redundancies found in DNA sequences. The "NML-0 + constr. NML-1" column shows the compression

rate of the program using the constrained order-1 model described in this article, and for comparison we include in the " NML-0 + NML-1" column the results of an earlier version [10], which did not account for the constraint on the contiguous seeds and used a different policy on encoding the first bit of the matching bit mask. It can be observed that the new method either improves or has the same average performance over each of the chromosomes, but these average improvements are only marginal. However, local improvements can be quite important, and they may signal better biological similarities.

6 Acknowledgements

We wish to thank Jorma Rissanen for generously sharing his ideas and for his enthusiasm in disseminating the NML model, which led to our fruitful joint work in the project on DNA compression.

References

[1] M.S. Bartlett, "The frequency goodness of fit test for probability chains", *Proceedings Cambridge Philosophical Society,* vol. 47, pp. 86–95, 1950.

[2] T.C. Bell, J.G. Cleary, and I.H. Witten, *Text Compression,* Prentice Hall, Englewood Cliffs, NJ, 1990.

[3] X. Chen, S. Kwong, and M. Li, "A Compression Algorithm for DNA Sequences", *IEEE Engineering in Medicine and Biology,* pp. 61–66, July/August 2001.

[4] X. Chen, M. Li, B. Ma, and J. Tromp, "DNACompress: fast and effective DNA sequence compression", *Bioinformatics,* vol. 18, pp. 1696–1698, 2002.

[5] L. Davisson, "Minimax noiseless universal coding for Markov sources", *IEEE Transactions on Information Theory*, vol. 29, no. 2, pp. 211 - 215, 1983.

[6] S. Grumbach and F. Tahi, "Compression of DNA sequences", *Data Compression Conference 1993,* DCC '93, pp. 340–350, 1993.

[7] S. Grumbach and F. Tahi, "A new challenge for compression algorithms: Genetic sequences", *J. Inform. Process. Manage.,* vol. 30, no. 6, pp. 875–886, 1994.

[8] P. Jacquet and W. Szpankowski, "Markov types and minimax redundancy for Markov sources", *IEEE Transactions on Information Theory*, vol. 50, no. 7, pp. 1393–1402, 2004.

[9] G. Korodi and I. Tabus, "An efficient normalized maximum likelihood algorithm for DNA sequence compression", *ACM Transactions on Information Systems,* vol. 23, no. 1, pp. 3–34, 2005.

[10] G. Korodi, I. Tabus, "Normalized maximum likelihood model of order-1 for the compression of DNA sequences", in Proc. IEEE Data Compression Conference, DCC'07, pp:33 - 42, Snowbird, 27-29 March 2007.

[11] B. Ma, J. Tromp, and M. Li, "PatternHunter: Faster and More Sensitive Homology Search", *Bioinformatics,* vol. 18, pp. 440–445, 2002.

[12] T. Matsumoto, K. Sadakane, and H. Imai, "Biological Sequence Compression Algorithms", *Genome Informatics Workshop,* Universal Academy Press, pp. 43–52, 2000.

[13] J.J. Rissanen and G.G. Langdon, "Arithmetic Coding", *IBM J. Research and Development,* vol. 23, no. 2, pp. 149-162, March 1979.

[14] J. Rissanen, "Fisher information and stochastic complexity", *IEEE Transactions on Information Theory*, vol. 42, pp. 40–47, Jan. 1996.

[15] J. Rissanen and T. Roos, " Conditional NML universal models". In Information Theory and Applications Workshop (ITA-07), January 29–February 2, 2007, University of California, San Diego. Electronic publication. Available from http://eprints.pascal-network.org/archive/ 00002972/01/CNMLITA.pdf.

Chromosome	Wildcards included		Wildcards omitted	
	NML-0 + NML-1	*NML-0 + constr. NML-1*	*NML-0 + NML-1*	*NML-0 + constr. NML-1*
chr1	1.466	1.464	1.644	1.641
chr2	1.621	1.619	1.664	1.662
chr3	1.624	1.621	1.672	1.669
chr4	1.610	1.608	1.653	1.651
chr5	1.618	1.616	1.650	1.648
chr6	1.626	1.624	1.664	1.661
chr7	1.575	1.570	1.614	1.609
chr8	1.621	1.620	1.670	1.668
chr9	1.377	1.376	1.608	1.607
chr10	1.583	1.582	1.641	1.640
chr11	1.595	1.592	1.647	1.644
chr12	1.601	1.600	1.653	1.651
chr13	1.413	1.412	1.689	1.687
chr14	1.380	1.380	1.667	1.666
chr15	1.311	1.310	1.618	1.616
chr16	1.398	1.396	1.574	1.573
chr17	1.517	1.516	1.599	1.598
chr18	1.638	1.638	1.709	1.708
chr19	1.296	1.295	1.482	1.481
chr20	1.582	1.582	1.694	1.694
chr21	1.228	1.228	1.701	1.701
chr22	1.118	1.118	1.610	1.610
chrX	1.500	1.498	1.550	1.548
chrY	0.513	0.513	1.149	1.149
Average	1.450	1.449	1.618	1.616

Table 1: Comparison of NML-based algorithms on the human genome, when the matching pattern is encoded by the winner between order-0 NML and order-1 NML (column marked as NML-0 + NML-1) and by the winner between order-0 NML and constrained order-1 NML (column marked as NML-0 + constr. NML-1). The compression efficiency is expressed as the ratio between the compressed size measured in bits, and the number of symbols (the smaller the better).

[16] É. Rivals, J.P. Delahaye, M. Dauchet, and O. Delgrange, "A Guaranteed Compression Scheme for Repetitive DNA Sequences", LIFL Lille I Univ., Tech. Rep. IT–285, 1995.

[17] Y.M. Shtarkov. Universal sequential coding of single messages. Translated from *Problems of Information Transmission*, Vol. 23, No. 3, 3–17, July-September 1987.

[18] I. Tabus, G. Korodi, and J. Rissanen, "DNA sequence compression using the normalized maximum likelihood model for discrete regression", *Data Compression Conference 2003,* DCC '03, pp. 253–262, 2003.

[19] P. Whittle, "Some distribution and moment formulae for the Markov chain", *Journal of the Royal Statistical Society. Series B (Methodological),* vol. 17, no. 2, pp. 235–242, 1955.

[20] The Human Genome,
`ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/April_14_2003`

Factorized NML Models

Petri Myllymäki Teemu Roos Tomi Silander Petri Kontkanen
Complex Systems Computation Group
Helsinki Institute for Information Technology
University of Helsinki and Helsinki University of Technology
e-mail: *firstname.lastname@hiit.fi*

Henry Tirri
Nokia Research Center
e-mail: *henry.tirri@nokia.com*

Abstract

We consider probabilistic graphical models where a directed acyclic graph represents a factorization of a joint probability distribution: the joint probability of the variables is represented as a product of conditional probabilities, one for each variable conditioned on its immediate parents in the graph. For this type of models, computing the normalized maximum likelihood (NML) is computationally very demanding. We suggest a computationally feasible alternative to NML, the factorized NML, where the normalization is done locally for each conditional distribution, and not globally.

1 Introduction

The Complex Systems Computation research group[1] (CoSCo) was established in the early 1990's at the Department of Computer Science of University of Helsinki. The group was first led by Professor Henry Tirri until 2002, and after that by Professor Petri Myllymäki. The first contact between CoSCo and Jorma Rissanen took place in 1996, in an evaluation of the HYPE research project, which was a part of a large research programme on Adaptive and Intelligent Systems, funded by Tekes, the Finnish Funding Agency for Technology and Innovation. For the evaluation, Jorma interviewed Henry in a one-to-one meeting, which did not go along quite the way we had expected. Namely, the very first thing Jorma did was to write the formula for Jeffreys prior on the board, and ask "What is this?". When Henry recognized the formula Jorma commented that he has read the papers and evidently Henry knows them so let's do some science. Then the rest of the session was spent on a pleasant conversation on recent developments of MDL. As a memento of this meeting, we still keep on the wall of our institute the drawings done during the session (see Figure 1).

All in all, it was apparent that Jorma had very carefully studied the material we had sent him beforehand, and he already had a clear opinion of our work. In his evaluation report, Jorma commends our work and points out that

The material in this paper was presented in part at the 2008 Information Theory and Applications Workshop (ITA-08), San Diego, CA, January–February 2008.

[1] `http://cosco.hiit.fi`

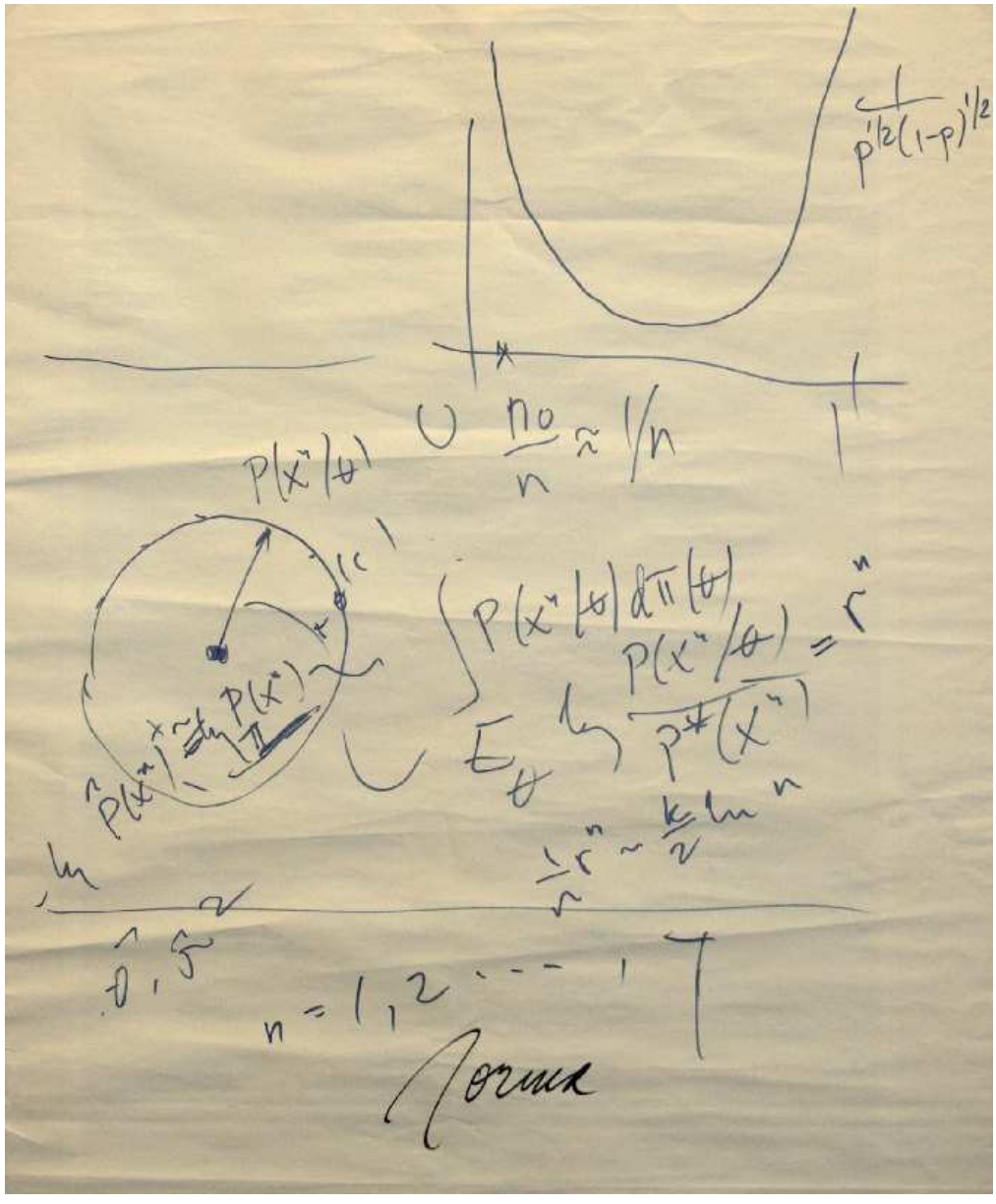


Figure 1: Jorma's notes from his first meeting with Henry Tirri in 1996.

> " It is particularly noteworthy that the difficult and important problem of determining the proper complexity of the models is done by new information-theoretic methods rather than resorting to usual ad hoc ones. "

He also had a quite clear opinion of the overall research programme, which focused on neural networks and genetic algorithms, which were popular at the time. Indeed, it is well known that Jorma is not scared to express his opinion quite directly, even if it is a negative one. In the evaluation of the research programme as a whole, Jorma chose to express his dissatisfaction somewhat indirectly, formulated cleverly in a seemingly positive statement:

> " On the whole, the research level of the teams using mainly the neural network techniques is in my opinion comparable to the general international level, which itself with a few exceptions, such as the work of A. Barron, is not particularly high. "

We were kind of an oddball in the programme, as we were just in the middle of a process of moving from neural networks and case-based reasoning to parametric probabilistic models. For the models we had started to explore—Bayesian networks, finite mixture models, Naive Bayes and other logistic regression type of classifiers—model regularization was clearly one of the central problems, and we were immediately intrigued by MDL. This interest had nothing to do with Jorma being Finnish, perhaps it was the information-theoretic approach that appealed to us as computer scientists. Actually, for a long time we discussed with Jorma in English only, and only later we have started to use more Finnish, at least for less technical discussions (involving often important topics like good food, beer, and soccer).

Our view of MDL was initially pretty ad hoc, and we, as many researchers still do, first employed the simple two-part/BIC types of codes, and later the "Bayes mixture" approach with various parameter priors [1–4]. Nevertheless, we soon felt increasing uneasiness with the arbitrariness of choosing the parameter priors, and we shared with Jorma the feeling that taking the subjective Bayesian approach is not as unproblematic as people often think, and that playing with the parameter priors is not an intuitively easy task after all, leading easily to anomalies in practical applications.

We kept seeing Jorma more and more often either in Helsinki or somewhere around the globe, and our appreciation towards him as a person and as a scientist was increasing. In addition to the pleasure of having a personal contact with Jorma, our work on MDL was greatly influenced by Peter Grünwald from CWI, Amsterdam, who met Petri Myllymäki in 1996 in a workshop organized by the NeuroCOLT working group of the European Union. Peter helped the CoSCo people to understand the new theoretical framework behind MDL, like the normalized maximum likelihood code, and we started working together in this field. Peter also came to Helsinki for a two month visit in 1997. Our joint work concentrated on issues like supervised learning, predictive distributions and choosing the parameter priors [5–11]. Quite interestingly, we were already then considering sequential (predictive) variants of MDL, which have recently gained popularity—more about them later. The co-operation between CWI and Helsinki has continued to this day, e.g. in the Pascal Network of Excellence[2], where Myllymäki and Grünwald are currently leading the Pascal Special Interest Group on Information-Theoretic Modeling. We are also jointly maintaining a popular web site[3] offering a (hopefully) useful portal to MDL-related work world-wide.

One of the active research areas in CoSCo nowadays is to study how to compute the NML criterion for Bayesian networks. This parametric model has become quite popular, and one

[2] `http://www.pascal-network.org`
[3] `http://www.mdl-research.org`

of the most popular freely available tools, the B-Course software[4], was developed and is being maintained by CoSCo. However, for practical applications, this model family introduces a couple of serious problems. First, the model structures are represented as acyclic directed graphs, which are superexponential in number. This makes the search for the best model structure a most difficult problem, which can currently be solved in reasonable time only for moderate size networks [12]. Nevertheless, perhaps even more crucial problem than how to find a good model, is the question of optimality: good in what sense?

Traditionally, in the Bayesian network community the models are evaluated by their posterior probability, which in the discrete Multinomial-Dirichlet setting can be computed in closed form, which leads to the popular BDe (Bayesian Dirichlet equivalent) score [13]. However, our recent work shows that the shape of the posterior is quite sensitive with respect to the choice of the hyperparameters of the Dirichlet prior [14]. NML would of course avoid this problem by offering a non-informative score that is not dependent on any parameter prior, but unfortunately, no efficient method for computing NML for Bayesian networks in general has yet been discovered. In the CoSCo group, we have gradually moved towards this goal by developing computationally efficient algorithms for independent multinomial variables, or equivalently, a Bayesian network with no arcs [15], for the Naive Bayes model [16, 17], and for tree-structured Bayesian networks [18, 19]. As an interesting application of the algorithm for computing the NML efficiently in the multinomial case, we can mention the minimax-optimal histogram density estimator suggested in [20].

As another active area of collaboration with Jorma, we have been focusing on MDL-based approaches to signal denoising. Starting from the original MDL denoising paper [21], we have been able to develop improved denoising methods [22], which are more robust with different levels of noise, achieve better frequency adaptivity, and employ the "soft thresholding" technique found very useful in denoising methods based on other approaches. For an illustration of denoising, see Figure 2. (The image in the example represents an Inter Milan soccer player in the 1950's. As many of us know, Jorma has always been a great fan of soccer, and a talented player himself: he even got an invitation in the early 1950's for a try-out in Milan, but the entrance examination for the Helsinki University of Technology was at the same time, and Jorma made, according to his own words, "a wrong decision" and chose science over football. Later he hurt his knee doing pole vault during his military service in the Finnish army, which finally ended any ideas about a potential career as a professional football player. This was a lucky strike for the IBM soccer team, who enjoyed having Jorma play for them for many years.)

One of the conclusions of the still ongoing work on denoising is the observation that the "model index", identifying the optimal subset of wavelet coefficients, forms a practically important part of the overall code length, and should not be ignored like was done in the original denoising paper. A similar phenomenon was observed already in the context of clustering [16]. However, Jorma was not after all very surprised by the result: he had of course always been aware of the missing part of his code, he just never thought it would make a difference in practice.

As the problem of computing NML for Bayesian networks is so difficult, we started to consider alternative solutions, other similar type of scoring functions that could be used instead of NML. It is probably appropriate to point out that also the non-informative Bayesian solution of using the Jeffreys prior is computationally NP-hard [10]. As already noted, we were already early on quite interested in predictive forms of MDL, while Jorma did not seem to share our interest. "Forget about prediction" was a frequently heard comment made by him when we tried to suggest exploring this area. One could have thought that Jorma did not want to touch the elaborate

[4] `http://b-course.hiit.fi`

Figure 2: The MDL denoising methods in action. *Top row* (from left) Original (size 128×128); noisy (noise std.dev. 20.0); original MDL denoising [21]. *Bottom row:* Left to right, gradual improvements of the MDL denoising method [22].

NML framework he had created, but as it would turn out, nothing was further from the truth. Already since 2004–2005, having studied a paper by Takimoto and Warmuth [23], we had started discussing the idea of sequential type NML variants in our group. Even though we found the topic potentially worthwhile, we couldn't see any obvious extensions beyond the basic idea. When we finally introduced the idea to Jorma in 2006, he was suddenly full of new ideas, leading to sequential NML (see Sec. 3 below) and many other novel innovations, and he was more than ready to abandon the "old" NML as obsolete—much more than we were! All in all, Jorma has often proved to be so fast and dynamic in his work that we, many being less than half of his age, have had hard time trying to keep up.

As the most recent result of our research on NML-like universal models for Bayesian networks, we introduce in this paper the *factorized NML* (fNML) model. The rest of the paper is organized as follows: In Sections 2 and 3 we discuss the normalized maximum likelihood (NML) and sequentially normalized maximum likelihood (sNML) models, respectively. In Section 4 we review the basics of Bayesian networks. The factorized NML model is introduced in Section 5, where it is also shown to be computationally feasible for all Bayesian networks. The new model is philosophically a relative of the sequential NML models discussed in Section 3. Finally, in Section 6, we present experimental results, demonstrating that fNML compares favorably in a model selection task, relative to the current state-of-the-art.

2 Normalized Maximum Likelihood Models

Before describing the sequential NML and factorized NML models, we fix some notation and review some basic properties of the well-known NML model. Let

$$x^n := \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,m} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,m} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_{1,:} \\ \mathbf{x}_{2,:} \\ \vdots \\ \mathbf{x}_{n,:} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_{:,1} \mathbf{x}_{:,2} \cdots \mathbf{x}_{:,m} \end{pmatrix} \ ,$$

be a data matrix where each row, $\mathbf{x}_{i,:} = (x_{i,1}, x_{i,2}, \ldots, x_{i,m}), 1 \leq i \leq n$, is an m-dimensional observation vector, and columns of x^n are denoted by $\mathbf{x}_{:,1}, \ldots, \mathbf{x}_{:,m}$.

A parametric probabilistic model $\mathcal{M} := \{p(x^n \ ; \ \theta) \ : \ \theta \in \Theta\}$, where Θ is a parameter space, assigns a probability mass or density value to the data. A *universal model* for $\mathcal{M}$ is a single distribution that, roughly speaking, assign almost as high a probability to any data as the the maximum likelihood parameters $\hat{\theta}(x^n)$.

Formally, a universal model $\hat{p}(x^n)$ satisfies

$$\lim_{n \to \infty} \frac{1}{n} \ln \frac{p(x^n \ ; \ \hat{\theta}(x^n))}{\hat{p}(x^n)} = 0 \ , \tag{1}$$

i.e., the log-likelihood ratio, often called the 'regret', is allowed to grow sublinearly in the sample size n. The celebrated *normalized maximum likelihood* (NML) universal model [24, 25]

$$p_{\mathrm{NML}}(x^n) := \frac{p(x^n \ ; \ \hat{\theta}(x^n))}{C_n} \ , \qquad C_n = \int_{\mathcal{X}^n} p(x^n \ ; \ \hat{\theta}(x^n)) \, dx^n$$

is the unique minimax optimal universal model in the sense that the worst-case regret is minimal. In fact, it directly follows from the definition that the regret is a constant dependent only on the sample size n:

$$\ln \frac{p(x^n \ ; \ \hat{\theta}(x^n))}{p_{\mathrm{NML}}(x^n)} = \ln C_n \ .$$

For some model classes, the normalizing factor is finite only if the range $\mathcal{X}^n$ of the data is restricted, see e.g. [21, 24, 26]. For discrete models, the normalizing constant, C_n, is given by a sum over all data matrices of size $m \times n$:

$$C_n = \sum_{x^n \in \mathcal{X}^n} p(x^n \ ; \ \hat{\theta}(x^n)) \ .$$

The practical problem arising in applications of the NML universal model is then to evaluate the normalizing constant. For continuous models the integral can be solved in closed form for only a few specific models. For discrete models, the time complexity of the naive solution, i.e., summing over all possible data matrices, grows exponentially in both n and m, and quickly becomes intractable. Even the second-most naive solution, summing over equivalence classes of matrices, sharing the same likelihood value, is usually intractable even though often polynomial in n.

The usual Fisher information approximation [24]

$$\ln C_n = \frac{k}{2} \ln \frac{n}{2\pi} + \ln \int_{\Theta} \sqrt{\det I(\theta)} \, d\theta + o(1) \ ,$$

where k is the dimension of the parameter space, is also non-trivial to apply due to the integral involving the Fisher information $I(\theta)$. Using only the leading term (with or without 2π), i.e., the BIC criterion [27], gives a rough approximation which, as a rule, performs worse in model selection tasks than more refined approximations or, ideally, the exact solution, see e.g. [28, Chap. 9].

3 Sequentially Normalized ML Models[4]

A recent family of variants of NML, called the *sequentially* (or *conditional*) *normalized maximum likelihood* (sNML) [29, 30] has similar minimax properties like NML but is often significantly easier to use in practice.

For data matrix $x^n = (\mathbf{x}_{1,:}, \mathbf{x}_{2,:}, \ldots, \mathbf{x}_{n,:})'$, the sNML-1 model is defined as

$$p_{\mathrm{sNML1}}(x^n) := \prod_{i=1}^{n} \frac{p(\mathbf{x}_{i,:} \mid x^{i-1} \,;\, \hat{\theta}(x^i))}{K_i(x^{i-1})} \ , \tag{2}$$

$$K_i(x^{i-1}) := \int p(\mathbf{x}_{i,:} \mid x^{i-1} \,;\, \hat{\theta}(x^i))\, d\mathbf{x}_{i,:} \ , \tag{3}$$

where normalization ensures that each factor in the product is a proper density function.

In some cases it is necessary to use a separate density, say $q(x^{n_0})$, for the first n_0 observations, with n_0 large enough, so that the maximimized likelihood is well-defined for longer sequences x^i with $i > n_0$. For instance, in linear regression n_0 has to be at least the number of regressor variables plus one.

Second variant (sNML-2). There is also another variant of sNML, which we call here sNML-2. It can be defined in analogy with (2) as follows:

$$p_{\mathrm{sNML2}}(x^n) := \prod_{i=1}^{n} \frac{p(x^i \,;\, \hat{\theta}(x^i))}{K'_i(x^{i-1})} \ , \tag{4}$$

$$K'_i(x^{i-1}) := \int p(x^i \,;\, \hat{\theta}(x^i))\, d\mathbf{x}_{i,:} \ .$$

Using the sNML-2 model is equivalent to predicting the ith observation using the standard NML model defined for sequences of length i. Formally we have

$$p_{\mathrm{NML}}(\mathbf{x}_{i,:} \mid x^{i-1}) = p_{\mathrm{sNML2}}(\mathbf{x}_{i,:} \mid x^{i-1}) \ .$$

Note that the standard NML model is not in general a stochastic process, which makes it possible that

$$p_{\mathrm{NML}}(\mathbf{x}_{i,:} \mid x^{i-1}) \neq \sum_{\mathbf{x}_{i+1,:}} p_{\mathrm{NML}}(\mathbf{x}_{i,:}, \mathbf{x}_{i+1,:} \mid x^{i-1}) \ , \tag{5}$$

and hence, typically two NML models, defined for sequences of different lengths, give different predictions. In contrast, both sNML-1 and sNML-2 are by definition stochastic processes, so that for them we always have an equality in (5).

[4] This section is mostly based on as yet unpublished work by Rissanen, Myllymäki, and Roos.

Regrets Visualized. Figure 3 gives a visualization of the regrets of four universal models in the Bernoulli case: the Laplace predictor ("add-one"), the Krichevsky–Trofimov predictor ("add-half"), sNML-2, and NML. For NML, the initial sequence probabilities, $q(x^t)$, are obtained from a fixed NML model, defined for $n = 5$, by summing over the possible continuations of length $n - t$.

Note that for NML, while the intermediate regrets, for $t < n$, depend on the prefix x^t, the total regret for x^n is a constant. For sNML, the difference between the regret for x^t and x^{t+1} is constant with respect to x_t but varies with x^{t-1}; in the figure this means that each pair of edges originating from the same branching point are of equal length, but their length depends on the path from the origin. For the Bernoulli model, SNML-1 is equivalent to the Laplace predictor. Figure 4 shows the regrets with $n = 5$ as a function of the number of 1s.

Related Work. The sNML-2 model has been analysed earlier in conjuction with discrete Markov models, including as a special case the Bernoulli model, by Shtarkov [25] (see his Eq. 45). Also, Takimoto and Warmuth [23] analyze a slightly more restricted minimax problem, the solution of which agrees with sNML-2 for Markov models. Grünwald [29] uses the term "conditional NML" (CNML) for a family of universal models, conditioned on an initial sequence without considering the joint model obtained as a product of such conditional densities. Our sNML-1 corresponds to his CNML-3, and our sNML-2 corresponds to his CNML-2. The conditional mixture codes studied by Liang and Barron [31] are also closely related to sNML, and have similar minimax properties.

4 Bayesian Networks

In Sec. 5, we describe a new NML variant, similar to the sNML models discussed in the previous section. This new variant gives a computationally feasible universal model, and a corresponding model selection criterion, for general Bayesian network models. This section presents the necessary background in Bayesian networks.

First, let us associate with the columns, $\mathbf{x}_{:,1}, \ldots, \mathbf{x}_{:,m}$, a directed acyclic graph (DAG), $\mathcal{G}$, so that each column is represented by a node. Each node, $X_j, 1 \leq j \leq m$, has a (possibly empty) set of *parents*, Pa_j, defined as the set of nodes with an outgoing edge to node X_j. Without loss of generality, we require that all the edges are directed towards increasing node index, i.e., $\mathrm{Pa}_j \subseteq \{1, \ldots, j-1\}$. If this is not the case, the columns in the data, and the corresponding nodes in the graph, can be simply relabeled, which does not change the resulting model. Figure 5 gives an example.

The idea is to model dependencies among the nodes (i.e. columns) by defining the joint probability distribution over the nodes in terms of *local distributions*: each local distribution specifies the conditional distribution of each node given its parents, $p(X_j \mid \mathrm{Pa}_j), 1 \leq j \leq m$. It is important to notice that these are *not* dependencies among the subsequent rows of the data matrix x^n, but dependencies 'inside' each row, $\mathbf{x}_{i,:}, 1 \leq i \leq n$. Indeed, in all of the following, we assume that the rows are independent realizations of a fixed (memoryless) source.

The local distributions can be modeled in various ways, but here we focus on the discrete case. The probability of a child node taking value $x_{i,j} = r$ given the parent nodes' configuration, $\mathrm{pa}_{i,j} = \mathbf{s}$, is determined by the parameter

$$\theta_{j|\mathrm{Pa}_j}(r, \mathbf{s}) = p(x_{i,j} = r \mid \mathrm{pa}_{i,j} = \mathbf{s}\,;\, \theta_{j|\mathrm{Pa}_j}) \quad , \quad 1 \leq i \leq n, 1 \leq j \leq m \ ,$$

where the notation $\theta_{j|\mathrm{Pa}_j}(r, \mathbf{s})$ refers to the component of the parameter vector $\theta_{j|\mathrm{Pa}_j}$ indexed by the value r and the configuration $\mathbf{s}$ of the parents of X_j. For empty parent sets, we let $\mathrm{pa}_{i,j} \equiv 0$.

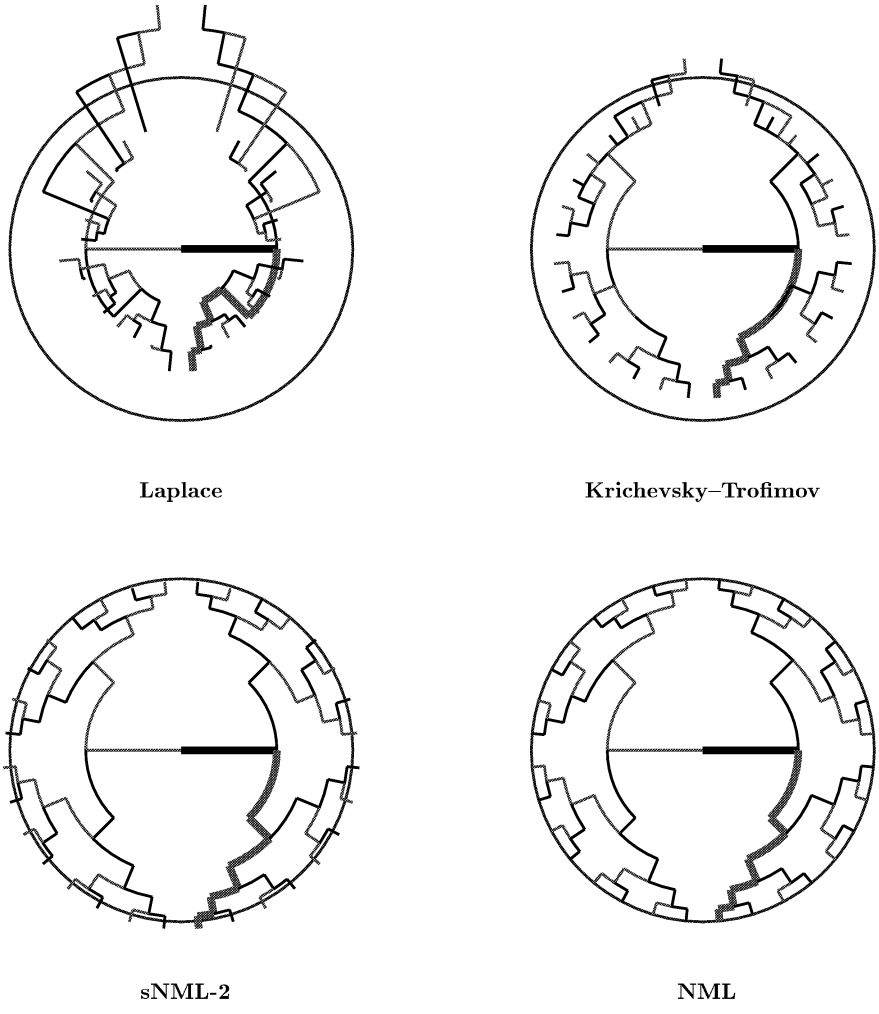


Figure 3: Regrets of four universal models in the Bernoulli case. Each path from the origin (center) to the boundary represents a binary sequence of length $n = 5$. Red edges correspond to `1`s, black edges to `0`s. The path for sequence `01111` is emphasized. The distances from the origin of the branching points are given by the regrets $\ln[p(x^t\ ;\ \hat{\theta}(x^t))/q(x^t)]$ for each prefix x^t. The blue circle shows the regret of NML. Note the similarity between sNML-2 and NML.

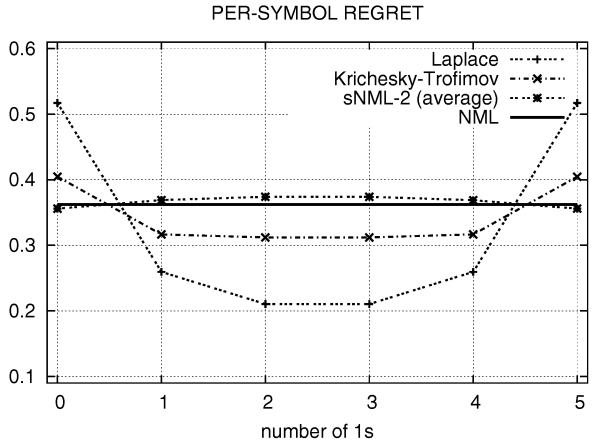


Figure 4: Per-symbol regrets of four universal models in the Bernoulli case as a function of the number of 1s in the sequence with $n = 5$ (for the same figure with $n = 30$, see [30]). For sNML-2 the regret depends not only on the number of 1s, but also on the actual sequence. (The dependency is *very* slight, see Fig. 3.) The graph shows the average regret.

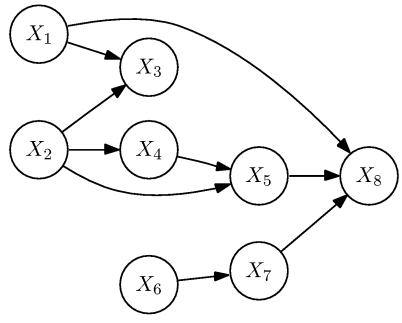


Figure 5: An example of a directed acyclic graph (DAG). The parents of node X_8 are $\{X_1, X_5, X_7\}$. The descendants of X_4 are $\{X_5, X_8\}$.

For instance, consider the graph of Fig. 5; on each row, $1 \leq i \leq n$, the parent configuration of column $j = 8$ is the vector $\mathrm{pa}_{i,8} = (x_{i,1}, x_{i,5}, x_{i,7})$; the parent configuration of column $j = 1$ is $\mathrm{pa}_{i,1} = 0$, etc.

The joint distribution is obtained as a product of local distributions:

$$p(x^n \; ; \; \theta) = \prod_{j=1}^{m} p(\mathbf{x}_{:,j} \mid \mathrm{Pa}_j \; ; \; \theta_{j|\mathrm{Pa}_j}) \; . \tag{6}$$

This type of probabilistic graphical models are called Bayesian networks [32]. Factorization (6) entails a set of conditional independencies, characterized by so called Markov properties, see [33]. For instance, the *local Markov property* asserts that each node is independent of its non-descendants given its parents, generalizing the familiar Markov property of Markov chains.

It is now possible to define the NML model based on (6) and a fixed graph structure $\mathcal{G}$:

$$p_{\mathrm{NML}}(x^n \; ; \; \mathcal{G}) = \frac{\prod_{j=1}^{m} p(\mathbf{x}_{:,j} \mid \mathrm{Pa}_j \; ; \; \hat{\theta}(x^n))}{C_n} \; , \tag{7}$$

where

$$C_n = \sum_{x^n} \prod_{j=1}^{m} p(\mathbf{x}_{:,j} \mid \mathrm{Pa}_j \; ; \; \hat{\theta}(x^n)) \; . \tag{8}$$

The required maximum likelihood parameters are easily evaluated since it is well known that the ML parameters are equal to the relative frequencies:

$$\hat{\theta}_{j|\mathrm{Pa}_j}(r, \mathbf{s}) = \frac{\big|\{i \; : \; x_{i,j} = r, \mathrm{pa}_{i,j} = \mathbf{s}\}\big|}{\big|\{i' \; : \; \mathrm{pa}_{i',j} = \mathbf{s}\}\big|} \; , \tag{9}$$

where $|S|$ denotes the cardinality of set S. However, as pointed out in Sec. 2, summing over all possible data matrices is not tractable except in toy problems where n and m are both very small. Efficient algorithms have been discovered only recently for restricted graph structures [17–19].

5 Factorized NML Models

As a computationally less demanding alternative to NML in the context of Bayesian networks, we define the *factorized NML* (fNML) in a similar spirit as sNML. We let the joint probability distribution be given by a product of *locally* normalized maximum likelihood distributions:

$$p_{\mathrm{fNML}}(x^n \; ; \; \mathcal{G}) := \prod_{j=1}^{m} \frac{p(\mathbf{x}_{:,j} \mid \mathrm{Pa}_j \; ; \; \hat{\theta}(x^n))}{Z_j(\mathrm{Pa}_j)} \tag{10}$$

$$= \frac{\prod_{j=1}^{m} p(\mathbf{x}_{:,j} \mid \mathrm{Pa}_j \; ; \; \hat{\theta}(x^n))}{Z(x^n)} \; , \tag{11}$$

where each of the local normalizing factors

$$Z_j(\mathrm{Pa}_j) = \sum_{X'_j} p(X'_j \mid \mathrm{Pa}_j \; ; \; \hat{\theta}(X'_j, \mathrm{Pa}_j)) \tag{12}$$

is a sum over all possible instantiations of column $\mathbf{x}_{:,j}$, and the global normalizing factor

$$Z(x^n) = \prod_{j=1}^{m} \sum_{X'_j} p(X'_j \mid \mathrm{Pa}_j \; ; \; \hat{\theta}(X'_j, \mathrm{Pa}_j)) \tag{13}$$

is a product of the local normalizing factors. The local normalizing factors $Z_j(\mathrm{Pa}_j)$ can be decomposed further into simple multinomial NML normalization constants, one for each parent configuration in Pa_j. Using the recently discovered linear-time algorithm [15] for the multinomial case, the total computation time becomes feasible even for large sample sizes and for many variables (columns).

In practice, we not only want to evaluate the likelihood of the data under a given model class, but we also wish to find the structure that maximizes the likelihood of the data. This is made hard by the fact that the number of possible DAG structure is superexponential. Unlike the standard NML criterion, the fNML criterion is 'modular' in the sense that it decomposes column-wise into independent terms. This enables the use dynamic programming techniques that find the global optimum in $o(n2^n)$ time, see [12], which is manageable for networks with up to about 30 nodes. For larger networks, local search heuristics are necessary.

Note that, as can be seen from (9), the maximum likelihood parameters of each local distribution, $\theta_{j|\mathrm{Pa}_j}$, depend only on column $\mathbf{x}_{:,j}$ and column(s) Pa_j. In particular, since we require $\mathrm{Pa}_j \subseteq \{1, \ldots, j-1\}$, we have

$$p(\mathbf{x}_{:,j} \mid \mathrm{Pa}_j \;;\; \hat{\theta}(x^n)) = p(\mathbf{x}_{:,j} \mid \mathrm{Pa}_j \;;\; \hat{\theta}(\mathbf{x}_{:,1}, \ldots, \mathbf{x}_{:,j})) = p(\mathbf{x}_{:,j} \mid \mathrm{Pa}_j \;;\; \hat{\theta}(\mathbf{x}_{:,j}, \mathrm{Pa}_j)) \;, \quad (14)$$

of which the second form, where only the first j columns appear, is the one that should be used in (10) by analogy with (2). Due to the above identity, the expressions are used interchangeably.

The sum-product view. It is interesting to compare the NML and fNML models. Consider Eqs. (7) and (11): the constant normalizer of NML, C_n, an exponential *sum of products*, is replaced in fNML by $Z(x^n)$, a *product of sums* that depends on the data. The fNML model can therefore be seen as 'cheating' by using a sum-product algorithm, where the distributive law (see [34])

$$\begin{cases} f(x_1, x_2) \equiv f(x_1) \\ g(x_1, x_2) \equiv g(x_2) \end{cases} \quad \Longrightarrow \quad \sum_{x_1, x_2} f(x_1, x_2) g(x_1, x_2) \quad = \quad \left(\sum_{x_1} f(x_1) \right) \left(\sum_{x_2} g(x_1) \right) \quad (15)$$

is applied to compute the sum in C_n even though the terms do not actually factor column-wise into independent parts. No cheating is necessary when the graph is empty, i.e., when $\mathrm{Pa}_j = \emptyset$ for all $1 \leq j \leq m$. This means that we have $Z(x^n) = C_n$, which by (7) and (11) implies that for empty graphs p_{NML} and p_{fNML} are equivalent.

The regrets of the two models are easily seen to be $\ln C_n$ and $\ln Z(x^n)$, for NML and fNML respectively. Notice also that the regret of fNML, $\ln Z(x^n)$, depends on the data only through the parents, $\mathrm{Pa}_j, 1 \leq j \leq m$, and hence, is independent of all the leaf nodes, i.e., nodes that have no descendants. Again, if the graph is empty, all nodes are leafs and $Z(x^n) = C_n$ for all x^n so that the NML and fNML models are equivalent.

Finally, we observe that for fNML the two variants of sNML, sNML-1 and sNML-2, coincide. Letting $x(j) := (\mathbf{x}_{:,1}, \mathbf{x}_{:,2}, \ldots, \mathbf{x}_{:,j})$ denote the first j columns, we obtain

$$\begin{aligned} p(x(j) \;;\; \hat{\theta}(x(j))) &= \prod_{l=1}^{j} p(\mathbf{x}_{:,l} \mid \mathrm{Pa}_l \;;\; \hat{\theta}(\mathbf{x}_{:,l}, \mathrm{Pa}_l)) \\ &= p(\mathbf{x}_{:,j} \mid \mathrm{Pa}_j \;;\; \hat{\theta}(x^n)) \prod_{l=1}^{j-1} p(\mathbf{x}_{:,l} \mid \mathrm{Pa}_l \;;\; \hat{\theta}(\mathbf{x}_{:,l}, \mathrm{Pa}_l)) \;, \end{aligned}$$

where both equalities depend on (14). The last factor on the right-hand side is independent of column $\mathbf{x}_{:,j}$. When the above is normalized with respect to $\mathbf{x}_{:,j}$, this factor cancels and we are left with $p(\mathbf{x}_{:,j} \mid \mathrm{Pa}_j \; ; \; \hat{\theta}(x^n))$, which is exactly what is normalized in (10). Hence, it doesn't matter whether we define fNML as in (10) or as the product over $1 \leq j \leq m$ of the normalized versions of $p(x(j) \; ; \; \hat{\theta}(x(j)))$, and sNML-1 is equivalent to sNML-2 for Bayesian network model classes.

6 Experiments

To empirically test performance of the fNML-criterion in Bayesian network structure learning task, we generated several Bayesian networks, and then studied how different model selection criteria succeeded in learning the model structure from data. The most often used selection criterion for the task is the BDe (Bayesian Dirichlet equivalent) score [13], but due to its sensitivity to the choice of prior hyperparameter, we chose two different versions of it: $\mathrm{BDe}_{0.5}$ and $\mathrm{BDe}_{1.0}$. We also included the Bayesian Information Criterion, BIC. All these scores can be interpreted as implementing some version of the MDL criterion or an approximation thereof.

We present the results for an experiment in which we generated 1800 different Bayesian network models, which we tried to learn back using the data generated from these models. We generated the networks using 5, 10 and 15 variables, and also varied the density and the parameters of the networks. We then generated 1000, 10000 and 10000 data vectors from each network, and tried to learn the models back using these data samples and different scoring criteria. It turned out that learning the models back with these sample sizes was practically possible only for smallest networks containing 5 nodes. However, varying the number of arcs and parameters did not seem to have a strong effect on the outcome. This made it possible us to concentrate on comparing the performance of different scoring criteria for different sample sizes (Figure 6).

The results clearly show that fNML excels with small sample sizes. With large sample sizes, the difference is not that big, which is hardly surprising, since asymptotically, they all converge to the data generating model. This result is significant, since BDe score(s) can be regarded as the current state-of-the-art. Furthermore, the fNML score is computationally no more demanding than the BDe score.

Acknowledgment

This work was supported in part by the Finnish Funding Agency for Technology and Innovation under projects Kukot and PMMA, by the Academy of Finland under project Civi, and by the IST Programme of the European Community, under the Pascal Network of Excellence, IST-2002-506778. We thank the reviewers for useful comments.

References

[1] P. Kontkanen, P. Myllymäki, T. Silander, and H. Tirri, "Comparing stochastic complexity minimization algorithms in estimating missing data," in *Proceedings of WUPES'97, the 4th Workshop on Uncertainty Processing*, Prague, Czech Republic, January 1997, pp. 81–90.

[2] ——, "On the accuracy of stochastic complexity approximations," in *Proceedings of the Causal Models and Statistical Learning Seminar*, London, UK, March 1997, pp. 103–117.

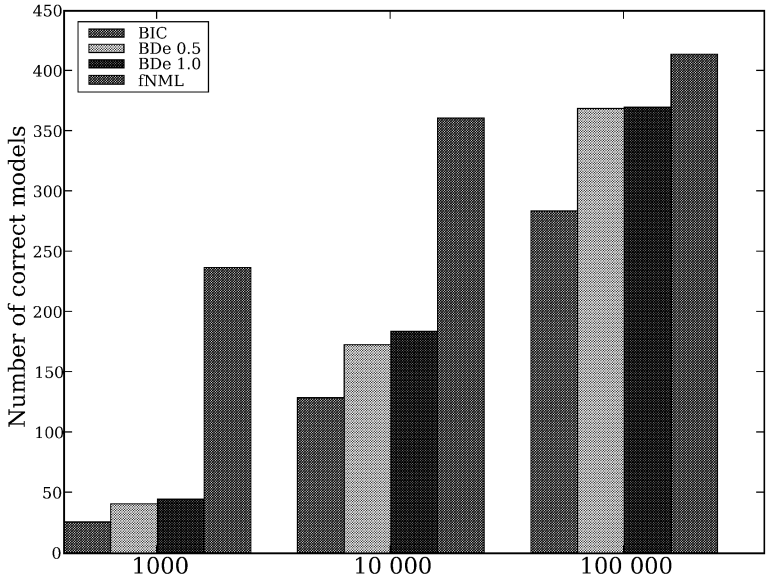


Figure 6: Number of correctly learned models in 1800 trials for sample sizes 1000, 10 000, and 100 000. For each sample size, the bars give the number of correctly learned models for (from left to right) the BIC, $\mathrm{BDe}_{0.5}$, $\mathrm{BDe}_{1.0}$, and fNML scores.

[3] P. Kontkanen, P. Myllymäki, and H. Tirri, "Comparing Bayesian model class selection criteria by discrete finite mixtures," in *Information, Statistics and Induction in Science*, D. Dowe, K. Korb, and J. Oliver, Eds. Proceedings of the ISIS'96 Conference, Melbourne, Australia: World Scientific, Singapore, August 1996, pp. 364–374.

[4] ——, "Experimenting with the Cheeseman-Stutz evidence approximation for predictive modeling and data mining," in *Proceedings of the Tenth International FLAIRS Conference*, D. Dankel, Ed., Daytona Beach, Florida, May 1997, pp. 204–211.

[5] P. Grünwald, P. Kontkanen, P. Myllymäki, T. Roos, H. Tirri, and H. Wettig, "Supervised posterior distributions," 2002, presented at the Seventh Valencia International Meeting on Bayesian Statistics, Tenerife, Spain.

[6] P. Grünwald, P. Kontkanen, P. Myllymäki, T. Silander, and H. Tirri, "Minimum encoding approaches for predictive modeling," in *Proceedings of the 14th International Conference on Uncertainty in Artificial Intelligence (UAI'98)*, G. Cooper and S. Moral, Eds. Madison, WI: Morgan Kaufmann Publishers, San Francisco, CA, July 1998, pp. 183–192.

[7] P. Kontkanen, P. Myllymäki, T. Silander, H. Tirri, and P. Grünwald, "Comparing predictive inference methods for discrete domains," in *Proceedings of the Sixth International Workshop on Artificial Intelligence and Statistics*, Ft. Lauderdale, Florida, January 1997, pp. 311–318.

[8] ——, "Bayesian and information-theoretic priors for Bayesian network parameters," in *Machine*

Learning: ECML-98, Proceedings of the 10th European Conference, ser. Lecture Notes in Artificial Intelligence, Vol. 1398, C. Nédellec and C. Rouveirol, Eds. Springer-Verlag, 1998, pp. 89–94.

[9] ——, "On the small sample behaviour of Bayesian and information-theoretic approaches to predictive inference," 1998, presented at the Sixth Valencia International Meeting on Bayesian Statistics, Alcossebre, Spain.

[10] ——, "On predictive distributions and Bayesian networks," *Statistics and Computing*, vol. 10, pp. 39–54, 2000.

[11] T. Roos, H. Wettig, P. Grünwald, P. Myllymäki, and H. Tirri, "On discriminative Bayesian network classifiers and logistic regression," *Machine Learning*, vol. 59, no. 3, pp. 267–296, 2005.

[12] T. Silander and P. Myllymäki, "A simple approach for finding the globally optimal Bayesian network structure," in *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*, R. Dechter and T. Richardson, Eds. AUAI Press, 2006, pp. 445–452.

[13] D. Heckerman, D. Geiger, and D. Chickering, "Learning Bayesian networks: The combination of knowledge and statistical data," *Machine Learning*, vol. 20, no. 3, pp. 197–243, September 1995.

[14] T. Silander, P. Kontkanen, and P. Myllymäki, "On sensitivity of the MAP Bayesian network structure to the equivalent sample size parameter," in *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence*, R. Parr and L. van der Gaag, Eds. AUAI Press, 2007, pp. 360–367.

[15] P. Kontkanen and P. Myllymäki, "A linear-time algorithm for computing the multinomial stochastic complexity," *Information Processing Letters*, vol. 103, no. 6, pp. 227–233, 2007.

[16] P. Kontkanen, P. Myllymäki, W. Buntine, J. Rissanen, and H. Tirri, "An MDL framework for data clustering," in *Advances in Minimum Description Length: Theory and Applications*, P. Grünwald, I. Myung, and M. Pitt, Eds. The MIT Press, 2006.

[17] T. Mononen and P. Myllymäki, "Fast NML computation for naive Bayes models," in *Proc. 10th International Conference on Discovery Science*, Sendai, Japan, October 2007.

[18] P. Kontkanen, H. Wettig, and P. Myllymäki, "NML computation algorithms for tree-structured multinomial Bayesian networks," *EURASIP Journal on Bioinformatics and Systems Biology*, 2007.

[19] H. Wettig, P. Kontkanen, and P. Myllymäki, "Calculating the normalized maximum likelihood distribution for Bayesian forests," in *Proc. IADIS International Conference on Intelligent Systems and Agents*, Lisbon, Portugal, July 2007.

[20] P. Kontkanen and P. Myllymäki, "MDL histogram density estimation," in *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, M. Meila and S. Shen, Eds., March 2007.

[21] J. Rissanen, "MDL denoising," *IEEE Transactions on Information Theory*, vol. 46, no. 7, pp. 2537–2543, 2000.

[22] T. Roos, P. Myllymäki, and J. Rissanen, "MDL denoising revisited," 2006, submitted for publication. Preprint arXiv cs.IT/0609138.

[23] E. Takimoto and M. Warmuth, "The last-step minimax algorithm," in *Proc. 11th International Conference on Algorithmic Learning Theory*, 2000, pp. 279–290.

[24] J. Rissanen, "Fisher information and stochastic complexity," *IEEE Transactions on Information Theory*, vol. 42, no. 1, pp. 40–47, January 1996.

[25] Y. Shtarkov, "Universal sequential coding of single messages," *Problems of Information Transmission*, vol. 23, pp. 3–17, 1987.

[26] S. de Rooij and P. Grünwald, "An empirical study of minimum description length model selection with infinite parametric complexity," *Journal of Mathematical Psychology*, vol. 50, no. 2, pp. 180–192, 2006.

[27] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, pp. 461–464, 1978.

[28] J. Rissanen, *Information and Complexity in Statistical Modeling.* Springer, 2007.

[29] P. Grünwald, *The Minimum Description Length Principle.* MIT Press, 2007.

[30] J. Rissanen and T. Roos, "Conditional NML models," in *Information Theory and Applications Workshop (ITA-07)*, San Diego, CA, January–February 2007.

[31] F. Liang and A. Barron, "Exact minimax strategies for predictive density estimation, data compression, and model selection," *IEEE Transactions on Information Theory*, vol. 50, no. 11, pp. 2708–2726, 2004.

[32] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann Publishers, San Mateo, CA, 1988.

[33] S. Lauritzen, *Graphical Models.* Oxford University Press, 1996.

[34] S. M. Aji and M. R. J., "The generalized distributive law," *IEEE Transactions on Information Theory*, vol. 46, no. 2, pp. 325–343, 2000.

Towards the multicomponent MDL denoising

Janne Ojanen[1], Jukka Heikkonen[1,2], Kimmo Kaski[1]

[1] Helsinki University of Technology TKK, Department of Biomedical Engineering and Computational Science, P.O.Box 9203 (Tekniikantie 14), FI-02015 TKK, Finland

[2] European Commission - Joint Research Centre, Institute for the Protection and Security of the Citizen (IPSC), G04 Maritime Affairs (Fishreg Sector), TP 051, Via Fermi 1, I-21020 Ispra (VA), Italy

contact: jiojanen@lce.hut.fi

Abstract

The minimum description length (MDL) wavelet denoising approach based on minimizing the normalized maximum likelihood code length can be extended to incorporate more than one non-noise component. We give an informal outline of a multicomponent approach, and present preliminary results showing that minimizing the code length for the three-component model leads to a separation of components with different characteristics.

1 Introduction

Removing noise from measurements is an important step in many applications. Wavelet denoising has been shown to give good results and it is no surprise that research in this field has been very active. A large number of methods based on different thresholding strategies have been proposed in the literature; see [1] for a comparison of performance of several popular methods. The minimum description length (MDL) principle has proven to be a useful tool also in wavelet denoising, for example, see [2, 3, 4, 5, 6, 7, 8].

Typically denoising is based on a model with an underlying informative signal and additive i.i.d. random noise. However, it is possible that in addition to the random noise there may be other disturbing signal elements, or that the informative signal is comprised of several different components which we may want to observe, separate or remove. For example, in monitoring the condition of a rotating machine a bearing fault may be seen as a new component in the measured vibration signal, or in image analysis the varying lighting conditions for a line camera may add a new component in the resulting image. If there is more than one informative component in the noisy measured data, a multicomponent approach may result in better performance than the denoising approach with an informative and a noise component, whether the aim is denoising or also the separation of components.

2 Wavelet denoising

Usually the wavelet denoising methods are based on the Discrete Wavelet Transform (DWT), $c^n = \mathbf{W}^T x^n$, of the data x^n. The wavelet basis vectors are often chosen to span a complete orthonormal basis

so that the Parseval's equality $||c^n|| = ||x^n||$ holds. In this article we consider the hard thresholding approach, in which the wavelet coefficients assumed to correspond to noise are set to zero, while the remaining coefficients are taken to represent the underlying informative signal. Typically k coefficients with largest magnitude are retained according to some optimality criterion. Given the modified coefficients $\hat{c}^n$ (k retained and $n-k$ zeroes) the reconstructed signal can be defined as the inverse DWT $\hat{x}^n = \mathbf{W}\hat{c}^n$. The solution by Rissanen [4] for obtaining $\hat{c}^n$ is to choose the subset of basis vectors resulting in the shortest description of the data x^n, determined as the normalized maximum likelihood (NML) code length. Calculating the NML code length requires evaluating a normalizing integral which is undefined unless the integration range is restricted. This is equivalent to introducing hyperparameters restricting the maximum likelihood estimates. The hyperparameters are removed by a renormalization and a second-level NML model is obtained. The resulting code length can be approximated by a simple criterion.

Roos et al. [6, 7, 8] have shown that a criterion similar to the renormalization result can be obtained by a different derivation, details of which can be found in [7, 8]. In short, they define a model

$$c_j \sim \begin{cases} N(0, \sigma_I^2), & \text{if } j \in \gamma \\ N(0, \sigma_N^2), & \text{otherwise} \end{cases} \tag{1}$$

for the wavelet coefficients, in which each coefficient is distributed according to a zero-mean Gaussian density with variance σ_I^2 if it belongs to the set of informative coefficients indexed by γ, or according to a zero-mean Gaussian density with variance σ_N^2 if it represents noise, with the restriction $\sigma_I^2 \geq \sigma_N^2$. The code length for the data given the model class indexed by γ is obtained as the normalized maximum likelihood code length

$$-\ln f_{\text{NML}}(x^n; \gamma) = -\ln \frac{f(x^n; \hat{\sigma}_I^2, \hat{\sigma}_N^2)}{\int_{y^n} f(y^n; \hat{\sigma}_I^2, \hat{\sigma}_N^2) dy^n} , \tag{2}$$

and the optimal index set γ is obtained by minimizing the joint code length,

$$\min_{\gamma} \left[-\ln f_{\text{NML}}(x^n; \gamma) + L(\gamma) \right], \tag{3}$$

where $L(\gamma)$ is the code length for the model class. Roos et al. [7, 8] suggest using code length $L(\gamma) = \ln \binom{n}{k}$, where k refers to the number of coefficients for which $j \in \gamma$.

The orthonormality of the DWT allows to calculate the integral in the c^n domain. In order to calculate the normalizing integral also this derivation requires hyperparameters to bound the maximum likelihood estimates of the variances. However, the effect of hyperparameters can be ignored and the minimization of the total code length can be approximated by

$$\min_{\gamma} \left[\frac{k}{2} \ln \frac{1}{k} \sum_{j \in \gamma} c_j^2 + \frac{n-k}{2} \ln \frac{1}{n-k} \sum_{j \notin \gamma} c_j^2 + \frac{1}{2} \ln k(n-k) + \ln \binom{n}{k} \right] . \tag{4}$$

The criterion for the NML code length in Eq. 4 (first three terms) is the same as the one proposed in [4].

3 Multicomponent MDL Denoising

The basic model described in Section 2 can be extended by adding another Gaussian signal component, such that there will be three different zero-mean Gaussian components with variances σ_1^2 and σ_2^2 for

the two informative components, while σ_3^2 refers to noise with the restriction $\sigma_1^2 \geq \sigma_2^2 \geq \sigma_3^2$ holding. Following the treatment in [7, 8], the extended basic model for the wavelet coefficients is given by

$$c_j \sim \begin{cases} N(0, \sigma_1^2), & \text{if } j \in \gamma_1 \\ N(0, \sigma_2^2), & \text{if } j \in \gamma_2 \\ N(0, \sigma_3^2), & \text{if } j \in \gamma_3 \end{cases} , \tag{5}$$

where index sets γ_1 and γ_2 define the two informative components and γ_3 defines the noise component. The Gaussian model for the informative components may not be the most realistic, but it allows calculating the NML code length.

The NML code length for this model can be calculated in a manner following the derivation in [7] for the two-component denoising criterion. The derivation turns out to be straightforward since the normalizing integral factors into three parts, each depending only on the coefficients determined by the respective index set γ_i. Therefore, we may use the same approach for calculating the integrals as in [7] for the two-component approach. Given the index sets γ_1, γ_2 and γ_3 we can approximate the joint code length with

$$\sum_{i=1}^{3} \left(\frac{k_i}{2} \ln \frac{1}{k_i} \sum_{j \in \gamma_i} c_j^2 + \frac{1}{2} \ln k_i \right) + L(\gamma_1, \gamma_2, \gamma_3) , \tag{6}$$

where k_i refers to the number of coefficients for which $j \in \gamma_i$ and $L(\gamma_1, \gamma_2, \gamma_3)$ is the code length for the model class. Following similar reasoning as in [7], we may use code length $L(\gamma_1, \gamma_2, \gamma_3) = \ln \binom{n}{k_1, k_2, k_3}$.

Instead of performing an extensive optimization over all possible index sets γ_1, γ_2 and γ_3, we perform the minimization so that we assume the coefficients in γ_1 to consist of the k_1 coefficients with largest magnitude and γ_2 of the k_2 next largest coefficients. This approximative approach is due to practical reasons: in addition to allowing fast computation, it also usually provides results where the restriction $\sigma_1^2 \geq \sigma_2^2 \geq \sigma_3^2$ holds.

This treatment is trivially generalized into arbitrary number of components by introducing a model with m Gaussian components. Following the same treatment as in the three-component model we obtain a criterion

$$\sum_{i=1}^{m} \left(\frac{k_i}{2} \ln \frac{1}{k_i} \sum_{j \in \gamma_i} c_j^2 + \frac{1}{2} \ln k_i \right) + L(\gamma_1, \ldots, \gamma_m) + m \log \log \frac{\sigma_{\max}^2}{\sigma_{\min}^2} + \text{const} , \tag{7}$$

where const refers to terms constant with respect to the index sets and m, and $\sigma_{\max}^2$ and $\sigma_{\min}^2$ are the hyperparameters for the maximum and minimum variance, respectively. The last two terms can be ignored if we wish to find the optimal m-component result. On the other hand, if we want to compare the results for two approaches with different number of components, for example $m_1 = 3$ and $m_2 = 4$, we cannot remove the term involving the hyperparameters as it affects the code length.

4 Experiments

We show example results for applying the three-component approach in two different test cases. The first test signal is a one-dimensional time series data set used in the European Symposium on Time Series Prediction 2007 (ESTSP07) prediction competition, available in [9]. In this signal we can see a varying baseline making a distinctive bump after time index 400, a periodic element and noise. The second test signal is a capillary electrophoresis signal resulting from DNA microsatellite genotyping application; for

more information on denoising electrophoresis signals see [10]. This very noisy test signal contains two important components: a varying baseline and a set of peaks.

The three-component approach results for the time series data are presented in Figures 1 and 2. Figure 1 shows the original signal and the three obtained components. Figure 2 shows the sum of the two informative components (the trend and the periodic component) plotted over the original data points. As can be seen the proposed three-component approach is able to separate well the components of interest in the given signal.

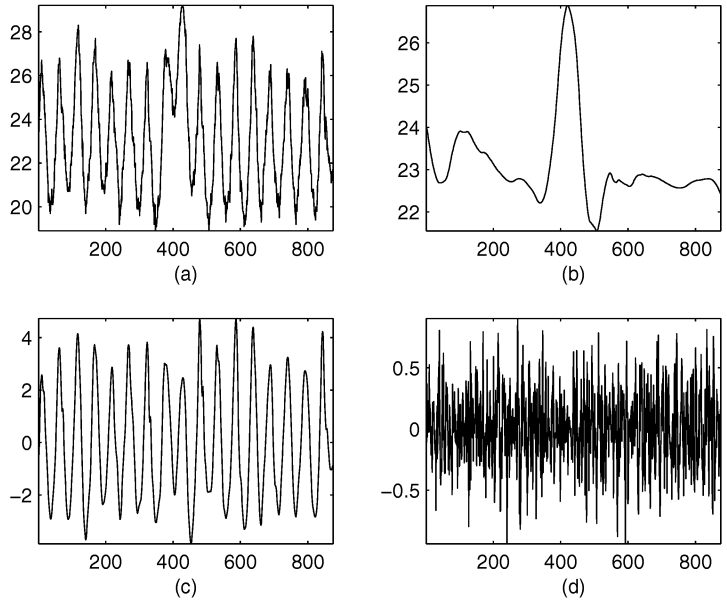


Figure 1: *The results for the ESTSP07 prediction competition data. The wavelet basis is the Daubechies 'db5', and the DWT has $N = 6$ levels. (a) The original signal; (b) the first component, which is seen to describe the varying trend in the signal; (c) the second component, which describes the periodic element; (d) the third component describing noise.*

The results for the capillary electrophoresis data are presented in Figures 3 and 4. Figure 3 shows the original signal and the three obtained components. Figure 4 presents the denoising result (two informative components summed) plotted over the original data. For comparison, the original MDL denoising result is also shown. It can be seen that the three-component approach improves the results significantly; the proposed three-component method is able to retain more of the important peaks, and it also retains the peak heights better than the original denoising approach.

5 Discussion

Here we have proposed a multicomponent MDL approach for wavelet denoising to be used in applications where the separation of signal components with different characteristics is needed. Although the

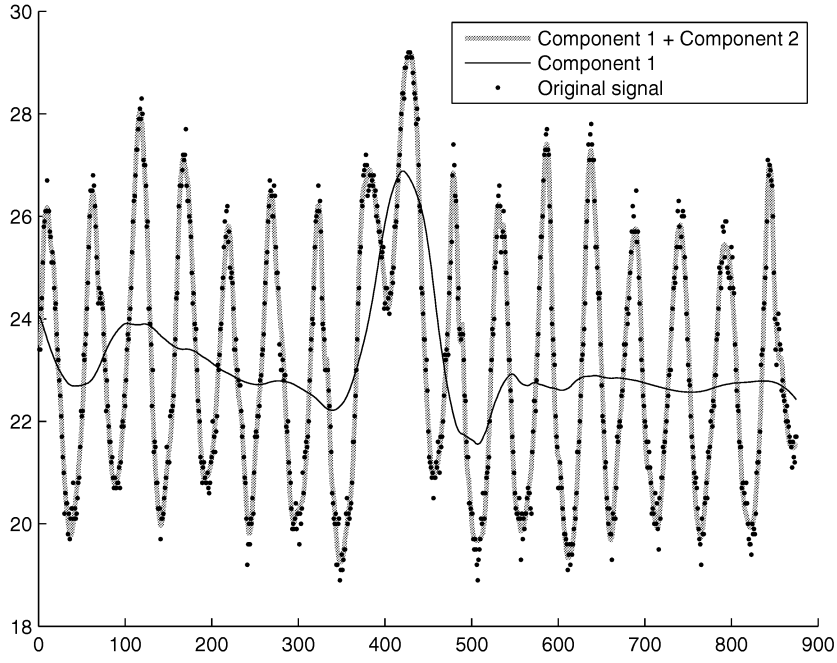


Figure 2: *The results for the ESTSP07 prediction competition data. The sum of the two informative components (Fig. 1 (b) and (c)) plotted over the original noisy data. The first component is also plotted alone to show the baseline.*

research in multicomponent MDL approach is at its first stages, the results are already promising. Modeling with several non-noise components can be useful in different ways: in some cases it may lead to better denoising results while in other cases the separation of the informative components is the more interesting result. For instance, in some applications the signal baseline removal may be critical in further analysis, which means that one is interested in using the resulting signal consisting of the other two components.

At the moment the criterion approximating the NML code length is calculated by a brute force approach, which is too slow for large signals such as images. Therefore, instead of an exhaustive optimization over the index sets the computations are performed so that the coefficients in the first component indexed by γ_1 are assumed to consist of the k_1 coefficients with largest magnitude, the coefficients in γ_2 of the k_2 next largest coefficients, and so on. In order to overcome these computational limits for example greedy approximations can be considered. Also, one way of improving the approximation could be to include information about the levels and subbands of the wavelet transform to the minimization problem.

In general, the properties of the NML based multicomponent approach are not yet known very well. However, it is known that the approach proposed here includes the same hyperparameters as the MDL denoising approach. These hyperparameters make comparison between models with different number of components difficult. The recently proposed conditional normalized maximum likelihood (CNML) [11] (also known as the sequential NML) approach could be potentially very useful also for multicomponent denoising, as it does not require hyperparameters. Furthermore, incorporating other density models than

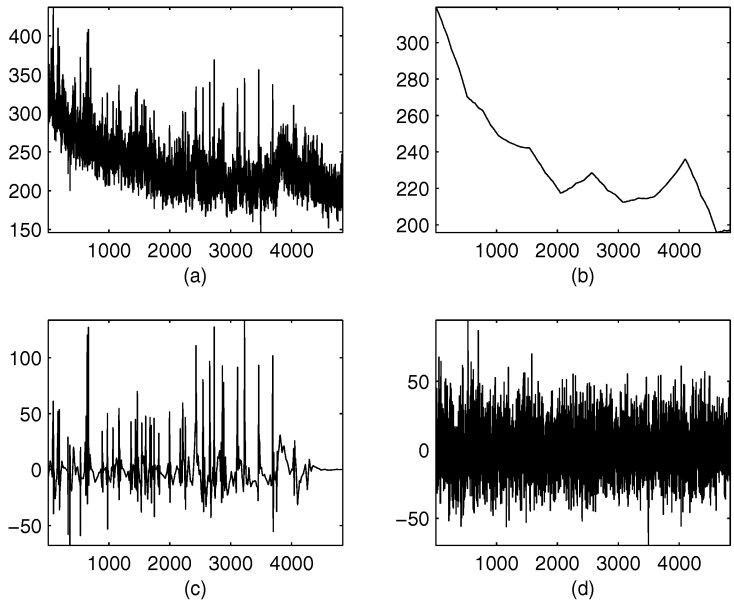


Figure 3: *The results for the capillary electrophoresis data. The wavelet basis is the Symmlet 'sym3', and the DWT has $N = 9$ levels. (a) The original signal; (b) the first component, which is seen to describe varying baseline; (c) the second component describing the peaks; (d) the third component describing noise.*

Gaussians should be easier in the CNML framework, so that more realistic models could be used.

Acknowledgements

This work was supported in part by the Graduate School in Computational Methods of Information Technology at Helsinki University of Technology, the Finnish Technology Agency under project KUKOT, and the Center of Excellence Program of the Academy of Finland.

References

[1] I. Fodor and C. Kamath, "Denoising through wavelet shrinkage: an empirical study," *Journal of Electronic Imaging*, vol. 12, no. 1, pp. 151–160, 2003.

[2] N. Saito, "Simultaneous noise suppression and signal compression using a library of orthonormal bases and the minimum description length criterion," in *Wavelets in Geophysics*, E. Foufoula-Georgiu and P. Kumar, Eds. New York: Academic, 1994, pp. 299–324.

[3] H. Krim and I. Schick, "Minimax description length for signal denoising and optimized representation," *IEEE Transactions on Information Theory*, vol. 45, no. 3, pp. 898–908, April 1999.

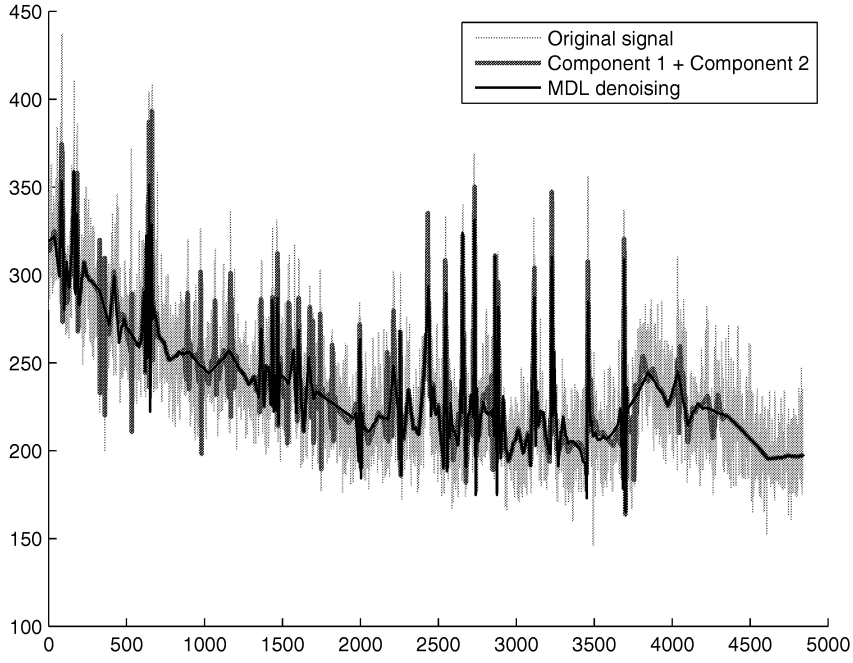


Figure 4: *The results for the capillary electrophoresis data. The results from the original MDL denoising and the proposed three-component MDL denoising are plotted over the original data.*

[4] J. Rissanen, "MDL denoising," *IEEE Transactions on Information Theory*, vol. 46, no. 7, pp. 2537–2543, 2000.

[5] M. Hansen and B. Yu, "Wavelet thresholding via MDL for natural images," *IEEE Transactions on Information Theory*, vol. 46, no. 5, pp. 1778–1788, August 2000.

[6] T. Roos, P. Myllymäki, and H. Tirri, "On the behavior of MDL denoising," in *AISTATS05*, R. G. Cowell and Z. Ghahramani, Eds. Society for Artificial Intelligence and Statistics, 2005, pp. 309–316.

[7] T. Roos, P. Myllymäki, and J. Rissanen, "MDL denoising revisited," *Submitted for publication*, 2006, preprint available at: http://www.arxiv.org/abs/cs.IT/0609138.

[8] T. Roos, "Statistical and information-theoretic methods for data-analysis," Ph.D. dissertation, University of Helsinki, 2007.

[9] "European symposium on time series prediction (ESTSP 2007) prediction competition dataset," Available at: http://www.estsp.org/files/competition_data.txt.

[10] J. Ojanen, T. Miettinen, J. Heikkonen, and J. Rissanen, "Robust denoising of electrophoresis and mass spectrometry signals with minimum description length principle," *FEBS Letters*, vol. 570, no. 1, pp. 107–113, July 2004.

[11] J. Rissanen and T. Roos, "Conditional NML universal models," *Information Theory and Applications Workshop, 2007*, pp. 337–341, Jan. 29 2007-Feb. 2 2007.

Context Adaptive Coding of Bi-level Images

dedicated to Jorma Rissanen on his 75th birthday

Søren Forchhammer
Technical University of Denmark
Email: sf@com.dtu.dk.

Abstract

With the advent of sequential arithmetic coding, the focus of highly efficient lossless data compression is placed on modelling the data. Rissanen's Algorithm Context provided an elegant solution to universal coding with optimal convergence rate. Context based arithmetic coding laid the grounds for the modern paradigm of data compression based on a modelling and a coding stage. One advantage of contexts is their flexibility, e.g. choosing a two-dimensional (2-D) context facilitates efficient image coding. The area of image coding has greatly been influenced by context adaptive coding, applied e.g. in the lossless JBIG bi-level image coding standard, and in the entropy coding of contemporary lossless and lossy image and video coding standards and schemes.

The theoretical work and analysis of universal context based coding has addressed sequences of data and finite memory models as Markov chains and sources. This paper discusses relations between context based coding of images and the context formation in some image models. Image models include Markov random fields (MRF), which have a non-causal description, and the special case of Pickard random fields, which are causal. These fields represent generalizations to 2-D of a finite memory source. Further developments of causal image models, e.g. to approximate MRF, lead to considering hidden states in the context formation. These causal image models provides image coding models and they are here related to context based image coding. The entropy of the image models is also considered.

Finally it is outlined how the techniques by duality may play a role in 2-D constrained coding for high density storage by switching the roles of encoding and decoding.

I. Introduction

Context based techniques for sequentially assigning conditional probabilities to elements of a data set is a very powerful tool. In this paper, context techniques as introduced in source coding will be treated in the light of lossless image coding. While the techniques are general, for practical image coding applications, they have mostly been used in their direct form for coding bi-level images. Combined with (pre-)processing of the data, the concepts are widely used in the entropy coding part of efficient image and even video coding schemes of today. The treatment will focus on the direct application of context based approaches to assigning conditional probabilities in images with the implicit assumption that these techniques may also be applied in the entropy coding of other image and video coding schemes.

By the minimum description length (MDL) principle the techniques may also be used more widely in modelling image data.

A. Lossless source coding

With the advent of sequential arithmetic coding the focus of highly efficient lossless data compression is placed on modelling the data. A very general and efficient approach is the use of contexts, which are given by a mapping of the causal data. For (bi-)level image coding, the mapping may simply be defined by a template in form of a selection of causal pixels.

Real world data including images are generally not generated according to any specific mathematical model. Thus the issue of determining a context function and its parameters is raised. A (coding) model class may be defined. In two part coding schemes, the model parameters are sent in a header as a preamble and thereafter the coding of the sequence based on these parameters. With this approach the parameters explicitly constitute a cost. In adaptive coding schemes, the parameters within the model class are learned adaptively. Rissanen's Algorithm Context provided an elegant solution of adaptively determining both the model within the class of tree-structured contexts functions and the parameters of the model. Algorithm Context provided a universal coding solution with optimal convergence rate. The universality applies to finite memory sources.

B. Image coding

The area of image coding has been greatly influenced by context adaptive coding, applied e.g. in contemporary lossless and lossy image image coding, as well as in video coding schemes. Context based arithmetic coding is directly applied in the international bi-level image coding standards, JBIG [8] and JBIG2, where the context function is given by a template of causal pixels. For lossless coding of (gray-level) images, the international standard JPEG-LS applies prediction to reduce the parameter space followed by context based coding. In JPEG-LS part 2 the actual coding may be based on arithmetic coding. For lossy (and lossless) coding of gray-level images, JPEG2000 applies a wavelet transform to the images as an initial decorrelation of the data. The wavelet coefficients are coded in bit-planes using context based coding. Also in the most recent video coding standard, MPEG4 part 10/H.264, context based arithmetic coding may be applied in the entropy coding for high-performance. The use of contexts has enabled flexible and efficient application to image coding. We shall consider issues of context based coding starting with template coding for (bi-level) images as well as the relations to image modelling. The application considered is compression, which for context based coding is based on conditional probabilities. Thus prediction and other estimation problems are closely related. The minimum description length principle formalizes some of the relations.

In Section II, lossless image coding and Algorithm Context are outlined. In Section III, Markov random fields and Pickard random fields (PRF) are presented. One approach to determining the parameters of a PRF is also presented. In Section IV, image models with hidden states are presented. Results from applying the image models to bi-level image coding is presented in Section V. A short remark is made on applying the techniques to coding for 2-D constrained coding in Section VI.

II. Lossless image coding

Arithmetic coding may take a sequence of symbols, $y_1^T = y^T = y_1, \ldots, y_T$, and corresponding probability assignments, $P(y_t|y^{t-1})$ and code the symbols, such that the total code length is very close to $-\sum_t \log P(y_t|y^{t-1})$.

This has lead to the modern paradigm of data compression separating the coding into a modelling step to attain $P(y_t|y^{t-1})$ and a coding step based on arithmetic coding. We shall pursue this and consider the modelling step for image data assuming arithmetic coding is used subsequently.

Let y_{ij} denote the picture element at (i, j) drawn from a finite alphabet, $\mathcal{A}$. For compression, the images are processed sequentially in a predefined scanning order. Let y_t denote element t in the sequential representation of the image and $y^T = y_1^T$ the whole image. The pixel position (i, j) is mapped to a sequence index t by a traversal of the image. We shall use the conventional row-by-row raster scan.

For coding the image, a probability is assigned by $P(y^T) = \prod P(y_t|y^{t-1})$, where $P(y_t|y^{t-1})$ is expressed by $P(y_t|y^{t-1}) = P(y_t|F(y^{t-1}))$. The function $F(y^{t-1})$ defines the context in terms of a mapping of the causal elements. We shall mainly consider context functions defined by a subset of elements of the past with given position relative to the current element, y_t at position (i, j),

but also context functions defined by calculations over past values possibly including hidden states are considered. Arithmetic coding is applied sequentially to the conditional probabilities. Thus the definition of the context function $F(y^{t-1})$ becomes crucial in the design of a lossless compression algorithm.

In context based adaptive coding the probability assignment $P(y_t|y^{t-1})$ is updated sequentially. Initially a probability assignment based on occurrence counts in each context is considered. Let $r_t = F(y^t)$ be the context value at t and $n_t(a|r_t)$ be the number of times that value a appears in context r_t the sequence y^{t-1}. The probability assignment is

$$P(a|r_t) = \frac{n_t(a|r_t) + 1/2}{\sum_{y\in\mathcal{A}} n_t(y|r_t) + |\mathcal{A}|/2}, \tag{1}$$

where $|\mathcal{A}|$ is the size of the alphabet.

The ideal code length of occurrences in context, s is given by

$$L(s) = -\sum_{t|r_t=s} \log P(y_t|r_t = s). \tag{2}$$

The ideal code length is given by summing $L(s)$ over all contexts.

A. Context based coding of binary images

The approach to compression outlined above may be applied to coding binary images achieving very efficient compression. In [9] sequential arithmetic coding based on conditional probabilities was applied to bi-level images. The contexts were defined by a template. For a sequence of variables $Y_1^t = Y_1, \ldots, Y_t$, the context is expressed by a subset of K of the variables, $(Y_{t-t_1}, Y_{t-t_2}, \ldots, Y_{t-t_K})$. The arithmetic coding was the so-called Q-coder avoiding multiplications by approximations. This approach was refined in the ISO JBIG standards. First JBIG defined a 10 pixel template and later JBIG2 followed up allowing different templates sizes up to 16 pixels of which 4 may be placed adaptively (in the causal part). The arithmetic coders used are further developments of the Q-coder. The probabilities assigned for coding are based on (1), but implemented by a finite state machine for higher speed.

Given an image, the decision of template size and pixel location may be determined by searching using multiple passes of the data if possible. A simple greedy search often provides good results, but also more advanced searching may be applied. For domain specific applications a predefined template will often provide good results as reflected by the standards.

B. Algorithm Context

In template based coding as outlined above, the context function is fixed, but the probability values are learned adaptively (1). To achieve universal coding the context function must also be learned from the data.

Algorithm Context [14] provided an elegant solution to this. It replaced the template by a dynamic context tree, where the context length is also decided adaptively. Given an (unbounded) ordering of context elements, $(Y_{t-t_1}, Y_{t-t_2}, \cdots)$ it consists of two steps: 1) A tree building step over the context elements defining (and limiting) the potential contexts and 2) for each new element, y_t, a node selection rule defining the context value, $F(y^{t-1})$ which in turn by (1) defines the probability assignment. Each node maintains occurrence counts of the context it represents.

The tree is updated after each new element, y_t, by following the path in the context tree given by the previous elements, $(y_{t-t_1}, y_{t-t_2}, \cdots)$, until a leaf node of the tree is reached. If the counts of the leaf node becomes at least two, a new node is added and the occurrence counts initialized, i.e the

counts defining the probability assignment reflects occurrences in the context in the past starting from the creation of the node.

The selection principle is given by choosing the up till now, in some sense, best node of the path. Efficient selection may be based on father-son comparisons. Different variations have been presented, e.g. [14], [15],[17]. Here we state the MDL based selection rule of [15] for a binary alphabet. Let s denote the father node and $s0$ and $s1$ the two son nodes. Calculate the ideal code lengths (2) of these contexts. Now pick the father node s over the sons if $L(s) \leq L(s0) + L(s1)$. (Note the code lengths are easily updated (2) and the decision may be based on $L(s) - L(s0) - L(s1)$.) A simple solution is given by evaluating, starting from the root until a father node is better than the sons [15]. For universal coding the full path should be evaluated.

In [14] the deepest node on the path, which is better than the sons is selected. In [17] the deepest father node which is better than the son on the path is selected. In both of these solutions where the full path is searched, there is a restriction to (the growth rate of) the context length (smaller than $c \log t$, where c is a constant), in order to ensure universality.

A machine defined by the nodes of a complete tree is called a tree machine. Such a machine may be defined by the nodes selected by Algorithm Context [17]. Algorithm Context is universal in the class of finite memory sources or tree machine sources. It provides optimal convergence both in the mean sense and almost surely [17]: Compared to any minimal complete tree, $\mathcal{T}$, with K leaves, and probabilities, $P(a|s) > 0$, the finite-memory source defined by Algorithm Context [17] asymptotically achieves a code length for y^n within $\frac{K}{2}(|\mathcal{A}| - 1) \log n + O(1/n)$ of the code length of the tree, $-\log P_{\mathcal{T}}(Y^T)$. Thus this term may be seen as (a bound on) the model cost paid for learning the model parameters. The proof is based on showing that the probability of over- or under-estimating the context length tends to zero and does so fast enough. Thus asymptotically the algorithm also identifies the correct context within the class almost surely.

In [17], the selection rule is formulated by entropies rather than code lengths as in the MDL formulation of a decision rule in [15]. The selection rule and the proof of universality [17] was based on deriving stationary probabilities from a Markov chain derived from the description.

The tree machine of Algorithm Context differs from finite state machines in that the next context is not necessarily given by the last context and the last symbol. As pointed out this distinction is important for image data [17]. For image data, Algorithm Context may be applied using say a raster-scan traversal of the image. When defining the context, an ordering of the causal data is applied. Finite context image models are considered in the next section. As for template based coding, Algorithm Context may readily capture two-dimensional (2-D) dependencies in images by defining the context in 2-D while performing sequential coding. This also generalizes to contexts in higher dimensions or the inclusion of any relevant side information.

Empirical results have indicated that to achieve good results on finite data sets, e.g. image date, the selection rule should have some bias towards long contexts. This was addressed in [17] for non-binary data. In [11] applying Algorithm Context to bi-level images was considered. In order to combine speed and high compression efficiency, the father-son comparison was extended to compare a father with the descendants in the path. This gave increased performance, using root to leaf traversal, but still allowing quick access to long contexts without being stopped by pixels on the path with little additional information. For analysis, a maximum context length was imposed as part of the selection rule.

C. Gray level images

In principle, template based coding and Algorithm Context may directly be applied to natural images as 8 bits per pixel or color images. In practice the alphabet size leads to a parameter space, which is too large for the desired context sizes in relation to typical image dimensions.

To alleviate the context dilution problem, two measures may be taken for gray level images, namely prediction, to initially decorrelate the image data, and context quantization [18]. While these two steps

are special cases of conditioning on the full context, the parameter space is greatly reduced, illustrating the potential importance of restricting the model class, for a more efficient parameterization in relation to a given data set.

The idea of Algorithm Context was applied to competing models in a tree-structured manner which is a generalization of the basic algorithm context [18]. These issues are not pursued further in this text.

III. Image models and adaptive code lengths

The concept of contexts has as noted enabled efficient application of context based techniques to image coding. In this section we consider, Markov random fields and the special case of Pickard random fields. These fields represent 2-D generalizations of a finite memory source. This leads to considering hidden states as part of the context formation of the (bi-level) image compression modelling stage. The non-causal models and models based on hidden states are related to the image coding schemes and expressions for the adaptive code length are given. For stationary sources entropy expressions are given.

In context-based image coding, the contexts provide the flexibility of capturing two-dimensional structure while coding the elements sequentially. As mentioned the per symbol code length of Algorithm Context asymptotically converges to the entropy for finite memory sources. Also template based coding implies a notion of finite memory for image. Methods to infer the conditional probabilities are discussed. Given probability estimates, expressions for the adaptive code length may be applied to single images regardless of whether the images belong to a model class or not.

A. Extending a local measure - finite contexts

Consider a template with three pixels, A, B, and C used for conditioning the current pixel D,

$$\begin{array}{cc} A & B \\ C & D \end{array}$$

We may code an image based on the conditional probability, $P(D|ABC)$ and given a probability distribution calculate an ideal code length for a given image. It is also possible to generate an image according to $P(D|ABC)$. Two questions arise: what characterizes image models where $P(D|ABC)$ describes the conditional probability of D given the past? Can we define $H(D|ABC)$ consistent with a stationary (shift invariant) distribution of $ABCD$ over the field?

Given the distribution on the 2×2 lattice, $(ABCD)$, one can extend this to a measure $\mu_{n\times m}$ on an $n \times m$ lattice $x = (x_{ij})$ in the following manner: First the element x_{11} is drawn according to the distribution (A). Then the first row $x_{12} \dots x_{1m}$ is drawn according to the conditional distribution $(B|A)$ one element at a time. Thereafter the first column $x_{21} \dots x_{n1}$ is drawn according to $(C|A)$ one element at a time. x_{22} can then be drawn using $(D|ABC)$. Proceeding in this manner one has (using shorthand notation for probabilities given by the argument):

$$\begin{aligned} \mu_{n\times m}(x) =& P(x_{11}) \cdot \Pi_{j=2}^{m} P(x_{1j}|x_{1,j-1}) \\ & \cdot \Pi_{i=2}^{n} P(x_{i1}|x_{i-1,1}) \\ & \cdot \Pi_{i=2}^{n} \Pi_{j=2}^{m} P(x_{ij}|x_{i-1,j-1}, x_{i-1,j} x_{i,j-1}). \end{aligned} \tag{3}$$

The union of the elements of the first row and the first column is called the boundary. The extended measure is *stationary* if the joint distribution of $(ABCD)$ does not depend on which 2×2 rectangle within the $n \times m$ rectangle we regard.

The context of element D is given by A, B, and C. Decomposed expressions as (3) with larger contexts (leading also to wider boundaries) are straightforward and the implicit model of template based coding, though the boundary need not be Markov chains. The question of stationarity or rather

Fig. 1. Markov random field neighborhoods divided in the causal and non-causal parts of a row-by-row scan of the field.

shift-invariance is naturally raised in relation to the probability of a given context. If we consider an image as a segment of a larger image, it would be natural to require that the context probabilities at a given point will not depend on where the segment starts nor the scanning policy, i.e the order of traversing the pixels. Shift-invariance is consistent with this requirement. Below this is considered for finite contexts.

B. Markov random fields

Markov random fields (MRF) may be defined by the probability of a pixel given the surrounding pixels called the cliques or neighborhood (Fig. 1). This leads to a natural extension of Markov chains to two dimensions. Consider an interior set of elements and a boundary given by the clique, such that the pixels of the cliques of any interior pixel is either an interior pixel or part of the boundary. An interior set with this property is independent of the exterior set in the sense that the conditional probability of the interior set conditioned on the complement is given by conditioning on the boundary. While the generalization given above is simple to state, the Markov random fields do not yield a simple causal description as desired in image coding. For a row-by-row traversal of an image, the context becomes infinite (or for an image of finite width, it will be given by the boundary of the past) given by one or more rows across the image depending on the neighborhood. Consider e.g. the neighborhood given by the four direct neighbors (Fig. 1). The infinite context, of a row-by-row scan, is given by a one pixel wide boundary across the plane, namely the last causal pixel in each column. This boundary will separate the past from the future. For both neighborhoods of Fig. 1, one row of the plane will separate the the field above this row from the field below the row.

The Markov random fields are max-entropic given the probabilities conditioning a pixel on the (non-causal) neighborhood. A drawback, in relation to analyzing image coding, is that it is not possible in general to compute the so-called partitioning function, which is a normalizing function equivalent to the entropy.

An exception are the Pickard random fields, which are considered next. We shall also consider ways to approximate the MRF in the next section. The starting point is to describe a number of rows by a Markov chain or a function of a Markov chain.

C. Pickard random fields

A special case is given by the Pickard random fields [12]. Let A, B, C, D be random variables over $\mathcal{A}$ in a 2×2 rectangle as above. A set of conditions on the probability distribution $(ABCD)$ shall be presented ensuring Markovian and stationary properties of extensions to measures (3) on rectangles of arbitrary size.

Let X, Y, Z be random variables and let $X \perp Y \mid Z$ denote that X and Y are independent given Z. The independence conditions $B \perp C \mid A$ and $B \perp C \mid D$ shall be assumed for the models considered in this section.

The model is completely specified by the probability distribution on (A) as well as the three conditional probability distributions $(B|A)$, $(C|A)$ and $(D|ABC)$.

The probabilities of $(ABCD)$ are expressed by

$$P(ABCD) = P(D|ABC)P(ABC) \tag{4}$$

and due to the independence condition $B \perp C \mid A$,

$$P(ABC) = P(B|A)P(C|A)P(A). \tag{5}$$

In order for the measure $\mu_{2\times 2}$ to be stationary on the 2×2 lattice, it is sufficient (and necessary) that the distributions on (AB) and (CD) be identical and the distributions on columns (AC) and (BD) be identical. The stationarity obviously implies that, the stationary distributions for $(B|A)$ and $(C|A)$ must be identical.

The following Theorem due to Pickard [12] gives a sufficient condition on $(ABCD)$ for the extended measures (3-4) to be stationary.

Theorem 3.1: Let $\mu_{2\times 2}$ be a stationary measure induced by $(ABCD)$ satisfying $B \perp C \mid A$. If $B \perp C \mid D$ then the extended measure $\mu_{n\times m}$ based on (3-4) is Markovian and stationary for any $n, m \geq 2$.

Theorem 3.1 provides sufficient conditions for the measure $\mu_{2\times 2}$ to be extended to a stationary measure. Since $B \perp C \mid D \Leftrightarrow P(BCD) = P(D)P(B|D)P(C|D)$, assuming stationarity, the right hand terms are readily derived from (ABC) (5).

The independence property (5) of the PRF leads to the finite memory Markovian property: $P(y_t = d|y^{t-1}) = P(y_t = d|abc)$. Further it follows [5] that

Theorem 3.2: The entropy per symbol of a stationary measure $\mu_{n\times m}$ defined by Theorem 3.1 is bounded by

$$H(\mu_F) \geq H(D|ABC). \tag{6}$$

In the limit $(n, m \to \infty)$ $H(\mu_F) = H(D|ABC)$ as the difference for the PRF is restricted to the one pixel wide upper and left boundary. $H(D|ABC)$ is readily defined and by the stationarity of the PRF, this is the expected contributions of all pixels of the interior.

A PRF has the property that any number of rows form a Markov chain [12]. For Pickard random fields, Algorithm Context provides a universal code achieving the PRF entropy asymptotically, assuming the nearest pixels, A, B and C are in the context set.

1) Consistency of PRF parameters: For context-based coding, the conditional probability, $P(D|ABC)$, is the main concern. Besides the conditional probability, the distribution on the boundary is given by the MCs on the rows and columns (5). In 1-D the stationary distribution of a MC may be derived from the transition probabilities. This is not the case in 2-D, e.g. for the PRF. Now we consider the consistency of the boundary distribution and the conditional probabilities of the interior. Given the conditional probabilities, $P(D|ABC)$, we would be interested in deriving a boundary distribution resulting in a stationary distribution. As this is not tractable, the boundary distribution is taken as the starting point.

Consider a stationary distribution on the four basic PRF variables, A, B, C, D. It is a necessary condition for (ABC) and (BCD) to be marginal distributions of $(ABCD)$, that their marginal distributions on (BC) be identical. This may be expressed by

$$\sum_{a\in\mathcal{A}} P(A = a, bc) = \sum_{d\in\mathcal{A}} P(bc, D = d), \forall (b, c) \in \mathcal{A}^2. \tag{7}$$

This condition may also be expressed in matrix form. The stationarity condition implies that the horizontal transition probabilities are identical, $P(A|B) = P(C|D)$, as well as the vertical transition probabilities, $P(C|A) = P(D|B)$. Let $\mathbf{R}$ and $\mathbf{S}$ denote the corresponding transition probability matrices, respectively.

The independence conditions, imply that the triples CAB and CDB are given by a horizontal and a vertical transition each but in reverse order. Thus (7) may be expressed by requiring the matrices to commute,

$$\mathbf{RS} = \mathbf{SR}, \tag{8}$$

which is the only way to ensure consistency of the PRF [4].

A simple solution starting from two commuting transition matrices is to derive the conditional probability from $\mathbf{RS}$, which leads to $P(D|BC)$, i.e. a two pixel template. For a fixed choice of $\mathbf{S}$ and $\mathbf{R}$, which commute, this is also the max-entropic choice as including conditioning on A besides BC will not increase the entropy once the distribution on ABC is fixed.

Having identical Markov chains horizontally and vertically, ($\mathbf{S} = \mathbf{R}$), these matrices obviously commute and a two-pixel template PRF may be derived from any (irreducible) Markov chain. We will now return to the full PRF, which has more modelling power.

2) Iterative techniques for stationary solutions: If a stationary boundary distribution, $P(ABC)$ (5) and a conditional distribution $P(D|ABC)$ is given, but the combination $Q(ABCD) = P(ABC)P(D|ABC)$ is not an PRF, the following PRF approximation may be derived.

The PRF stationarity conditions and the independence conditions provide sufficient conditions for a probability distribution $(ABCD)$ described by (4-5) to satisfy Theorem 3.1. How to determine parameters of the PRF model is considered next.

3) Iterative scaling: Given a boundary description in terms of the distribution (ABC) satisfying (8), iterative scaling [2] may be used to find conditional probabilities $P(D|ABC)$.

For each configuration $B = b, C = c$, consider $P(AD|bc)$. The distribution (ABC) determines $P(A|bc)$,

$$\begin{gathered}\alpha_i \equiv P(A = i|bc) = \\ P(A = i, bc)/\textstyle\sum_{j\in\mathcal{A}} P(A = j, bc), i \in \mathcal{A}.\end{gathered} \tag{9}$$

The distribution (ABC), the stationarity and the independence $B \perp C \mid D$ determines (BCD), which in turn determines $P(D|bc)$,

$$\begin{gathered}\beta_j \equiv P(D = j|bc) = \\ P(bc, D = j)/\textstyle\sum_{i\in\mathcal{A}} P(bc, D = i), j \in \mathcal{A}.\end{gathered} \tag{10}$$

The probabilities of $P(AD|bc)$, must satisfy the linear relations due to (9-10):

$$\sum_{j\in\mathcal{A}} P(A = i, D = j|bc) = \alpha_i, i \in \mathcal{A}. \tag{11}$$

$$\sum_{i\in\mathcal{A}} P(A = i, D = j|bc) = \beta_j, j \in \mathcal{A}. \tag{12}$$

Thus we seek a solution in the intersection, $\mathcal{L}$, of the two linear families defined by (11) and (12), respectively, each determined by a partition (which may be described as row- and column-sums of a matrix given by elements $p_{bc}(i, j) = P(A = i, D = j|bc)$).

Iterative scaling [2] may take an initial distribution, $Q(x)$ and a class of distributions as $\mathcal{L}$ and find a distribution, P^* which minimizes the divergence, $D(P^*||Q)$ for $P^* \in \mathcal{L}$, where the divergence is given by $D(P||Q) \equiv \sum P(x)\log(P(x)/Q(x))$, if a distribution satisfying (11) and (12) and possibly joint constraints exists.

This leads to the scalings defining new probabilities, $p^*_{bc}(i, j)$, within each of the families,

$$p^*_{bc}(i, j) = cp_{bc}(i, j), c = \alpha_i/\sum_{i\in\mathcal{A}} p_{bc}(i, j), j \in \mathcal{A}. \tag{13}$$

$$p^*_{bc}(i,j) = dp_{bc}(i,j), d = \beta_j / \sum_{j \in \mathcal{A}} p_{bc}(i,j), i \in \mathcal{A}. \quad (14)$$

The iterative scaling will converge to an approximative distribution to the initial distribution $Q(ABCD)$, in the case that this does not satisfy the PRF conditions.

In general we may have two distinct Markov chains defining the boundaries, $\mathbf{R}$ and $\mathbf{S}$. In case that given $B = b, C = c$, A and D are not independent (for constraints see the last section), the dependencies between A and D are determined in a consistent way. Given $Q(ABCD)$, the conditional distribution $P(AD|BC)$ may be calculated, such that in combination with $P(ABC)$ (5), the so derived $(ABCD)$ defines a PRF.

Theorem 3.3: It is a necessary condition for two irreducible Markov chains, with transition matrices $\mathbf{R}$ and $\mathbf{S}$, to form the boundary of a Pickard random field, that the matrices commute (8). If so iterative scaling of $P(AD|BC)$ determines whether a PRF with boundary given by the irreducible $\mathbf{R}$ and $\mathbf{S}$ exists and if so determines the conditional probabilities $P(D|ABC)$, which minimizes the divergence between $P(AD|BC)$ and $Q(AD|BC)$.

If the goal is to approximate a given $Q(ABCD)$ and if the description of valid boundary Markov chains are given by parameters, a search over these parameters may be applied to these to minimize $D(P(ABCD)||Q(ABCD))$ determining the optimal $P(AD|BC)$ by the iterative scaling.

For Pickard random fields (and other finite context 2-D models), Algorithm Context can determine the context pixels and the set of conditional probabilities, $P(D|ABC)$.

D. Block Pickard random fields

The PRF introduced above is restricted to the three pixel template, ABC, but may be defined on any finite alphabet. A block based PRF was considered in [4],[5], where (rectangular) blocks of pixels were treated by alphabet extension. Thus a larger context is defined. How to control that this description is shift-invariant on the original pixels (and not only on the blocks) is a question though.

IV. Image models with hidden states

In the previous section, we considered PRFs, which have the properties that the context is finite, the field is stationary and the rows form Markov chains. One may say that the PRF is defined by the stationary solution to a two-row Markov chain (the distribution on 2×2 elements) subject to certain conditions.

To extend this class, we consider the case where the rows are still described by Markov chains or more generally by functions of Markov chains, but the model context need not be finite.

First consider a stationary Markov chain defined on two-rows and further assume that the distribution is symmetric in the two rows. Thus the distribution in each row is identical and the construction may be a repeated adding a new row such that and two rows are drawn from the original Markov chain distribution. If the distribution does not possess the independence property, $B \perp C \mid A$, the next element in a row may be dependent on the the rest of the preceding row. This dependency may be treated by considering the non-causal elements of the current row as hidden variables. This will be treated as a special case of the more general frame work with hidden states introduced below.

To define the models, hidden states are introduced. The common grounds of the PRF and the other models is that the current row will be conditioned on one or more previous rows. Thus these rows separate the past and the future as it was the case for the MRF.

The aim is to develop models and thereby probability assignments for coding schemes of reasonable complexity (linear in the data length) which may use hidden states with unknown parameters. As part of this we consider models with a finite parameter set, whereas the model context is infinite.

Deriving a stationary image model from a description on N rows enables expressing the per symbol entropy, H, by

$$H = H_N - H_{N-1}, \tag{15}$$

where H_{N-1} and H_N are the per column entropies of $N-1$ and N rows, respectively. We shall consider how to express these entropies.

We start by introducing two parameterizations of a function of a Markov chain, which are later developed into two distinct image models. The approach may be described as combining hidden Markov models and contexts to introduce hidden states in image modelling. Let y denote the observable output variables and x the hidden part.

For a function of a Markov chain [1], we have, for k finite

$$P(y^{t+1}, x^{t+1}) = P(y^t, x^t)P(y_{t+1}, x_{t+1}|y_{t-k}^t, x_{t-k}^t), \tag{16}$$

for $t = k+1, ..., T-1$.

We define two context functions $s_t = F_S(y^t)$ and $r_t = F_R(y^t)$ which are used in the Partially Hidden Markov Model (PHMM) [6]. A (PHMM) is defined by the axiom schema

$$P(y^{t+1}, x^{t+1}) = P(y^t, x^t)P(y_{t+1}|x_{t+1}, r_t)P(x_{t+1}|x_t, s_t), \tag{17}$$

for $t = 0, 1, ..., T-1$, where $x_0 = y_0$ and therefore also $r_0 = s_0$ are taken as the null string.

The Markovian property of the combined variables (x, y) means that knowing the hidden variable x_t along with y_t will separate the past and the future of the observed sequence motivating the forward variable,

$$\alpha_t(x_t) \equiv P(y^t, x_t). \tag{18}$$

Efficient implementation may be based on a trellis structure, where the hidden states are nodes of the trellis and the transition probabilities are conditioned on the context given by a mapping of the causal part of the output.

The implications of (16) and (17) will be presented below. Image models will be derived by applying the recursions row by row leading to two distinct image models.

1) Function of a Markov chain: For Markov random fields, let $N-1$ refer to the number rows that will separate the current row from the past. Thus the current row may be described conditional on the previous $N-1$ rows. (For an image which is a segment of a Markov random field this will apply assuming special treatment at the boundaries.)

Here we consider a joint description of N rows (possibly including hidden states) and derive the expression for the current row conditioned on the $N-1$ previous rows from this.

To pursue this view point, we may consider the observations to be vectors given by the N variables in a column. For simplicity the rows are indexed 1 to N. The N elements $y_{1j}, ...y_{Nj}$ in column j are denoted $\mathbf{y}_j$. The output sequence is $\mathbf{y}_1^j$. Likewise let $\mathbf{x}_j$ denote the hidden variables associated with $\mathbf{y}_j$, such that $\mathbf{y}_j, \mathbf{x}_j$ defines the value of a state of a Markov chain. A forward variable is introduced to capture the influence of the first j vectors,

$$\alpha_j(\mathbf{x}_j) \equiv P(\mathbf{y}^j, \mathbf{x}_j). \tag{19}$$

A backward pass is introduced to capture the influence of the elements after vector j. As the non-causal elements in the current row are still not seen, these will be treated as hidden, i.e. there are only $N-1$ elements in the observed output vector, denoted $\mathbf{y}'_j$ and the element of the current row is included in the hidden vector denoted $\mathbf{x}'_j$. Let W denote the width of the image. A backward conditional probability is defined by $\beta'_j(\mathbf{x}_j) \equiv P(\mathbf{y}_{j+1}^W|\mathbf{x}_j, \mathbf{y}_j)$. The conditional probabilities $\beta'_j(\mathbf{x}_j)$ satisfy the recursion

$$\beta'_j(\mathbf{x}_j) = \sum_{\mathbf{x}_{j+1}} P(\mathbf{x}_{j+1}, \mathbf{y}_{j+1}|\mathbf{x}_j, \mathbf{y}_j)\beta'_{j+1}(\mathbf{x}_j). \tag{20}$$

Combining the forward variable and the backward pass, at j the probability of the causal part of the N rows may be expressed by

$$P(\mathbf{y}_1^{j-1}, \mathbf{y}_j^{'W}) = \sum_{\mathbf{x}_{j-1}} \sum_{\mathbf{x}'_j} \alpha_{j-1}(\mathbf{x}_{j-1}) P(\mathbf{x}'_j, \mathbf{y}'_j | \mathbf{x}_{j-1}, \mathbf{y}_{j-1}) \beta'_j(\mathbf{x}'_j). \tag{21}$$

Taking the ratio of the probabilities (21) one step apart gives the probability of the next element given the causal past,

$$P(y_{Nj} | \mathbf{y}_1^{j-1}, \mathbf{y}_j^{'W})) = \frac{P(\mathbf{y}_1^{j-1}, \mathbf{y}_j^{'W})}{P(\mathbf{y}_1^{j-1}, \mathbf{y}_j^{'W})} \tag{22}$$

The model may be applied to an image row by row, conditioning on the $N-1$ previous rows.

If there are no hidden elements in the states i.e. $(\mathbf{y}_j, \mathbf{x}_j) = \mathbf{y}_j$, the model amounts to a Markov chain description on N consecutive rows and the only hidden part in $\beta'_j(\mathbf{x}_j)$ is the non-causal elements of the current row. In this case the parameters and N may be determined by Algorithm Context, but the coding may still be based on (22), where β' will provide a backward pass over the elements of the current row to be coded.

2) Partially Hidden Markov Models: Returning to the context version of the function of a Markov chain, it is seen that the PHMM axioms (17) immediately imply the recursion

$$\alpha_t(x_t) \equiv P(y^t, x_t) = \sum_{x^{t-1}} P(y^t, x^t) = P(y_t | x_t, r_{t-1}) \sum_{x_{t-1}} P(x_t | x_{t-1}, s_{t-1}) \alpha_{t-1}(x_{t-1}), \tag{23}$$

where $\alpha_0(x_0) = 1$.

This gives

$$P(y^t) = \sum_{x_t} \alpha_t(x_t) \tag{24}$$

and

$$P(y_{t+1} | y^t) = \sum_{x_{t+1}} P(y_{t+1} | x_{t+1}, r_t) w_t(x_{t+1} | s_t) \tag{25}$$

where

$$w_t(x_{t+1} | s_t) = \frac{\sum_{x_t} P(x_{t+1} | x_t, s_t) \alpha_t(x_t)}{\sum_{x_t} \alpha_t(x_t)}. \tag{26}$$

It may be noted that $P(y_{t+1}|y^t)$ is a convex mixture of output distributions over the hidden states.

The partially hidden Markov model is described by the parameter set $\lambda = (\pi, A, B)$, where π gives the probabilities of the initial hidden state, A the conditional hidden state transition probabilities, and B the conditional output probabilities for each of the states.

A context formulation of the PHMM is used, where the parameters of the partially hidden Markov model $\lambda = (\pi, A, B)$ consist of the following

$$\pi_i = P(x_1 = i) \tag{27}$$

$$a(i, j, k) = P(x_{t+1} = j | x_t = i, s_t = k) \tag{28}$$

$$b_m(i, l) = P(y_t = m | x_t = i, r_{t-1} = l). \tag{29}$$

Let $a_t(i, j)$ denote $a(i, j, l)$ where l is given by s_t and let $b_t(i, l)$ denote $b_m(i, l)$ where m is given by y_t.

Given the model parameters, λ and the observation sequence until t, y^t, we may sequentially calculate the probability of being in a specific hidden state x_i at t,

$$\alpha_1(i) = \pi_i b_1(i, r_0), \qquad 1 \le i \le N \tag{30}$$

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_t(i,j)\right] b_{t+1}(j,k), \quad 1 \le t \le T-1, \quad 1 \le j \le N \tag{31}$$

If the context, which may be called the seen part, for each state can assume only one fixed value, we have the formulation of a Hidden Markov Model (HMM) given in e.g. [13].

The ideal code length is given by $-\log[P(y^T|\lambda)] = -\sum \log[P(y_{t+1}|y^t, \lambda)]$. An actual coding may be performed by applying arithmetic coding to the sequence of conditional probabilities.

3) Reestimation of PHMM parameters: If the model parameters are not given, we have the problem of estimating the parameters, including those involving the hidden states. For a given data set and a given model order of a HMM, the Baum-Welch method iteratively converges towards a local maximum of $P(O|\lambda)$ over the parameters λ [10]. The model parameters are reestimated after each pass of the data set. The reestimation formulas were generalized to the PHMM [6] as specified below.

We define a backward variable, $\beta_t(x_t) = P(y_{t+1}^T|x_t, y^t)$, which may be perceived as representing the completion of the α forward pass for a given hidden state, x_t, at t. (This differs from the β' backward pass by having the same set of hidden and seen variables as for the α forward pass.)

The conditional probabilities $\beta_t(x_t)$ satisfy the recursion

$$\beta_t(x_t) = \sum_{x_{t+1}} P(y_{t+1}|x_{t+1}, r_t) P(x_{t+1}|x_t, s_t) \beta_{t+1}(x_{t+1}), \tag{32}$$

where $\beta_T((x_t)) \equiv 1$.

Given the entire observation sequence and the model parameters λ, the forward variable may be supplemented with a backward variable to obtain the probability of being in a given hidden state x_t at time t,

$$\gamma_t(i) \equiv P(x_t|y^T) = \frac{\alpha(x_t)\beta(x_t)}{\sum_{x_t} \alpha(x_t)\beta(x_t)}. \tag{33}$$

The induction formula determining the backward variable β becomes

$$\beta_T(i) = 1, \qquad 1 \le i \le N. \tag{34}$$

$$\beta_t(i) = \sum_{j=1}^{N} a_t(i,j) b_{t+1}(j,k) \beta_{t+1}(j), \qquad t = T-1, T-2, ..., 1, \quad 1 \le i \le N. \tag{35}$$

The reestimation formula of π_i is given by

$$\bar{\pi}_i = \gamma_1(i) \tag{36}$$

For $a(i,j,k)$ the reestimation formula is

$$\bar{a}(i,j,k) = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)\delta(s_t,k)}{\sum_{t=1}^{T-1} \gamma_t(i)\delta(s_t,k)}, \tag{37}$$

where

$$\xi_t(i,j) = \frac{\alpha_t(i) a_t(i,j) b_{t+1}(j,k) \beta_{t+1}(j)}{P(y^T|\lambda)}, \tag{38}$$

and $\delta(i,j)$ is 1 if $i = j$ and 0 otherwise. For $b_m(j,l)$, the reestimation formula is

$$\bar{b}_m(j,l) = \frac{\sum_{t=1}^{T} \gamma_t(j)\delta(r_{t-1},l)\delta(y_t,m)}{\sum_{t=1}^{T} \gamma_t(j)\delta(r_{t-1},l)}. \tag{39}$$

If the seen parts of the states can assume only one fixed value, we have the reestimation formulation for the HMM given in e.g. [13].

As for the HMM, the PHMM reestimation formulas ensures convergence to a local maximum of $P(y^T|\lambda)$ [6]. This may be used in either a two-pass or an adaptive coding scheme [7].

In the PHMM, we may consider y_t and the real valued distribution over x_t given by α_t as a state. Assuming the contexts make this composite state independent from β_{t+1} given the contexts, r_t and s_t, the α values does determine the conditional probability of y_t given the past.

We can not expect this independence in general though. And even addressing the issue seems intractable. Instead a β' backward pass may be introduced and combined with the α forward pass. The reestimation formulas may still be applied using β in a learning phase followed by using the β' backward pass when coding or assigning an adaptive code length or a probability to an image. We define a backward variable, $\beta'_t(y_t, x'_t)$ for the PHMM, where the elements of the current row are hidden. Thus y_t is hidden and x'_t reflects that the context elements on the current row are hidden.

The conditional probabilities $\beta'_t(x'_t)$ satisfy the recursion

$$\beta'_t(y_t, x'_t) = \sum_{y_{t+1}, x'_{t+1}} P(y_{t+1}|x'_{t+1}, r_t) P(x'_{t+1}|x'_t, s_t) \beta_{t+1}(y_{t+1}, x'_{t+1}). \tag{40}$$

All the parameters are derived from the PHMM model, λ. The α and the β' variables maybe combined as in (21) leading to an estimate the probability of $P(y_{t+1}|y^t)$.

4) Functions of a Markov chain revisited: Consider a probability distribution on N rows. Stationarity is of interest both in order to have a translation invariant description and to be able to express the entropy. Stationarity is obtained by having identical distributions on the top and bottom $N-1$ rows of the N row description. The implications of this requirement is not clear, though. As mentioned for $N = 2$, a way to achieve this without hidden states, i.e. x_t just assumes one fixed value, is given by requiring that, the top and bottom row are symmetric with respect to the joint two-row Markov chain. A way to achieve this using hidden states is to place the joint states $(\mathbf{y}_t, \mathbf{x}_t)$ on a cylinder and require rotational invariance around the cylinder [3]. This is referred to as the Cylinder Partially Hidden Markov Model (CPHMM).

5) Entropy of models with hidden states: Once stationarity is obtained, the entropy may be expressed by (15). For PRF, both one and two rows form Markov chains, thus the entropies are easily measured and we have $H_2 - H_1 = H(D|ABC)$. If the N rows form a Markov chain, but not the $N-1$ rows, the latter is a function of a Markov chain. Finally both the $N-1$ and the N rows may be described by a function of a Markov chain, as for the CPHMM. If the $N-1$ rows and possibly the N rows are described by a function of a Markov chain, the entropies H_{N-1} and H_N maybe bounded. The entropy of a function of a Markov chain is bounded by [1]

$$H(Y_t|Y_1^{t-1}, X_1) \leq H(Y) \leq H(Y_t|Y_1^{t-1}), \tag{41}$$

where Y_t denotes the output variable at t and X_1 the hidden state at $t = 1$. Thus the entropies in (15) may be bounded by bounding the right hand terms which are given by functions of Markov chains.

A. Hierarchical structures, mixtures and models

Above an image model based on functions of Markov chains was introduced. Expressions for reestimation of the parameters were also presented, but not a universal coding scheme. For the subset, were N rows actually form Markov chains may be treated by using Algorithm Context to learn this Markov chain, its order and parameters. For the even more restricted subset defined by the PRF, Algorithm Context may be applied directly. If the image model class is unknown, all three approaches may be combined using the MDL principle. When using hidden states and a backward pass, the length of backward pass could be restricted again using an MDL approach.

Contexts may efficiently describe local relationships within images. Many natural images as well as documents also have high-level or long range structures. To some extent, the contexts may distinguish these as well but generally not completely.

The hidden states of the PHMM may also represent switching sources, where the hidden states represent different sources, the output coming from one of these or a mixture. Shtarkov [16] presented a universal coding scheme for sequences from switching sources (with unknown parameters) which include the HMM as a special case. Unfortunately, this universal coding scheme involves summing over exponentially many sequences involving hidden states. The PHMM provides an efficient albeit not universal approach to switching sources.

Another approach is model-based coding with a high-level model combined with local coding. An example is the application of pattern matching for coding binary documents. In JBIG2 a dynamic dictionary of patterns are created. Context based coding is used both for refinement coding when the matching is not perfect and for parts which are not represented efficiently by the patterns.

V. Bi-level image coding results

The results of applying context based techniques to two bi-level images are given for illustration. The two bi-level test images are denoted S06a400 (S06) and S09a400 (S09). S06 is a mixture of text and halftone ($4352 \times 3072 = 13,369,344$ bits). S09 is an error diffused image ($1024 \times 1024 = 1,048,576$ bits). These are challenging image, for which the binary image fax coding standard (MMR), prior to the use of context based methods, provides little or no compression [6].

Template based coding is represented by a 10 pixel template as in JBIG. The results are given both for 10 fixed pixel and 9 fixed and one adaptive pixel. Adaptive context pixel positions are chosen by greedy search. Algorithm Context is a fast full path version presented in [11]. Results are given for fixed predefined and adaptive image dependent context pixel positions. Results for PHMM coding is from [7]. Templates are used with 8 (F_S), 8 (F_R), and 4 (Hidden state) binary pixels for S06 and 6 (F_S), 5 (F_R) and, 4 (Hidden state) binary pixels for S09. The model was initialized by counting over the image (two-part coding) or part of the image (adaptive coding), where the hidden state in the initialization represents non-causal pixels given by the template. Details are given in [7].

Also results for the so called free tree coding is given [11]. The tree is heavily adapted to the data by selecting new context pixels dependent on the past context so far, i.e. for each node in the tree a context pixel decides the split, whereas for Algorithm Context, the same pixel is applied at a given level of the tree. A two part code, coding the tree structure and associated context pixels as part of the header is used. The conditional probabilities are updated sequentially.

The prior fax standard method (MMR) resulted in 7,970,024 bits for S06 and even expanded the code length for s09 to more than twice the uncoded code length, 1,048,576. Thus, all the context based methods gives efficient coding compared to the prior standard (MMR).

Further, it may be noted that the more elaborate context based models do provide better performance than the template based coding.

	Template 10 pix	Template 10 (9+1)	Context Fixed	Context Free	PHMM	Free Tree
S06	2049160	1703472	1254912	1212416	1376562	1054816
S09	595312	584688	587936	538816	547758	544080

TABLE I

Lossless code lengths (bits) for context based coding of bi-level images (S06 and S09).

By the MDL principle, optimizing an image model by adaptive code length will lead to an efficient model and possibly the image model class could also be selected based on the adaptive code length. Thus the advanced context based models including the use of hidden states represents new efficient approaches to image modelling. Hidden states in HMMs are widely used in recognition schemes for sequences.

VI. Coding for two-dimensional storage

The techniques and concepts in image coding may by duality be applied to two-dimensional constrained coding for storage applications.

Some configurations of pixel values may be undesirable due to the physical media if a high data density is to be achieved. One approach is to design a code, which avoids certain undesired patterns. A simple example for a binary field is the constraint that the four neighbors of a 1 must all be zero. This may be described by a Pickard field [3]. Another example for a binary field is the no isolated bits constraint, where a pixel may not have four four-neighbors all having the complement value. This is not readily described by a Pickard field but it may be if the alphabet is extended by considering blocks of pixels and a PRF is described on the extended alphabet [5].

For the constrained code, the task may be formulated by taking an i.i.d. stream and code it to the redundant format satisfying the constraint. This mapping may in principle be performed by the inverse of arithmetic coding, based on conditional probabilities derived for the 2-D constraint. In this case the entropy of the model describing the coding process determines the capacity of the storage process, so the goal is to optimize the entropy under the 2-D constraint.

Returning to the Pickard random field and iterative scaling, we now consider selecting the initial distribution, $Q(AD|BC)$, to be uniform over the configurations admissible according to the 2-D constraint. Consider the divergence $D(P||Q) = \sum_x P(x)\log(P(x)/Q(x))$. Assigning a fixed value of $Q(AD|BC)$ for admissible x and minimizing the divergence, $D(P^*||Q)$ for $P^* \in \mathcal{L}$, leads to the maximizing the entropy [4] [5].

Maximum entropy iterative scaling of $P(AD|bc)$ is defined by (13) and (14) with the initial distribution $Q(AD|bc)$ here set to a uniform distribution over the admissible configurations $abcd$ for each bc. For each bc a sequence of distributions is generated by iterating (13) and (14).

Thus by Theorem 3.3 we may check if two Markov chains can be the boundaries of a PRF and if so if a distribution $P(D|ABC)$ exists satisfying the PRF conditions. The entropy of the interior, $H(D|ABC)$, may be written as $H(D|ABC) = H(ABCD) - H(ABC) = H(BC) + H(AD|BC) - H(ABC)$. If a solution exists, the Maximum entropy iterative scaling will find the max-entropic solution to $H(AD|BC)$, which in turn will maximize $H(D|ABC)$ for the given boundary distribution and the given constraint.

VII. Discussion

The combination of arithmetic coding and context based techniques for assigning conditional probabilities provides efficient tools for source coding, e.g. Algorithm Context provides universal coding. Contexts are also efficiently used in image coding, e.g. template based coding.

The class of a finite memory sources, for which source coding schemes as Algorithm Context are proven universal, is not easy to generalize to 2-D image data. The exception of the Pickard random field was introduced. To extend the class, hidden states were introduced in the modelling. Issues as stationarity and parameterization were discussed with context adaptive image coding in mind, but e.g deriving a stationary solution from the conditional probabilities even for a PRF poses a challenge.

Open issues remain related to defining and describing stationary causal 2-D models for which the model parameters may efficiently be determined.

References

[1] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, New York: Wiley, 1991

[2] I. Csiszár and P.C. Shields, "Information theory and statistics: A tutorial," in *Foundations and Trends in Communications and Information Theory*, vol. 1, no. 4, pp. 422–527, 2004.

[3] S. Forchhammer, J. Justesen, "Entropy bounds for constrained two-dimensional random fields," *IEEE Trans. Inform. Theory*, vol. 45, pp. 118–127, Jan. 1999.

[4] S. Forchhammer and J. Justesen, "Block Pickard models for two-dimensional constraints," submitted to *IEEE Trans. Inform. Theory*.

[5] S. Forchhammer and T.V. Laursen, "A Model for the two-dimensional no isolated bits constraint," *Proc. IEEE Int'l Symp. Inform. Theory, ISIT*, Seattle, July, 2006.

[6] S. Forchhammer and J. Rissanen, "Partially hidden Markov models," *IEEE Trans. Inform. Theory*, vol. 42, pp. 1253–1256, July 1996.

[7] S. Forchhammer and T.S. Rasmussen, "Adaptive partially hidden Markov models with application to bi-level image coding," *IEEE Trans. Image Processing*, vol. 8, pp. 1516–1526, Nov. 1999.

[8] JBIG, "Progressive Bi-level Image Compression," *ISO/IEC International Standard 11544*, 1993.

[9] G.G. Langdon and J. Rissanen, "Compression of black-white images with arithmetic coding", *IEEE Trans. Commun.*, vol. 29, pp. 858-867, June 1981.

[10] S.E. Levinson, L.R. Rabiner, and M.M. Sondhi, "An introduction to the theory of probabilistic functions of a Markov process to automatic speech recognition," *Bell Syst. Tech. J.*, vol. 62, pp. 1035–1074, Apr. 1983.

[11] B. Martins and S. Forchhammer, "Tree coding of bilevel images," *IEEE Trans. Image Processing*, vol. 7, pp. 517-528, April 1998.

[12] D. Pickard, "Unilateral Markov fields," *Adv. Appl. Probability*, vol. 12, pp. 655–671, 1980.

[13] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, pp. 257–286, Feb. 1989.

[14] J. Rissanen, "A universal data compression system", *IEEE Trans. Inform. Theory*, vol. 29, 656–664, Sept. 1983.

[15] J. Rissanen, "Complexity of strings in the class of Markov sources", *IEEE Trans. Inform. Theory*, vol. 32, 526–532, July 1986.

[16] Y. Shtarkov, "Switching discrete sources and its universal encoding", *Problemy Peredachi Informatsii*, vol. 28, 95–111, (English pp. 282–296), July-Sept. 1992.

[17] M.J. Weinberger, J. Rissanen, and M. Feder, "A universal finite memory source", *IEEE Trans. Inform. Theory*, vol. 41, 643–652, May 1995.

[18] M.J. Weinberger, J. Rissanen, and R.B. Arps, "Applications of universal context modeling to lossless compression of gray-scale images," *IEEE Trans. Image Proc.*, vol. 5, 575–586, April 1996.

Estimation of sinusoidal regression models by stochastic complexity*

Ciprian Doru Giurcăneanu
Department of Signal Processing, Tampere University of Technology,
P.O. Box 553, FIN-33101 Tampere, Finland
E-mail:ciprian.giurcaneanu@tut.fi

Abstract

Stochastic complexity (SC), or equivalently, the negative logarithm of the NML (Normalized Maximum Likelihood) was proven to be successful for the estimation of model structure in the linear quadratic regression problem. Recently, the results have been extended to autoregressive (AR) and autoregressive moving average (ARMA) models, whereas most of the information theoretic methods currently applied for determining the number of sine-waves in additive Gaussian noise still rely on asymptotic two-terms formulae where the first term is given by the minus maximum log-likelihood, and the second one is a penalty coefficient that depends on the number of parameters and the sample size. Additionally, the noise is assumed to be white, which is not realistic in most of the practical applications. Our main purpose is to apply sharper approximations of SC for estimating the number of sinusoidal terms in a time series contaminated by AR noise. This is known to be challenging because we have to solve a mixed-spectrum estimation problem. We elaborate on two different SC criteria that involve the Fisher information matrix (FIM) of the investigated model. For small and moderate sample sizes, the experimental results show that SC compares favorably with other well-known criteria such as: Bayesian information criterion (BIC), corrected Kullback information criterion (KICc) and the generalized Akaike information criterion (GAIC).

1 Introduction and preliminaries

We address the estimation of the number of sinusoids observed in additive noise with unknown correlation structure. To formulate the problem, we consider the data model

$$\begin{aligned} y_t &= x_t + e_t,\ t \in \{0, \ldots, N-1\}, \\ x_t &= \sum_{k=1}^{K} \alpha_k \cos(\omega_k t + \phi_k), \end{aligned} \tag{1}$$

*The contribution extends the results of the paper "Stochastic complexity for the estimation of sine-waves in colored noise", authored by C.D. Giurcaneanu and presented at ICASSP 2007, Honolulu, Hawaii, USA.

where y_t denotes the measurements, x_t is the noise-free signal and e_t is the colored Gaussian noise.

To ensure the identifiability of the parameters, we assume as usual that the amplitudes α_k are strictly positive and the frequencies ω_k belong to the interval $(0, \pi)$ [1]. The frequencies are distinct and, without loss of generality, $\omega_1 < \cdots < \omega_K$. Both the amplitudes and the frequencies are non-random parameters that will be estimated from the available measurements.

Two different hypotheses will be considered for modeling the phases $\phi_k \in [-\pi, \pi)$: H_{dp} - the phases are unknown deterministic constants; H_{rp} - the phases are independent and uniformly distributed random variables that are also independent of e_t. For both assumptions, the statistical properties of y_t have been investigated in previous studies, and more details can be found, for example, in [2].

In line with the approach from [2][3][4] and the references therein, we model the noise e_t as a stable autoregressive (AR) process with order M:

$$e_t \;=\; -\sum_{m=1}^{M} a_m e_{t-m} + w_t, \tag{2}$$

where w_t is a sequence of independent and identically distributed (i.i.d.) Gaussian random variables with zero mean and variance τ. Since we consider only the case of real sinusoids in real AR noise, we emphasize that the white random process w_t and the coefficients a_m, $1 \le m \le M$, are real-valued.

When the noise is white, or equivalently $M = 0$, it is well-known the definition of the local SNR for the k-th sinusoid: $\mathrm{SNR}_k = \dfrac{\alpha_k^2/2}{\tau}$. An extension of this definition, namely $\mathrm{SNR}_k = \dfrac{\alpha_k^2/2}{|H(\exp(j\omega_k))|^2}$, was also introduced in the literature [5] for the case when the additive noise is modeled as the output of an exponentially stable and invertible linear filter $H(q^{-1})$ whose input is a sequence of i.i.d. Gaussian random variables with zero mean and variance τ. We note that q^{-1} is the unit delay operator and $j = \sqrt{-1}$. For the AR noise defined in equation (2), we get immediately

$$\mathrm{SNR}_k \;=\; \frac{\alpha_k^2/2}{\tau/|A(\omega_k)|^2}, \tag{3}$$

where $A(\omega_k) = 1 + \sum_{m=1}^{M} a_m \exp(-jm\omega_k)$. A similar formula can be written without difficulties for the case when the additive noise is a moving average process.

Based on (1) and (2), we observe that the parameters of the model are $\boldsymbol{\theta} = [\boldsymbol{\xi}\ \mathfrak{a}\ \tau]^\top$, where $\boldsymbol{\xi} = [\boldsymbol{\xi}_1^\top \cdots \boldsymbol{\xi}_K^\top]^\top$ with the convention $\boldsymbol{\xi}_k = [\alpha_k\ \omega_k\ \phi_k]^\top$ for the k-th sine-wave. The notation $\mathfrak{a}$ is employed for the vector of the AR coefficients $[a_1 \cdots a_M]^\top$.

Because the model structure $\gamma = (K, M)$ is not known a priori, we resort to the traditional model selection procedure that comprises two steps:

(a) for all pairs of integers $\gamma = (K, M)$ that satisfy $0 \le K \le K_{max}$ and $0 \le M \le M_{max}$, estimate the model parameters $\hat{\boldsymbol{\theta}}_\gamma$ from the observations $y^N = y_0, \ldots, y_{N-1}$;

(b) evaluate an information theoretic criterion for all γ's considered at the first step, and choose the model structure $\hat{\gamma}$ that minimizes the criterion.

The most popular rules for model selection can be reduced to a common form with two terms: the first one is the minus maximum log-likelihood, and the second one is a penalty coefficient that depends on the number of parameters of the model and, for some criteria, also on the sample size N [6]. In general, the criteria used in practical applications are derived for $N \to \infty$, and the asymptotic approximations could potentially yield false conclusions when the sample size is small or moderate.

During recent years, the advances in stochastic complexity (SC) have led to new exact formulae or to sharper approximations for large classes of models [7][8][9], but the use of the new results in signal processing is scarce. We illustrate next how SC can be applied to estimate the structure for the model of sine-waves in Gaussian AR noise.

The rest of the paper is organized as follows. In Section 2, two different approximative formulae for SC are revisited: one proposed by Rissanen in [8], and another one introduced by Qian and Künsch in [7]. As the computation of the Rissanen sharp approximation is difficult for the sinusoidal regression model, we focus on the Qian and Künsch formula, and we investigate its properties in Section 3. Because the SC expression involves the determinant of the Fisher information matrix (FIM), the calculation of FIM is addressed in Section 4. SC and three other well-known model selection criteria are compared in Section 5 to evaluate their performances in estimating the number of sinusoids from simulated data.

2 SC for sine-waves in AR noise

We focus on the expression of SC for the class of the density functions $\{f(y^N;\boldsymbol{\theta})\}$ defined by the equations (1) and (2). For the maximum likelihood (ML) estimates we employ the notation $\hat{\boldsymbol{\theta}}(y^N)$ whenever it is necessary to emphasize on the data set. If it is clear from the context which measurements are used for estimation, then the simpler notation $\hat{\boldsymbol{\theta}}$ is preferred to $\hat{\boldsymbol{\theta}}(y^N)$. Therefore $\ln f(y^N;\hat{\boldsymbol{\theta}}) = \ln f(y^N;\hat{\boldsymbol{\theta}}(y^N))$ is the maximum log-likelihood, Θ denotes the parameter space, and $\mathbf{J}_N(\boldsymbol{\theta}) = E\left[-\dfrac{\partial^2 \ln f(y^N;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\right]$ is the Fisher information matrix (FIM). The Normalized Maximum Likelihood (NML) density function is given by [10][8],

$$\hat{f}(y^N;K,M) = \frac{f(y^N;\hat{\boldsymbol{\theta}}(y^N))}{\int_{x^N:\hat{\boldsymbol{\theta}}(x^N)\in\Theta} f(x^N;\hat{\boldsymbol{\theta}}(x^N))\mathrm{d}x^N},$$

and the stochastic complexity is defined as

$$\mathrm{SC}(y^N;K,M) = \ln\left(1/\hat{f}(y^N;K,M)\right).$$

The NML criterion has two important optimality properties [11] that recommend it to be used as a yardstick in model selection. The application of the NML criterion is appealing, but its computation is not very easy for all classes of models. Under mild assumptions on ML estimates, SC is approximated in [8] with a formula that involves the integral of the squared root of the

FIM determinant. The approximation is valid only if FIM divided by N, the number of samples, has a finite limit as $N \to \infty$. The condition is verified for most of the models used in signal processing, but not for the sinusoidal regression model [6]. We show next how the results from [8] can be extended to the sinusoidal regression model, and we also point out the difficulties with evaluating the integral term. Due to the troubles with the integral, we resort to another SC approximation that was introduced by Qian and Künsch in [7].

2.1 Sharp approximations of SC

As the Rissanen formula involves the asymptotic FIM, the following result is very useful for our application: when N is large, under both H_{dp} and H_{rp}, $\mathbf{J}_N(\boldsymbol{\theta})$ is block-diagonal such that the block $\mathbf{J}_N(\boldsymbol{\xi}_k)$ corresponds to the parameters of the k-th sine-wave and the block $\mathbf{J}_N(\mathfrak{a}, \tau)$ corresponds to the parameters of the AR noise [2][5]. More precisely, we have

$$\mathbf{J}_N(\boldsymbol{\theta}) \;=\; \begin{bmatrix} \mathbf{J}_N(\boldsymbol{\xi}_1) & & & \\ & \ddots & & \\ & & \mathbf{J}_N(\boldsymbol{\xi}_K) & \\ & & & \mathbf{J}_N(\mathfrak{a}, \tau) \end{bmatrix} \tag{4}$$

where

$$\mathbf{J}_N(\boldsymbol{\xi}_k) \;=\; \mathbf{Q}_N \mathbf{G}(\boldsymbol{\xi}_k, \mathfrak{a}, \tau) \mathbf{Q}_N, \tag{5}$$

$$\mathbf{Q}_N = \begin{bmatrix} N^{1/2} & 0 & 0 \\ 0 & N^{3/2} & 0 \\ 0 & 0 & N^{1/2} \end{bmatrix} \quad \text{and} \quad \mathbf{G}(\boldsymbol{\xi}_k, \mathfrak{a}, \tau) = \mathrm{SNR}_k \begin{bmatrix} 1/\alpha_k^2 & 0 & 0 \\ 0 & 1/3 & 1/2 \\ 0 & 1/2 & 1 \end{bmatrix}. \tag{6}$$

Here SNR_k denotes the local SNR for the k-th sinusoidal component and its formula is given in (3). The entries of $\mathbf{J}_N(\mathfrak{a}, \tau)$ are not influenced by the parameters $\boldsymbol{\xi}$, hence $\mathbf{J}_N(\mathfrak{a}, \tau)$ is the same as in the pure AR case. Based on results from [12], we can write

$$\mathbf{J}_N(\mathfrak{a}, \tau) = \frac{N}{\tau} \begin{bmatrix} \mathbf{R}(\mathfrak{a}) & \mathbf{0} \\ \mathbf{0} & 1/(2\tau) \end{bmatrix}, \tag{7}$$

where $\mathbf{R}(\mathfrak{a}) = \begin{bmatrix} r_0 & \cdots & r_{M-1} \\ \vdots & \ddots & \vdots \\ r_{M-1} & \cdots & r_0 \end{bmatrix}$ is the covariance matrix of the AR process defined in (2). Additionally we define the diagonal matrix $\mathbf{C}_N = \begin{bmatrix} \mathbf{I}_K \otimes \mathbf{Q}_N & \mathbf{0} \\ \mathbf{0} & N^{1/2} \times \mathbf{I}_{M+1} \end{bmatrix}$, where the symbol $\otimes$ denotes the Kronecker product, and for a strictly positive integer p, $\mathbf{I}_p$ is the $p \times p$ identity matrix. We adopt the convention that $\mathbf{0}$ denotes a null vector/matrix of appropriate dimensions. Based on (4)-(7), we note that $\lim_{N\to\infty} \frac{1}{N} \mathbf{J}_N(\boldsymbol{\theta})$ is not finite, whereas $\mathbf{J}(\boldsymbol{\theta}) = \lim_{N\to\infty} \mathbf{C}_N^{-1} \mathbf{J}_N(\boldsymbol{\theta}) \mathbf{C}_N^{-1}$ is finite [6]. Moreover, the ML estimates satisfy the Central Limit Theorem: the distribution

of $\mathbf{C}_N(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ converges to the Gaussian distribution of mean zero and covariance $\mathbf{J}(\boldsymbol{\theta})^{-1}$ [4]. These properties allow us to extend the results from [8] to the sinusoidal regression model for which the SC formula is given by

$$\mathrm{SCr}(y^N; K, M) = -\ln f(y^N; \hat{\boldsymbol{\theta}}) + \frac{5K + M + 1}{2} \ln \frac{N}{2\pi} + \ln \int_{\Theta} |\mathbf{J}(\boldsymbol{\theta})|^{1/2} \mathrm{d}\boldsymbol{\theta} + o(1). \tag{8}$$

We use the notation SCr to differentiate this particular approximation of SC by other formulae that will be discussed later.

Remark that $\mathbf{J}(\boldsymbol{\theta})$ is a block-diagonal matrix, and $\frac{1}{\tau}\mathbf{R}(\mathfrak{a}, \tau)$ is the block corresponding to the parameters of the Gaussian autoregressive noise. Computing the integral of $\left|\frac{1}{\tau}\mathbf{R}(\mathfrak{a}, \tau)\right|^{1/2}$ over the parameter space is a problem that arose also in the context of order estimation for AR processes [13]. Since it is hard to find a closed-form expression of the integral, the authors of [13] resorted to Monte Carlo techniques for its evaluation. Our task here is even more difficult because the other blocks of $\mathbf{J}(\boldsymbol{\theta})$ must be also considered when calculating the integral term. Hence the computational burden discourages us to apply formula (8) for estimating the number of sine-waves in Gaussian autoregressive noise. We show next that SC expression (8) becomes simpler when the noise is white ($M = 0$). In this case, $\boldsymbol{\theta} = [\boldsymbol{\xi}^\top \ \tau]^\top$, and elementary calculations lead to

$$\mathrm{SCr}(y^N; K, 0) = -\ln f(y^N; \hat{\boldsymbol{\theta}}) + \frac{5K + 1}{2} \ln \frac{N}{2\pi} + \ln \int_{\Theta} \frac{1}{2^{1/2} 96^{K/2}} \frac{\prod_{k=1}^{K} \alpha_k^2}{\tau^{(3K+2)/2}} \mathrm{d}\boldsymbol{\theta} + o(1). \tag{9}$$

To ensure that the integral term is finite, we have to assume that all amplitudes have an upper bound, $\alpha_k < \alpha_{max} < \infty$, and the noise variance has a strictly positive lower bound, $\tau > \tau_{min} > 0$. Once these conventions are adopted, the estimated number of sine-waves will depend on α_{max} and τ_{min}. Note that α_{max} and τ_{min} are just arbitrary values if we do not have a priori knowledge on the analyzed signals. The troubles with the computation of SCr make us to prefer the SC formula that was derived in [7]:

$$\mathrm{SC}(y^N; K, M) = -\ln f(y^N; \hat{\boldsymbol{\theta}}) + \ln |\mathbf{J}_N(\hat{\boldsymbol{\theta}})|^{1/2} + \sum_{i=1}^{3K+M+1} \ln(|\hat{\theta}_i| + N^{-1/4}) \tag{10}$$

A similar approximation of SC was already utilized in [14] to estimate K, the number of sinusoids. In [14], the noise variance τ is treated as a nuisance parameter in the sense that the code length to describe it is not included in the SC formula. Here we consider in (10) the cost for transmitting the value of the τ parameter, and this is the main difference between our approach and the one from [14]. We check in the Appendix how the conditions for the derivations from [7] are fulfilled for the model of sinusoids in Gaussian AR noise. We also give in the Appendix more details on the accuracy of the approximation in formula (10).

In the next Section, we investigate the asymptotic behavior of the SC criterion and we show its relation with well-known selection rules like Bayesian information criterion (BIC) [15], Minimum Description Length (MDL) [16], and the *maximum a posteriori* (MAP) probability criterion [17]. During the asymptotic analysis, we check also the necessary conditions for the consistency [18] of SC. For small and moderate sample sizes, we draw a parallel between SC and two other recently introduced model selection methods: Conditional Model Estimator (CME) [19] and the Exponentially Embedded Families (EEF) [20].

3 Some properties of SC and its relation to other model selection criteria

3.1 BIC, MDL and MAP

Based on the results from the Appendix, we obtain readily the well-known asymptotic identity $\lim_{N\to\infty} \ln |\mathbf{J}_N(\hat{\boldsymbol{\theta}})|^{1/2} = \dfrac{5K+M+1}{2} \ln N$, and it is easy to notice that the sum of the first two terms in SC is equivalent with the Bayesian information criterion:

$$\mathrm{BIC}(y^N; K, M) = -\ln f(y^N; \hat{\boldsymbol{\theta}}) + \frac{5K+M+1}{2} \ln N. \tag{11}$$

More details on the derivation of BIC can be found in [6] and the references therein. In [21], it is investigated the possibility of improving the performances of BIC for small and moderate sample sizes by considering two terms that are neglected in the asymptotic formula (11): the first one involves the logarithm of the determinant of the observed FIM, and the second one is mainly determined by a prior over the family of the analyzed models. As the sinusoidal regression model is not discussed in [21], we restrict our interest to the celebrated BIC selection rule (11), and we do not consider in simulations any sharp approximation of the Bayesian information criterion. We mention for completeness that formula (11) was also obtained in [3] as a crude version of the MDL, and its consistency was demonstrated in the same study. In [17], the use of the MAP methodology in conjunction with asymptotic approximations led also to (11) for the particular case of white noise.

3.2 A short note on the consistency of SC for $M = 0$

For ease of presentation we investigate the consistency of the criterion $\mathrm{SC}'(y^N; K, 0) = \mathrm{SC}(y^N; K, 0) - \frac{1}{2} \ln N$. It is evident that SC′ and SC are equivalent selection rules because $\frac{1}{2} \ln N$ is independent of K. We focus on the last term in (10), and for simplicity we assume $M = 0$. If zero does not belong to the domain of the parameter θ_i, then $\hat{\theta}_i \neq 0$ and $\ln(|\hat{\theta}_i| + N^{-1/4})$ is much smaller than $\frac{1}{2} \ln N$ [7]. Hence the term $\ln(|\hat{\theta}_i| + N^{-1/4})$ becomes important only when $\hat{\theta}_i \approx 0$. Since among the $\boldsymbol{\xi}$ parameters only the phases can be equal to zero, the penalty term in SC′ formula takes values between $\frac{9K}{4} \ln N$ and $\frac{5K}{2} \ln N$ when N is large. Based on formula (10), we can write

$$\mathrm{SC}'(y^N; K, 0) = -\ln f(y^N; \hat{\boldsymbol{\theta}}) + K\zeta(N, \hat{\boldsymbol{\theta}}),$$

and asymptotically $\frac{9}{4} \ln N \leq \zeta(N, \hat{\boldsymbol{\theta}}) \leq \frac{5}{2} \ln N$. Thus $\lim_{N\to\infty} \dfrac{\zeta(N, \hat{\boldsymbol{\theta}})}{N} = 0$ and $\liminf_{N\to\infty} \dfrac{\zeta(N, \hat{\boldsymbol{\theta}})}{\ln N} > 1$. If supplementarily the model (1) verifies $\frac{\omega_k}{2\pi} \in \left\{ \frac{1}{N}, \cdots, \frac{(N-1)/2}{N} \right\} \forall k \in \{1, \ldots, K\}$, all the conditions for the application of the Theorem from [18] are satisfied. We select $\hat{K}$ to be the minimum nonnegative integer for which $\mathrm{SC}'(y^N; K, 0) < \mathrm{SC}'(y^N; K+1, 0)$, and the Theorem guarantees that $\hat{K}$ converges almost surely to the true number of sinusoids.

3.3 CME, EEF and an example from [20]

In [20], it was shown that using the determinant of FIM as a penalty term could lead to modest results when the sample size is small. As the example from [20] involves sinusoidal signals, we briefly discuss it in the sequel: the noise-free signal x_t is generated like in (1) by a sum of $K = 3$ sine-waves whose parameters are $\boldsymbol{\xi}_1 = [1\ 0.2\pi\ 0]^\top$, $\boldsymbol{\xi}_2 = [1\ 0.22\pi\ 0]^\top$ and $\boldsymbol{\xi}_3 = [1\ 0.24\pi\ 0]^\top$. The white noise e_t is Gaussian with variance $\tau = 10$, and the selection is restricted to the class of nested models $\mathcal{M}_\kappa$, $\kappa \in \{2, 4, \ldots, 16\}$, defined by

$$\mathcal{M}_\kappa \ : \ y_t = \sum_{k=1}^{\kappa/2} \alpha_k \cos(\omega_k t + \phi_k) + e_t, \ t \in \{0, \ldots, N-1\},$$

where $\omega_k = 2\pi \left(0.1 + \frac{k-1}{100}\right)$. Since the frequencies are known, $\mathcal{M}_\kappa$ reduces to the linear regression for which the observation matrix has the expression

$$\mathbf{H}_\kappa = \begin{bmatrix} 1 & 0 & \cdots & 1 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \cos(\omega_1(N-1)) & \sin(\omega_1(N-1)) & \cdots & \cos(\omega_{\kappa/2}(N-1)) & \sin(\omega_{\kappa/2}(N-1)) \end{bmatrix},$$

and the vector of the unknown parameters is given by $\mathbf{v}_\kappa = [A_1\ B_1 \cdots A_{\kappa/2}\ B_{\kappa/2}]^\top$, where $A_k = \alpha_k \cos\phi_k$ and $B_k = -\alpha_k \sin\phi_k$ for all $k \in \{1, \ldots, \kappa/2\}$. The noise variance τ is assumed to be known. Remark that the number of parameters for the $\mathcal{M}_\kappa$ model is κ. Applying the CME criterion is equivalent with choosing the model $\mathcal{M}_{\hat{\kappa}}$ that minimizes [19]

$$\mathrm{CME}(y^N; \kappa) = \frac{\mathrm{RSS}_\kappa}{\tau} + \ln \left| \frac{\mathbf{H}_\kappa^\top \mathbf{H}_\kappa}{2\pi\tau} \right|,$$

where RSS_κ is the residual sum of squares obtained when fitting the $\mathcal{M}_\kappa$ model to the observations y^N. In [20], it was utilized the approximation $\mathbf{H}_\kappa^\top \mathbf{H}_\kappa \approx (N/2)\mathbf{I}_\kappa$ to show that the second term in the equation above is negative when $N < 125$, thus the penalty term of the CME criterion decreases when κ increases. Since for the models considered in this example, $\frac{1}{\tau}\mathbf{H}_\kappa^\top \mathbf{H}_\kappa$ coincides with the FIM [1], Kay concluded in [20] that all criteria whose penalty factor is given by the determinant of FIM will always choose the most complex model when the sample size is small or moderate. To circumvent such difficulties, he introduced the EEF criterion that, for linear regression models, amounts to select the $\mathcal{M}_{\hat{\kappa}}$ model that minimizes

$$\mathrm{EEF}(y^N; \kappa) = \left[-Q_\kappa + \kappa \left(\ln \frac{Q_\kappa}{\kappa} + 1 \right) \right] u\left(\frac{Q_\kappa}{\kappa} - 1 \right), \tag{12}$$

where $Q_\kappa = \dfrac{\|\mathbf{H}_\kappa \hat{\mathbf{v}}_\kappa\|^2}{\tau}$, the entries of $\hat{\mathbf{v}}_\kappa$ are the ML estimates of the parameters, and $u(\cdot)$ is the step unit function [20].

We use the same example to investigate if similar drawbacks appear when the model selection relies on the SC criterion. The FIM-based SC approximation [8] was computed in [22] for the linear regression case, and it involves the ranges of the parameters, which is not convenient as

we have already pointed out in Section 2.1. Fortunately we do not need to resort to such an approximation because Rissanen gave in [9] a very elegant solution to the problem of evaluating SC for the linear regression model. For the analyzed example, we prefer to apply the result from [9] in the form that was worked out in [23]:

$$\mathrm{SClr}(y^N;\kappa) = \frac{1}{\tau}\sum_{i=0}^{N-1} y_i^2 - Q_\kappa + \kappa\left(\ln\frac{Q_\kappa}{\kappa}+1\right) + \ln\kappa,$$

where the notations are the same like in (12). We observe that unlike the CME criterion, SClr does not contain the term given by the determinant of FIM. Moreover, the expressions of SClr and EEF are very similar. It is easy to note that $-Q_\kappa$ decreases with κ. Let us consider first the case when $\frac{Q_\kappa}{\kappa} > 1$. In general, the term $\kappa\left(\ln\frac{Q_\kappa}{\kappa}+1\right)$ increases with κ [20], hence it is a penalty term for both EEF and SClr. Remark that in this case, due to the $\ln\kappa$ term, the penalty will be more stringent for SClr than for EEF. Formula (12) can be re-written as $\mathrm{EEF}(y^N;\kappa) = \kappa h(Q_\kappa/\kappa)$, where $h(x) = -x + \ln x + 1, \forall x \in (0,\infty)$. Because $h(x)$ is strictly negative for $x > 1$, the criterion EEF has the same property. Whenever $\frac{Q_\kappa}{\kappa} \le 1$, $\mathrm{EEF}(y^N;\kappa)$ takes value zero, and consequently the model $\mathcal{M}_\kappa$ will not be selected. For SClr, if $\frac{Q_\kappa}{\kappa}$ is small, the term $\kappa\left(\ln\frac{Q_\kappa}{\kappa}+1\right)$ could become negative and $\ln\kappa$ will remain the only penalty term.

For Gaussian linear regression with known noise variance, another SC criterion was derived in [24] by using the universal mixture model instead of the NML:

$$\mathrm{SChy}(y^N;\kappa) = \frac{1}{2\tau}\sum_{i=0}^{N-1} y_i^2 + \frac{1}{2}\left[-Q_\kappa + \kappa\left(\ln\frac{Q_\kappa}{\kappa}+1\right) + \ln N\right] u\left(\frac{Q_\kappa}{\kappa}-1\right). \tag{13}$$

As it was pointed out in [24], SChy coincides up to the $\frac{1}{2}\ln N$ additive term with the empirical Bayesian selection rule proposed in [25]. Comparing (12) and (13) we also note that EEF and SChy are essentially the same.

4 Computational issues

The use of (10) is very appealing from computational viewpoint, but it was already pointed out in [7] that (10) is not invariant under re-parametrization. Due to this reason, we prefer to use as parameters for the AR noise the magnitudes and the angles of the poles instead of the coefficients.

More precisely, let us assume that the poles of the AR noise model are $g_1,\ldots,g_M$: if the poles $g_1,\ldots,g_{M_1}$ are real-valued, then the pure complex poles $g_{M_1+1},\ldots,g_M$ occur in complex conjugate pairs because the coefficients $\mathfrak{a}$ are real-valued. Instead of $\boldsymbol{\theta} = [\boldsymbol{\xi}\ \mathfrak{a}\ \tau]^\top$, we will apply the parametrization $\boldsymbol{\eta} = [\boldsymbol{\xi}\ \mathfrak{g}\ \tau]^\top$, where $\mathfrak{g} = [g_1 \ldots g_{M_1}\ |g_{M_1+1}|\ \psi_{g_{M_1+1}} \ldots |g_{M-1}|\ \psi_{g_{M-1}}]^\top$, and for a complex pole g_i, the symbol ψ_{g_i} denotes its angle. Remark the range of the entries of $\mathfrak{g}$: we have $g_i \in (-1,1)$ for $1 \le i \le M_1$, and for the rest of the parameters $|g_i| \in (0,1)$ and $\psi_{g_i} \in (0,\pi)$.

SC	Hypothesis	$\mathbf{J}_N(\boldsymbol{\xi})$	$\mathbf{J}_N(\mathfrak{g})$
SCp	$\mathrm{H_{rp}}$	exact	exact
SCa	$\mathrm{H_{rp}}/\mathrm{H_{dp}}$	asymptotic	asymptotic
SCe	$\mathrm{H_{dp}}$	exact	exact

Table 1: Nomenclature for SC when various formulae for FIM are used in calculations.

To calculate the determinant of the FIM with the new parametrization, we use the general result on the transformation of parameters [1] in conjunction with the result of equation (15) from [2]. For writing the equations in a more compact form, we define the $(M+1)\times(M+1)$ matrix $\left[\frac{D(\mathfrak{a},\tau)}{D(\mathfrak{g},\tau)}\right]$ whose (m,n)-th element is $\frac{\partial \mathfrak{a}_m}{\partial \mathfrak{g}_n}$ if $1\le m,n\le M$, it is one if $m=n=M+1$, and otherwise takes value zero. Next we obtain the following identities:

$$\begin{aligned}|\mathbf{J}_N(\boldsymbol{\eta})| &= \left|\begin{array}{cc}\mathbf{I}_{3K} & \mathbf{0}\\ \mathbf{0} & \left[\frac{D(\mathfrak{a},\tau)}{D(\mathfrak{g},\tau)}\right]^\top\end{array}\right|\left|\begin{array}{cc}\mathbf{J}_N(\boldsymbol{\xi}) & \mathbf{0}\\ \mathbf{0} & \mathbf{J}_N(\mathfrak{a},\tau)\end{array}\right|\left|\begin{array}{cc}\mathbf{I}_{3K} & \mathbf{0}\\ \mathbf{0} & \left[\frac{D(\mathfrak{a},\tau)}{D(\mathfrak{g},\tau)}\right]\end{array}\right|\\ &= |\mathbf{J}_N(\boldsymbol{\xi})||\mathbf{J}_N(\mathfrak{g},\tau)|\end{aligned}$$

The block $\mathbf{J}_N(\boldsymbol{\xi})$ that corresponds to the signal parameters can be evaluated with the fast algorithms from [2]: the exact $\mathbf{J}_N(\boldsymbol{\xi})$ is different for $\mathrm{H_{dp}}$ and $\mathrm{H_{rp}}$, but the asymptotic $\mathbf{J}_N(\boldsymbol{\xi})$ has the same form under both hypotheses. This asymptotic form is well-known [5], and it is also given in Section 2.1. $\mathbf{J}_N(\mathfrak{a},\tau)$ has the same expression as in the pure AR case, and for its calculation we resort to the exact and the asymptotic formulae from [12]. The conversion from $\mathbf{J}_N(\mathfrak{a},\tau)$ to $\mathbf{J}_N(\mathfrak{g},\tau)$ can be easily performed with the results from [26]. A discussion on the asymptotic form of $\mathbf{J}_N(\mathfrak{g},\tau)$ can be found in [13].

Applying the exact or asymptotic formulae for $\mathbf{J}_N(\boldsymbol{\xi})$ and $\mathbf{J}_N(\mathfrak{g},\tau)$ leads to various expressions for SC. In Table 1, we explain the nomenclature for SC when FIM in (10) is evaluated with various formulae.

For better understanding the differences between SCp, SCa and SCe we resort to one of the examples used in [2] to analyze the Cramer-Rao bound (CRB). Let us consider the case of one single sinusoid ($K=1$) in AR noise with order $M=2$. We choose $\alpha_1=1$, $\omega_1=\pi/2$, the modulus of the AR poles is $|g_1|=0.9$, and the sample size is $N=35$. The angle ψ_{g_1} takes values between 0.02 and $(\pi-0.02)$, and the variance τ is selected such that to keep constant $\mathrm{SNR}_1=3$ dB. Evaluating the differences between SCp, SCa and SCe reduces to calculate $\ln|\mathbf{J}_N(\boldsymbol{\eta})|^{1/2}$ with various formulae. Because under $\mathrm{H_{dp}}$ the exact $\mathbf{J}_N(\boldsymbol{\xi})$ depends on the phase ϕ_1, for each ψ_{g_1} we compute $\ln|\mathbf{J}_N(\boldsymbol{\eta})|^{1/2}$ for sixty different values of ϕ_1 that are equally spaced in $[-\pi,\pi)$, and the largest ($\bigtriangledown$) and the smallest ($\bigtriangleup$) results are plotted in Figure 1. We plot in the same Figure the values of $\ln|\mathbf{J}_N(\boldsymbol{\eta})|^{1/2}$ used in the calculation of SCp (dash-dot line) and SCa (continuous line). For sake of comparison, we draw also a horizontal line that corresponds to $\frac{5K+M+1}{2}\ln N$. We can easily extend the conclusions on CRB drawn in [2], by observing the significant difference between the asymptotic approximation of $\ln|\mathbf{J}_N(\boldsymbol{\eta})|^{1/2}$ and its exact value when the line spectrum is close to the spectral peak of the noise. Remark also in Figure 1 that

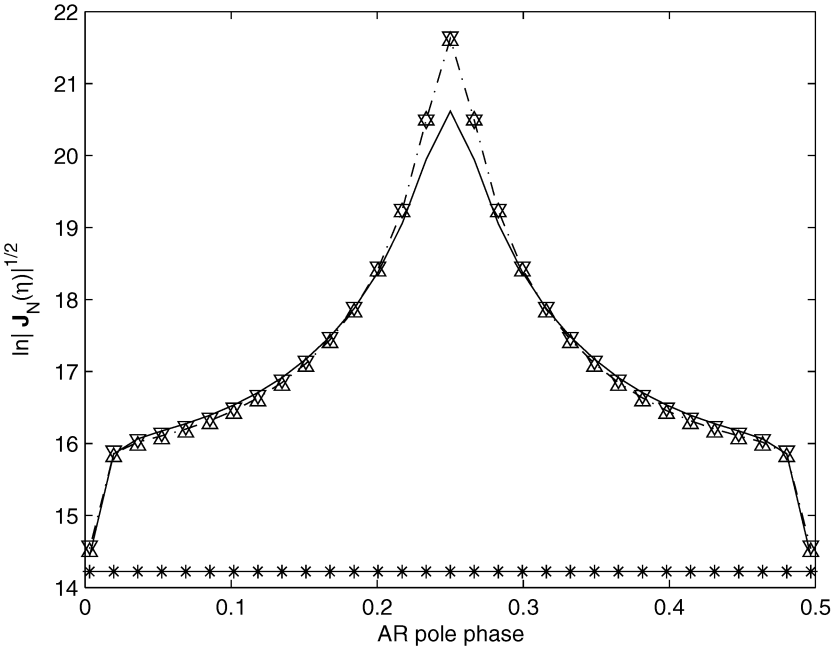


Figure 1: The term $\ln|\mathbf{J}_N(\boldsymbol{\eta})|^{1/2}$ versus the phase ψ_{g_1} of the AR pole when the sample size is $N = 35$. In the case of the SCe formula, $\ln|\mathbf{J}_N(\boldsymbol{\eta})|^{1/2}$ is calculated for sixty different values of $\phi_1 \in [-\pi, \pi)$, and the largest ($\bigtriangledown$) and the smallest ($\bigtriangleup$) results are plotted. The dash-dot line and the continuous line are for the values of $\ln|\mathbf{J}_N(\boldsymbol{\eta})|^{1/2}$ as they are used in the evaluation of the SCp and SCa, respectively. The horizontal line with a $\star$ at each data point corresponds to $\frac{5K+M+1}{2}\ln N$.

the value of $\ln|\mathbf{J}_N(\boldsymbol{\eta})|^{1/2}$ used to compute SCp is approximately equal with the average of the maximum and the minimum of $\ln|\mathbf{J}_N(\boldsymbol{\eta})|^{1/2}$ employed in the calculation of SCe.

In the next Section we investigate how the structure estimation performances of SC are influenced by the use of various formulae for FIM.

5 Experimental results

In all the examples presented next, we resort to the RELAX algorithm that performs a decoupled parameter estimation for the sinusoids and the AR noise [4]. In our simulations we have used for the implementation of RELAX the Matlab functions that are publicly available at *http://www.uni-kassel.de/fb16/hfk/neu/toolbox*.

Asymptotically both RELAX and the maximum likelihood (ML) yield statistically efficient estimates, and the use of RELAX is recommended due to its lower computational burden [4][5].

For $\gamma = (K, M)$, let $\hat{\boldsymbol{\xi}}_k$ be the parameters of the k-th sinusoid estimated with RELAX. We denote $\hat{e}_t = y_t - \sum_{k=1}^{K} \hat{\alpha}_k \cos(\hat{\omega}_k t + \hat{\phi}_k)$, and let $\hat{\mathbf{a}}$ be the coefficients of the AR noise determined from the sequence $\hat{e}_0, \ldots, \hat{e}_{N-1}$. We further define the residual sum of squares as $\mathrm{RSS}_\gamma = \sum_{t=0}^{N-1} \left[\hat{e}_t + \sum_{m=1}^{M} \hat{a}_m \hat{e}_{t-m}\right]^2$, with the convention that $\hat{e}_t = 0$ for $t < 0$.

The performances of SC are compared in our simulations with BIC (11) and two other criteria: GAIC and KICc. GAIC is a generalized Akaike Information Criterion that was traditionally used in conjunction with the RELAX algorithm [4]. It seeks for the model structure γ that minimizes

$$\mathrm{GAIC}(y^N; K, M) = N \ln \mathrm{RSS}_\gamma + 8(3K + M + 1) \ln(\ln N).$$

KICc was derived in [27] as a unbiased Kullback Information Criterion for linear regression models with i.i.d. Gaussian noise. Since then its application was extended also to other classes of models, see for example [28] and the references therein. Applying KICc is equivalent with selecting the model structure γ that minimizes [27]

$$\mathrm{KICc}(y^N; K, M) = -2 \ln f(y^N; \hat{\boldsymbol{\theta}}) + 2\frac{(\kappa + 1)N}{N - \kappa - 2} - N\psi\left(\frac{N - \kappa}{2}\right) + N \ln \frac{N}{2}, \tag{14}$$

where $\kappa = K + M$ and $\psi(\cdot)$ is the *digamma* function [29]. In SC (10), BIC (11) and KICc (14), $-\ln f(y^N; \hat{\boldsymbol{\theta}})$ is evaluated as $\frac{N}{2} \ln \mathrm{RSS}_\gamma$ after discarding the terms that do not depend on γ.

In our settings, the maximum number of sinusoids is $K_{max} = 8$, and the maximum order of the AR process depends on the number of the available measurements: $M_{max} = \lfloor \ln^2 N \rfloor - 1$. The formula for M_{max} is derived from the condition used in [3] to ensure the consistency of the BIC criterion. Supplementarily, each pair (K, M) must verify the inequality $3K + M < N - 2$ to be a candidate for the model structure.

Examples 1-3 are taken from [3], where the estimation results are reported only for $N \geq 128$. Since our main interest is on small and moderate sample sizes, we evaluate the performances of the information theoretic criteria for $N \in \{25, \ldots, 100\}$ and various levels of the local SNR. In Examples 1-3, we consider $K = 2$ sinusoids whose parameters are $\boldsymbol{\xi}_1 = [2^{1/2}\ 1\ 0]^\top$ and $\boldsymbol{\xi}_2 = [2^{-1/2}\ 2\ 0]^\top$. The additive noise is generated as follows:
Example 1: $e_t = \varepsilon_t$ (white noise),
Example 2: $e_t = -0.81e_{t-2} + \varepsilon_t$ (autoregressive noise),
Example 3: $e_t = \varepsilon_t + 1.6\varepsilon_{t-1} + 0.64\varepsilon_{t-2}$ (moving average noise),
where ε_t is a sequence of i.i.d. Gaussian random variables with zero mean and variance chosen such that the local SNR's take the desired values.
Example 4 is taken from [4] and modified such that the observations y^N are real-valued. The number of sinusoids is $K = 3$ and their parameters are $\boldsymbol{\xi}_1 = [2\ 0.10\pi\ 0]^\top$, $\boldsymbol{\xi}_2 = [2\ 0.80\pi\ 0]^\top$ and $\boldsymbol{\xi}_3 = [2\ 0.84\pi\ 0]^\top$. The noise is simulated by the autoregressive process $e_t = 0.85e_{t-1} + \varepsilon_t$, where the significance of ε_t is the same as above.

We focus on the capabilities of the tested criteria to estimate correctly the number of sinusoids K. For the Examples 1-4, we count the number of correct estimates for 100 runs when the local SNR's and the sample size N take various values. The results are reported in Tables 2-5.

SNR_2=-3.00 dB								
N	30	40	50	60	70	80	90	100
SCp	41	61	**79**	**86**	**94**	**94**	**97**	**100**
SCa	41	63	76	84	**94**	93	**97**	**100**
SCe	26	52	64	63	78	84	89	89
BIC	25	41	58	68	87	85	90	96
KICc	**54**	**80**	77	74	61	45	44	40
GAIC	3	6	13	22	38	59	65	67
SNR_2=-1.00 dB								
N	30	35	40	45	50	60	80	100
SCp	62	72	**88**	**93**	**95**	**99**	98	99
SCa	64	67	71	89	85	91	94	98
SCe	44	59	71	83	82	85	86	96
BIC	43	56	69	78	80	91	**99**	**100**
KICc	**81**	**81**	79	77	78	67	49	38
GAIC	7	19	40	40	53	73	93	**100**
SNR_2=0.00 dB								
N	25	30	35	40	45	50	75	100
SCp	60	79	**93**	**96**	**97**	**98**	**98**	99
SCa	60	68	76	76	86	88	94	97
SCe	45	68	75	80	81	90	92	96
BIC	37	51	68	82	86	88	**98**	99
KICc	**75**	**89**	89	92	91	80	56	40
GAIC	6	18	37	52	67	74	97	**100**

Table 2: Example 1: the counts indicate for 100 runs the number of times the number of sinusoids was correctly estimated by each criterion. The best result for each sample size N is represented with bold font.

We note that the estimation results are similar with those reported in [14]. SCp is the best among the SC formulae and its performances are closely followed by SCa. For both SCp and SCa, FIM of the sinusoidal components are decoupled [2], which is a serious computational advantage. From the results reported in [2] together with the outcome of the Example discussed in Section 4, we can draw the conclusion that the shape of the noise spectrum has more influence on SCp than on SCa, and this explains the superiority of the SCp criterion. The performances of SCe are very modest because FIM used in SCe can be ill-conditioned for small and moderate sample size when the number of sinusoids is two or larger [2].

When the sample size N is smaller than 80, SCa is superior to BIC and GAIC. This is a straightforward consequence of the asymptotic approximations applied in the derivations of the BIC and GAIC criteria. KICc estimates for the number of sinusoids are remarkably correct when $N \leq 40$, but the number of correct estimations yield by KICc declines when N increases such that for $N \geq 80$ the reported results are very modest.

SNR_2=1.00 dB								
N	30	40	50	60	70	80	90	100
SCp	**26**	**63**	**76**	**81**	**90**	**87**	87	92
SCa	25	53	74	78	87	84	86	92
SCe	24	34	74	79	89	**87**	87	91
BIC	9	41	61	76	89	86	**89**	**93**
KICc	23	36	35	37	42	36	33	43
GAIC	5	12	28	42	64	78	**89**	**93**
SNR_2=3.00 dB								
N	30	35	40	45	50	60	80	100
SCp	**50**	**61**	**76**	**96**	**92**	**95**	96	97
SCa	44	52	60	86	83	93	95	97
SCe	41	46	34	50	88	88	96	97
BIC	22	41	53	71	77	91	95	**98**
KICc	29	44	45	38	42	43	46	45
GAIC	7	17	28	38	54	75	**97**	**98**
SNR_2=5.00 dB								
N	25	30	35	40	45	50	75	100
SCp	36	**63**	**81**	**88**	**90**	**95**	95	99
SCa	**41**	58	73	51	67	91	93	99
SCe	27	54	57	21	15	89	94	99
BIC	16	28	65	57	77	83	99	**100**
KICc	29	40	57	54	47	52	50	44
GAIC	14	16	32	47	60	72	**100**	**100**

Table 3: Example 2: the performances in estimating the number of sine-waves reported with the same conventions as in Table 2.

We extend our analysis by counting the Type I and Type II errors. Let $f_k = \omega_k/(2\pi)$ and similarly $\hat{f}_k = \hat{\omega}_k/(2\pi)$. Since K and $\hat{K}$ are not necessarily equal, we take $\mathcal{K} = \min(K, \hat{K})$. We select the indices $\{i_1, \ldots, i_{\mathcal{K}}\} \subseteq \{1, \ldots, K\}$ and $\{j_1, \ldots, j_{\mathcal{K}}\} \subseteq \{1, \ldots, \hat{K}\}$ such that $|f_{i_1} - \hat{f}_{j_1}|, \ldots, |f_{i_{\mathcal{K}}} - \hat{f}_{j_{\mathcal{K}}}|$ are the smallest entries of the set $\{|f_i - \hat{f}_j| \ : \ 1 \le i \le K, 1 \le j \le \hat{K}\}$. For each $k \in \{1, \ldots, \mathcal{K}\}$, $\hat{f}_{j_k}$ is deemed to be the estimate for f_{i_k}. As usual, a Type I error is counted in connection with the frequency f_k if none of the estimated frequencies are assigned to f_k, and a Type II error is counted whenever $\hat{K} > K$. We compute also the mean-squared errors (MSE) for the frequency estimates.

For brevity, we report in Tables 6-9 the Type I and Type II errors together with the MSE for one single experiment conducted in each Example. In our comparisons, we consider SCp and the asymptotic criteria BIC and GAIC.

Because in Example 2 the simulated noise is an autoregressive process, we propose to analyze more carefully the data shown in Table 7. Remark that only GAIC has difficulties in recovering the first harmonic when $N > 35$, and recovering the second harmonic whose local SNR is smaller

SNR_1=1.00 dB								
N	30	40	50	60	70	80	90	100
SCp	**64**	**85**	**85**	**91**	**78**	**88**	**88**	**85**
SCa	51	58	61	69	67	80	82	81
SCe	59	75	83	**91**	77	85	**88**	**85**
BIC	32	55	58	73	70	78	70	78
KICc	60	69	65	75	59	58	58	59
GAIC	11	39	45	62	63	72	63	73
SNR_1=3.00 dB								
N	30	35	40	45	50	60	80	100
SCp	**93**	**86**	**94**	**96**	**92**	**94**	**93**	87
SCa	60	60	60	61	67	71	84	86
SCe	85	74	81	91	89	**94**	**93**	87
BIC	56	55	72	80	82	85	89	**91**
KICc	82	74	75	76	77	72	62	57
GAIC	50	60	70	84	80	84	90	89
SNR_1=5.00 dB								
N	25	30	35	40	45	50	75	100
SCp	**97**	**96**	**95**	**97**	**94**	**96**	92	94
SCa	61	58	61	68	61	59	82	89
SCe	83	86	85	86	88	**96**	92	94
BIC	53	71	72	85	80	86	93	93
KICc	90	89	81	88	79	81	63	59
GAIC	45	75	85	95	93	93	**95**	**95**

Table 4: Example 3: the performances in estimating the number of sine-waves reported with the same conventions as in Table 2.

posses problems to all the criteria. Note for SCp that the number of Type I errors connected with f_2 decreases fast with the increase of the sample size. For GAIC, the number of Type II errors is always small, but many Type I errors occur even for $N = 60$. This is a clear sign that, for small N, GAIC underestimates the number of sinusoids. The computed MSE is almost the same for all the investigated criteria and this is natural because the evaluation of SCp, BIC and GAIC is based on the estimates provided by the RELAX algorithm.

Final remarks

The new results on SC for the sinusoidal regression model illustrate very nicely the main idea that SC is not just the minus maximum log-likelihood term penalized with $\frac{k}{2} \ln N$, where k is the number of parameters and N is the number of samples. The most important achievement is to show that, for small and moderate sample sizes, the adequate use of SC could improve the estimation performances even for problems that have been intensively researched in the past, as

SNR_1=-5.00 dB								
N	30	40	50	60	70	80	90	100
SCp	24	65	**81**	**74**	**70**	**70**	**81**	84
SCa	22	65	**81**	**74**	**70**	**70**	**81**	**85**
SCe	32	64	70	66	63	66	79	83
BIC	15	53	79	64	65	69	70	79
KICc	**49**	**75**	79	70	**70**	67	73	71
GAIC	0	3	19	39	49	64	64	72
SNR_1=-3.00 dB								
N	30	40	50	60	70	80	90	100
SCp	20	82	**88**	**86**	**90**	**93**	**91**	**95**
SCa	20	**83**	**88**	**86**	**90**	91	84	89
SCe	33	71	80	78	83	87	86	92
BIC	13	72	**88**	79	78	83	90	94
KICc	**51**	**83**	84	81	80	78	74	72
GAIC	0	23	48	56	66	70	75	81
SNR_1=-1.00 dB								
N	30	35	40	45	50	60	80	100
SCp	30	80	**90**	**93**	**99**	**90**	**94**	92
SCa	30	80	86	**93**	97	88	77	73
SCe	40	72	74	80	90	83	91	87
BIC	27	83	85	87	95	**90**	92	95
KICc	**61**	**91**	89	89	90	82	76	65
GAIC	0	13	38	68	79	74	90	**97**

Table 5: Example 4: the performances in estimating the number of sine-waves reported with the same conventions as in Table 2.

it is the case with the mixed-spectrum estimation.

Acknowledgements

This work was supported by the Academy of Finland, project No. 113572, 118355 and 213462. The author is thankful to Gabriel Dospinescu from Ecole Normale Superieure, Paris, for the help with the proof of the Remark 2 in the Appendix.

Freq.	N		30	35	40	45	50	60	80	100
f_1	Err.1	SCp	0	0	0	0	0	0	0	0
		BIC	19	11	8	4	4	3	0	0
		GAIC	73	48	25	4	4	0	0	0
	MSE	SCp	-55.92	-56.45	-57.59	-60.01	-59.30	-62.62	-65.32	-68.79
		BIC	-56.15	-56.37	-57.52	-60.01	-59.17	-62.65	-65.32	-68.79
		GAIC	-56.61	-56.60	-58.30	-59.59	-59.04	-63.12	-65.32	-68.79
f_2	Err.1	SCp	38	25	10	7	5	0	0	0
		BIC	51	40	26	21	20	8	1	0
		GAIC	93	81	60	60	47	27	7	0
	MSE	SCp	-49.04	-51.43	-52.38	-54.00	-54.82	-56.64	-59.96	-63.19
		BIC	-46.09	-51.59	-51.90	-53.67	-54.47	-56.67	-59.92	-63.19
		GAIC	-46.08	-52.16	-51.71	-54.58	-55.88	-56.15	-59.97	-63.19
Err.2		SCp	0	3	2	0	0	1	2	1
		BIC	6	4	5	1	0	1	0	0
		GAIC	0	0	0	0	0	0	0	0

Table 6: Type I and Type II errors for Example 1 when SNR_2 = -1.00 dB. MSE is computed for the estimates of the frequencies and it is expressed in dB. The results are reported for 100 runs.

Freq.	N		30	35	40	45	50	60	80	100
f_1	Err.1	SCp	5	1	0	0	0	0	0	0
		BIC	9	8	1	0	0	1	0	0
		GAIC	28	20	14	12	9	9	1	0
	MSE	SCp	-58.96	-60.13	-61.20	-63.05	-64.48	-65.73	-71.68	-80.15
		BIC	-59.54	-59.87	-61.72	-63.53	-64.04	-65.69	-71.68	-80.15
		GAIC	-59.20	-60.62	-61.99	-63.66	-64.35	-65.66	-71.10	-80.15
f_2	Err.1	SCp	32	21	14	1	2	1	0	0
		BIC	24	29	20	7	7	4	1	0
		GAIC	83	74	66	61	42	22	1	0
	MSE	SCp	-49.97	-53.42	-55.07	-56.00	-42.96	-59.76	-64.03	-64.98
		BIC	-40.71	-39.15	-55.34	-56.82	-42.73	-60.11	-64.09	-64.98
		GAIC	-50.24	-53.27	-54.20	-56.21	-58.35	-59.84	-64.09	-64.98
Err.2		SCp	18	18	10	3	6	4	4	3
		BIC	54	30	27	22	16	5	4	2
		GAIC	10	9	6	1	4	3	2	2

Table 7: Type I and Type II errors for Example 2 when SNR_2 = 3.00 dB.

Freq.	N		30	35	40	45	50	60	80	100
f_1	Err.1	SCp	0	1	0	0	1	0	0	0
		BIC	5	8	1	2	3	1	0	2
		GAIC	43	32	20	14	18	9	4	6
	MSE	SCp	-52.69	-55.06	-56.26	-58.26	-59.12	-60.96	-64.09	-66.39
		BIC	-53.15	-54.38	-56.05	-58.34	-59.23	-60.97	-64.09	-66.50
		GAIC	-52.54	-55.38	-56.47	-58.01	-59.51	-60.93	-64.02	-66.52
f_2	Err.1	SCp	2	3	2	0	3	1	4	5
		BIC	20	21	16	11	13	11	10	7
		GAIC	50	40	27	16	19	15	10	11
	MSE	SCp	-53.61	-54.90	-58.91	-57.97	-60.02	-62.41	-65.47	-69.76
		BIC	-53.43	-56.14	-59.50	-58.48	-60.90	-62.32	-65.48	-69.67
		GAIC	-53.23	-54.98	-59.15	-57.59	-60.48	-62.21	-65.70	-69.87
Err.2		SCp	5	11	4	4	5	5	3	8
		BIC	24	24	12	9	5	4	1	2
		GAIC	0	0	3	0	1	1	0	0

Table 8: Type I and Type II errors for Example 3 when SNR_1 = 3.00 dB.

Freq.	N		30	40	50	60	70	80	90	100
f_1	Err.1	SCp	79	17	5	8	3	2	2	1
		BIC	85	23	8	18	17	13	6	4
		GAIC	100	77	52	43	34	29	22	17
	MSE	SCp	-42.86	-45.75	-51.69	-54.32	-55.19	-57.47	-58.26	-60.92
		BIC	-42.75	-46.07	-51.54	-47.16	-55.20	-57.33	-58.22	-60.76
		GAIC	-	-47.81	-52.23	-54.07	-56.13	-46.15	-58.84	-61.14
f_2	Err.1	SCp	79	6	1	0	0	0	0	0
		BIC	84	8	3	1	0	0	0	0
		GAIC	100	62	46	17	4	1	1	0
	MSE	SCp	-48.32	-55.79	-54.60	-59.31	-65.52	-65.26	-65.58	-65.73
		BIC	-50.14	-56.16	-54.60	-59.28	-65.49	-65.24	-65.58	-65.73
		GAIC	-	-55.54	-55.39	-59.32	-65.41	-65.24	-65.59	-65.73
f_3	Err.1	SCp	79	6	1	0	0	0	0	0
		BIC	84	8	3	1	0	0	0	0
		GAIC	100	62	46	17	4	1	1	0
	MSE	SCp	-45.49	-61.30	-58.20	-60.70	-68.24	-70.78	-82.14	-82.14
		BIC	-46.99	-61.45	-58.25	-60.31	-68.24	-70.78	-82.14	-82.14
		GAIC	-	-60.73	-59.16	-60.13	-68.40	-70.90	-82.14	-82.14
Err.2		SCp	0	1	7	6	7	5	7	4
		BIC	0	5	4	3	5	4	4	2
		GAIC	0	0	0	1	0	1	3	2

Table 9: Type I and Type II errors for Example 4 when SNR_1 = -3.00 dB.

APPENDIX

On the derivation of SC formula (10)

To check the conditions for the applicability of the SC formula of Qian and Künsch in our particular case, we resort to the closed-form expression of $\mathbf{J}_N(\boldsymbol{\theta})$ from the equations (4)-(7). We list below the conditions as they are given in [7]:

C1. $\mathbf{J}_N(\boldsymbol{\theta})$ is positive definite.

It is easy to check that all the eigenvalues of $\mathbf{J}_N(\boldsymbol{\xi}_k)$ are strictly positive. The covariance matrix $\mathbf{R}(\mathfrak{a})$ is positive definite for any M [30], therefore $\mathbf{J}_N(\mathfrak{a},\tau)$ is also positive definite, and the condition C1 is verified.

C2. The minimum eigenvalue of $\mathbf{J}_N(\boldsymbol{\theta})$ is of order $O(N)$ as $N \to \infty$.

Two of the eigenvalues of $\mathbf{J}_N(\boldsymbol{\xi}_k)$ are $O(N)$ and the third one is $O(N^3)$. As each eigenvalue of $\mathbf{J}_N(\mathfrak{a},\tau)$ is $O(N)$, we conclude that C2 is satisfied.

C3. $|\mathbf{J}_N(\boldsymbol{\theta}_1)|^{-1}||\mathbf{J}_N(\boldsymbol{\theta}_1)| - |\mathbf{J}_N(\boldsymbol{\theta}_2)|| \leq c||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2||,\ \forall \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$, where c is a finite constant.

For any $\boldsymbol{\theta}$, we have

$$|\mathbf{J}_N(\boldsymbol{\theta})| = N^{5K+M+1} \frac{|\mathbf{R}(\mathfrak{a})|}{2\tau^{M+2}} \prod_{k=1}^{K} |\mathbf{G}(\boldsymbol{\xi}_k, \mathfrak{a})|, \tag{15}$$

which implies that the left-hand-side term in the inequality C3 is finite and it does not depend on N. As the condition C3 is easily verified for $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$, we analyze only the case $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$. Thus we have $\min_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| > \delta$, where δ is given by the precision used to store the values of the parameters. To circumvent some technical difficulties, we consider firstly one sine-wave ($K = 1$) in white noise ($M = 0$). Without loss of generality, we assume $0 < \alpha_{min} < \alpha_1 < \alpha_{max} < \infty$ and $0 < \tau_{min} < \tau < \tau_{max} < \infty$. Elementary calculations lead to the inequality $\max_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2} |\mathbf{J}_N(\boldsymbol{\theta}_1)|^{-1}||\mathbf{J}_N(\boldsymbol{\theta}_1)| - |\mathbf{J}_N(\boldsymbol{\theta}_2)|| < \Delta$, where $\Delta = (\alpha_{max}/\alpha_{min})^4 (\tau_{max}/\tau_{min})^5$. Therefore, condition C3 is verified by selecting $c = \Delta/\delta$. To gain more insight, we assume next $K = 1$ and $M = 1$. As the noise model is stable, the AR coefficient is a non-zero number from the interval $(-1, 1)$. If supplementarily, the precision δ is used to store the value of the AR coefficient, then we get immediately $a_1 \in [-1 + \delta, -\delta] \bigcup [\delta, 1 - \delta]$. Taking α_1 and τ to be bounded as in the white noise case, it is not difficult to show that $\max_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2} |\mathbf{J}_N(\boldsymbol{\theta}_1)|^{-1}||\mathbf{J}_N(\boldsymbol{\theta}_1)| - |\mathbf{J}_N(\boldsymbol{\theta}_2)|| < \Upsilon$, where $\Upsilon = (\alpha_{max}/\alpha_{min})^4 (\tau_{max}/\tau_{min})^5 ((2-\delta)/\delta)^6$. Since $\min_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| > \upsilon$, we choose $c = \Upsilon/\upsilon$ and the condition C3 is verified. We emphasize that the precision used in this proof for the model parameters does not depend on the number of samples N.

C4. $\ln |\mathbf{J}_N(\boldsymbol{\theta})| = o(N)$.

Using the expression (15) for $|\mathbf{J}_N(\boldsymbol{\theta})|$, we readily obtain $\lim_{N\to\infty} \dfrac{\ln |\mathbf{J}_N(\boldsymbol{\theta})|}{N} = 0$, thus C4 is verified.

We apply next the SC criterion from [7]. For simplicity we ignore the terms that do not depend on N, and the SC formula becomes:

$$-\log f(y^N;\hat{\boldsymbol{\theta}}) + \log|\tilde{\mathbf{J}}_N(\hat{\boldsymbol{\theta}},y^N)|^{1/2} + \sum_{i=1}^{3K+M+1}\log(|\hat{\theta}_i|+N^{-1/4}) + \sum_{i=1}^{3K+M+1} r^*(N^{1/4}|\hat{\theta}_i|+1) + O(N^{-1/4}), \tag{16}$$

where $\log(\cdot)$ is the logarithm base 2, $\hat{\boldsymbol{\theta}}$ denotes the ML estimates, and $\tilde{\mathbf{J}}_N(\hat{\boldsymbol{\theta}},y^N) = -\dfrac{\partial^2 \ln f(y^N;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$ is the observed FIM. For any $x>0$, $r^*(x) = \log(\log x) + \log(\log(\log x)) + \cdots$, where the sum continues as long as the iterated logarithms are strictly positive. The approximative formula (10) is obtained from (16) after operating the following changes:

- $\tilde{\mathbf{J}}_N(\hat{\boldsymbol{\theta}},y^N)$ is replaced with $\mathbf{J}_N(\hat{\boldsymbol{\theta}})$.
- An $O\left((3K+M+1)\log\log N\right)$ term is discarded.
- $\log(\cdot)$ is replaced with $\ln(\cdot)$.

Remark 1 The two-step encoding procedure adopted in [7] employs first a uniform quantization of Θ that is performed with the same precision for all the parameters. The term $N^{-1/4}$ in (10) is due to the option from [7] to select this precision based on the minimum eigenvalue of FIM.
Remark 2 It is recommended in [7] to consider in the SC expression also the term given by the number of parameters divided by two and multiplied by $\log\rho$, where ρ is the largest eigenvalue of $\mathbf{J}_N(\hat{\boldsymbol{\theta}})^{-1/2}\tilde{\mathbf{J}}_N(\hat{\boldsymbol{\theta}},y^N)\mathbf{J}_N(\hat{\boldsymbol{\theta}})^{-1/2}$. We prove below that, under mild conditions, ρ does not depend on N, hence we ignore the $\log\rho$ term in (10).

Inspired by the expression of the asymptotic FIM, we assume there exist the non-singular matrices $\mathbf{A}$ and $\mathbf{B}$, and the diagonal matrix $\mathbf{C}_N$ such that $\mathbf{J}_N(\hat{\boldsymbol{\theta}}) = \mathbf{C}_N\mathbf{A}\mathbf{C}_N$ and $\tilde{\mathbf{J}}_N(\hat{\boldsymbol{\theta}},y^N) = \mathbf{C}_N\mathbf{B}\mathbf{C}_N$. Supplementarily all the diagonal entries of $\mathbf{C}_N$ are powers of N, and the entries of $\mathbf{A}$ and $\mathbf{B}$ do not depend on N. With the notation $\mathbf{Z} = \mathbf{J}_N(\hat{\boldsymbol{\theta}})^{-1/2}\tilde{\mathbf{J}}_N(\hat{\boldsymbol{\theta}},y^N)\mathbf{J}_N(\hat{\boldsymbol{\theta}})^{-1/2}$, we have $\mathbf{Z} = \mathbf{J}_N(\hat{\boldsymbol{\theta}})^{-1/2}\left(\tilde{\mathbf{J}}_N(\hat{\boldsymbol{\theta}},y^N)\mathbf{J}_N(\hat{\boldsymbol{\theta}})^{-1}\right)\mathbf{J}_N(\hat{\boldsymbol{\theta}})^{1/2}$, thus $\mathbf{Z}$ and $\tilde{\mathbf{J}}_N(\hat{\boldsymbol{\theta}},y^N)\mathbf{J}_N(\hat{\theta})^{-1}$ are similar. Moreover, $\tilde{\mathbf{J}}_N(\hat{\boldsymbol{\theta}},y^N)\mathbf{J}_N(\hat{\boldsymbol{\theta}})^{-1} = \mathbf{C}_N\mathbf{B}\mathbf{A}^{-1}\mathbf{C}_N^{-1}$, which leads to the conclusion that $\mathbf{Z}$ and $\mathbf{B}\mathbf{A}^{-1}$ are also similar. As the eigenvalues of $\mathbf{B}\mathbf{A}^{-1}$ do not depend on N, ρ is also independent of N.

References

[1] S. M. Kay, *Fundamentals of statistical signal processing: estimation theory.* Prentice Hall, 1993.

[2] M. Ghogho and A. Swami, "Fast computation of the exact FIM for deterministic signals in colored noise," *IEEE Trans. Signal. Proces.*, vol. 47, no. 1, pp. 52–61, Jan. 1999.

[3] L. Kavalieris and E. Hannan, "Determining the number of terms in a trigonometric regression," *Journal of time series analysis*, vol. 15, no. 6, pp. 613–625, 1994.

[4] J. Li and P. Stoica, "Efficient mixed-spectrum estimation with applications to target feature extraction," *IEEE Trans. Signal. Proces.*, vol. 44, no. 2, pp. 281–295, Feb. 1996.

[5] P. Stoica and A. Nehorai, "Statistical analysis of two nonlinear least-squares estimators of sine-wave parameters in the colored-noise case," *Circuits, Systems, and Signal Processing*, vol. 8, no. 1, pp. 3–15, 1989.

[6] P. Stoica and Y. Selen, "A review of information criterion rules," *IEEE Signal. Proces. Mag.*, vol. 21, no. 4, pp. 36–47, Jul. 2004.

[7] G. Qian and H. Künsch, "Some notes on Rissanen's stochastic complexity," *IEEE Trans. Inf. Theory*, vol. 44, no. 2, pp. 782–786, Mar. 1998.

[8] J. Rissanen, "Fisher information and stochastic complexity," *IEEE Trans. Inf. Theory*, vol. 42, no. 1, pp. 40–47, Jan. 1996.

[9] ——, "MDL denoising," *IEEE Trans. Inf. Theory*, vol. 46, no. 7, pp. 2537–2543, Nov. 2000.

[10] A. Barron, J. Rissanen, and B. Yu, "The minimum description length principle in coding and modeling," *IEEE Trans. Inf. Theory*, vol. 44, pp. 2743–2760, Oct. 1998.

[11] J. Rissanen, *Information and complexity in statistical modeling*. Springer Verlag, 2007.

[12] B. Friedlander and B. Porat, "The exact Cramer-Rao bound for Gaussian autoregressive processes," *IEEE Tr. on Aerospace and Electronic Systems*, vol. AES-25, pp. 3–8, 1989.

[13] C. Giurcăneanu and J. Rissanen, "Estimation of AR and ARMA models by stochastic complexity," in *Time series and related topics.*, H.-C. Ho, C.-K. Ing, and T. L. Lai, Eds. Institute of Mathematical Statistics Lecture Notes-Monograph Series, 2006, vol. 52, pp. 48–59.

[14] C. Giurcăneanu, "Stochastic complexity for the estimation of sine-waves in colored noise," in *Proc. of 2007 Int. Conf. on Acoustics, Speech, and Signal Processing*. Honolulu, Hawaii, USA: IEEE, Apr. 2007, vol. 3, pp. 1097–1100.

[15] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, Mar. 1978.

[16] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[17] P. M. Djuric, "A model selection rule for sinusoids in white Gaussian noise," *IEEE Trans. Signal. Proces.*, vol. 44, no. 7, pp. 1744–1751, Jul. 1996.

[18] B. G. Quinn, "Estimating the number of terms in a sinusoidal regression," *Journal of time series analysis*, vol. 10, no. 1, pp. 70–75, 1989.

[19] S. Kay, "Conditional model order estimation," *IEEE Trans. Signal. Proces.*, vol. 49, no. 9, pp. 1910–1917, Sep. 2001.

[20] ——, "Exponentially embedded families-new approaches to model order estimation," *IEEE Trans. on Aerospace and Electronic Systems*, vol. 41, no. 1, pp. 333–345, Jan. 2005.

[21] A. Neath and J. Cavanaugh, "Regression and time series model selection using variants of the Schwarz information criterion," *Communications in Statistics - Theory and Methods*, vol. 26, pp. 559–580, 1997.

[22] A. Hanson and P. C.-W. Fu, "Aplications of MDL to selected families of models," in *Advances in Minimum Description Length: theory and applications*, P. Grünwald, I. Myung, and M. Pitt, Eds. MIT Press, 2005, ch. 5, pp. 125–150.

[23] E. Liski, "Normalized ML and the MDL principle for variable selection in linear regression," in *Festschrift for Tarmo Pukkila on his 60th birthday*, E. Liski, J. Isotalo, J. Niemelä, S. Puntanen, and G. Styan, Eds. Univ. of Tampere, 2006, pp. 159–172.

[24] M. Hansen and B. Yu, "Minimum description length model selection criteria for generalized linear models," in *Science and statistics: a festchrift for Terry Speed*, D. Goldstein, Ed. Institute of Mathematical Statistics Lecture Notes-Monograph Series, 2002, vol. 40, pp. 145–164.

[25] E. George and D. Foster, "Calibration and empirical Bayesian variable selection," *Biometrika*, vol. 87, no. 4, pp. 731–747, 2000.

[26] S. Bruzzone and M. Kaveh, "Information tradeoffs in using the sample autocorrelation function in ARMA parameter estimation," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. ASSP-32, no. 4, pp. 701–715, Aug. 1984.

[27] A.-K. Seghouane and M. Bekara, "A small sample model selection criterion based on Kullback's symmetric divergence," *IEEE Trans. Signal. Proces.*, vol. 52, pp. 3314–3323, 2004.

[28] M. Bekara, L. Knockaert, A.-K. Seghouane, and G. Fleury, "A model selection approach to signal denoising using Kullback's symmetric divergence," *Signal Processing*, vol. 86, pp. 1400–1409, 2006.

[29] J. Bernardo, "Psi (digamma) function," *Appl. Statist.*, vol. 25, pp. 315–317, 1976.

[30] P. Stoica and R. Moses, *Introduction to Spectral Analysis*. New Jersey: Prentice-Hall, 1997.

A Stochastic Complexity Perspective of Induction in Economics and Inference in Dynamics

K. Vela Velupillai*
Department of Economics†
University of Trento
Via Inama, 5
I-381 00 Trento
Italy
&
Girton College
Cambridge CB3 0JG
UK

April 7, 2008

*I am very particularly indebted to Francesco Luna, John McCall, Shu-Heng Chen and Stefano Zambelli. They encouraged my early and continuing interest in Jorma Rissanen's work from their own particular viewpoints of Bayesian statistics, Gold's model of learning, de Finetti's theory of probability, *MDL* and Chaitin's algorithmic information theory. However, they are not responsible for any of the remaining infelicities. My own view was, from the outset, shaped by my determination to fashion the subject of computable economics on the foundations of recursion theoretic and constructive mathematics - i.e., on the works of Turing, Kolmogorov, Brouwer and Bishop. Thus, I came to stochastic complexity, as Rissanen originally did, from 'Kolmogorov complexity' theory.

†Corresponding e-mail address: kvelupillai@gmail.com

Abstract

Rissanen's fertile and pioneering *minimum description length* principle (*MDL*) has been viewed from the point of view of statistical estimation theory, information theory, as stochastic complexity theory[1] – i.e., a computable approximation to Kolomogorov Complexity – or Solomonoff's recursion theoretic induction principle or as analogous to Kolmogorov's sufficient statistics. All these – and many more – interpretations are valid, interesting and fertile. In this paper I view it from two points of view: those of an algorithmic economist and a dynamical system theorist. From these points of view I suggest, first, a recasting of Jevons's sceptical vision of induction in the light of MDL; and a complexity interpretation of an undecidable question in dynamics.

[1] I am using 'stochastic complexity' in a kind of 'generic' way. Rissanen has, over the past three decades, gradually refined the exact formal meaning of the phrase and I believe his most mature views are now represented in [1]. The kind of meaning I have in mind is what I learned from Rissanen's early writings on MDL, for example in [24], p.1080, emphasis in the original:

> *"[I]f ... a shortest description of the data, to be called* STOCHASTIC COMPLEXITY *is found in terms of the models of a selected class, there is nothing much further anyone can teach us about the data; we know all there is to know."*

1 A Personal Preamble

> To paraphrase the famous reply of Laplace to Napoleon, who wondered why the word 'God' did not appear in *Mécanique Céleste*, we could state that 'the assumption of a 'true' distribution is not needed in this theory'.
>
> Jorma Rissanen

I have shared many moments of intellectual and personal splendour with Jorma Rissanen. One serendipitous conjunction relates to my first published, technical, article, which was in 1978, in Volume 14 of *Automatica*, [27]. I did not, of course, know, then, that Jorma Rissanen's first published, pioneering, paper on *stochastic complexity* – or, the *Minimum Description Length* principle (henceforth, *MDL*) – was also in that same volume of the same *Journal*[2]! Justifiably, that paper on *MDL* spawned a path-breaking research program that has, in one strand, developed into *Algorithmic Statistics*. It gives me great and undiluted pleasure to state that my own, much humbler piece, in that same volume of Automatica, was also the fountainhead for what I have developed into the research program on *Algorithmic Economics*!

In the intervening 30 years, particularly in its second half, I have had the pleasure and privilege of hosting Jorma Rissanen at numerous venues, exotic and otherwise, trying to make his fertile and fascinating research program more familiar to obdurate economists. I believe a measure of success can now be seen, albeit taking place at snail's pace.

I had read, quite by chance (sic!), an expository piece on *stochastic complexity* in an issue of the *IBM Research Magazine*[3] around the time I was trying to establish the *Center for Computable Economics* (*CCE*), in the department of economics, at UCLA, in the academic year 1990-91. The modest initial success, together with funds for a seminar series on Computable Economics, gave me the chance to invite Jorma Rissanen to give a talk at the *CCE* seminar series, as one of its first speakers, in autumn, 1991. Soon after that I organised a 'Summer School' in Computable Economics, in July, 1992, sponsored by Aalborg University in Denmark, at the beautiful *Dronninglund Slot* in Nordjylland. Naturally, Jorma Rissanen was one of the key speakers at that event.

[2] Rissanen's classic was published in the September issue of Volume 14, 1978 and mine in the November issue of the same Volume ([23], [27]). Mine had been presented an an IFAC meeting in Vienna the year before. My path towards what I now call *Algorithmic Economics* began with *computational complexity theory*. I like to think there is a further *serendipity* even here: one strand of the tradition from which Jorma Rissanen created MDL arose from *algorithmic complexity theory*, as is well documented in several articles tracing his thought on these matters.

[3] The 'expository' piece, by Rowan Dordick of IBM's 'communications department', [7], had an eye-catching title – *Understanding the 'go' of it*, quoting Maxwell – and an attractive blurb, (with a photograph of Jorma Rissanen at the blackboard (those were the days...!), which said:

> "A novel approach to statistical inference – the theory of stochastic complexity – holds that the best description of data is the shortest one. "

Most recently I set up the *Computable and Behavioural Economics Research Axis*[4] (*COBERA*) in the department of economics at the National University of Ireland, in Galway. One of the first events sponsored by COBERA was a 'Spring School' on Computable Economics, in March, 2005. Naturally, Jorma Rissanen was again one of the key lecturers at this event, too.

In all of the above events the audience was predominantly made up of advanced graduate students, senior and junior faculty and interested outside participants, almost all of whom were economists. However, the distinguished speakers – like Jorma Rissanen – were not all economists; apart from Jorma Rissanen, there were recursion theorists (Piergiorgio Odifreddi, F.A. Doria), algorithmic information theorists (Greg Chaitin), game theorists (Ken Binmore), dynamical system theorists (Ralph Abraham, Joe McCauley), and others, all of whom were united by being motivated by an algorithmic approach to theory and application in the sciences, both pure and applied.

Jorma Rissanen was always a persuasive lecturer and an engaging participant at all of these events. Economists of widely varying persuasions – in statistical methodology and mathematical epistemology – were always fascinated by his wonderful lectures, always prepared with utmost care and delivered with immaculate clarity. On many occasions his talks were interrupted by genuinely perplexed members of the audience who were struggling to absorb a whole new set of concepts with which to understand a fascinating framework and methodology. On occasions there was also one or another famous, but obdurate, economists, entrenched in orthodoxy, who was unable to dissociate himself from the traditional frameworks that shackled his thoughts and practice.

I would like to end this brief personal preamble with a pleasant recollection of an event that I have had occasion to repeat almost every time I have chaired a session where Jorma Rissanen has been the lecturer. On this particular occasion, after Jorma Rissanen's beautifully crafted lecture on stochastic complexity and statistical estimation, the following brief dialogue occurred between a very distinguished game theorist (referred to as DGE), not known for any competency in statistical methodology, and Jorma (JR):

DGE: (In an irritated tone), 'You seem to be suggesting that your method is the only one around. You must know that there are many other methods, and some have survived the test of time, too.'

Pin-drop silence in the lecture hall (at Dronninglund Slot).

JR: (In a perfectly calm and conciliatory tone), 'Oh, I am so sorry; I did not mean to suggest that MDL was the only statistical method around. I do apologise if I gave that impression.'

Pause and continued silence in the lecture hall; not even a whisper or a murmur among the distinguished collections of lecturers and auditors, among whom were very famous economists like Bob Clower, Axel Leijonhufvud, Michael Intrilligator, John McCall; computer scientists like Berc Rustem, Greg Chaitin and Piergiorgio Odifreddi; and so on.

Then, after only a brief pause, which seemed like eternity:

[4] Now defunct.

JR: '*But it is the best* [*method available*]!'

The whole hall erupted in appreciative and almost unanimous (i.e., except one member of the audience!) laughter and applause.

Jorma Rissanen continues a distinguished Finnish tradition of making Induction a scientifically respectable enterprise, free of the nihilistic scepticism propagated by ill-informed scholars of Hume and Mill[5], particularly in economics and the philosophy of science. His great predecessors and contemporaries in the rich *Finnish tradition* of the mathematical epistemology of induction are, *among others*, of course Georg Henrik von Wright, Jaako Hintikka, Ilkka Niniluoto and Risto Hilpinen[6]. In my own economics education at Cambridge in the early 1970s, under the inspiring supervision of Richard Goodwin, I was advised, wisely as it turned out later, to attend the lectures given by Ian Hacking in the philosophy department. Fortunately, Hacking was just then lecturing, broadly, on issues of induction and probability and, of course, the works of Wittgenstein's immediate successor as the Knightbridge Professor of Philosophy at Cambridge, Georg Henrik von Wright, were often brought into focus. Margaret Anscombe was often at those lectures and, occasionally, a brief dialogue took place between Hacking and Anscombe, to which we – students – were privileged auditors.

It is a pleasure and a privilege to pay homage to a pioneer scientist of uncompromising integrity and undiluted personal warmth.

The paper is divided into four subsequent sections. A brief methodological discussion, of lessons learned from Rissanen's modelling philosophy, is the content of the next section. In section 3, the main, substantively economic section of the paper, I try to reinterpret a celebrated sceptical – even hostile – vision of inductive inference by one of the pioneers of modern economic theory from the point of view of MDL. In section 4, motivated by an issue in economic dynamics, I try to pose an undecidable problem in dynamical systems theory as an inference problem and formulate its Kolmogorov complexity. The concluding section 5 consists of speculative thoughts on Algorithmic Economics as a companion in arms of Algorithmic Statistics, Algorithmic Randomness and Algorithmic Information Theory.

[5] I have in mind, in particular, Jevons in eocnomics and Popper in the philosophy of science. What I ahve to say about Jevons is given in section 3, below; I have had my say on Popper, *from the point of view of MDL* in [30]. Li and Vitanyi quite pungently, but accurately (I think) note, [31], p.448:

> "Unsatisfactory solutions [to the problem of scientific inference] have been provided by philosophers like R.Carnap and K.Popper."

I suppose they should ahve been a little more precise and designated these two worthy individuals as 'philosophers of science'! In any case, my case against Popper from the point of view of MDL, substantiating the Li-Vitanyi claim, is fully described and discussed in detail in [30].

[6] von Wright's magisterial exposition of induction in the probabilistic tradition is in [32] & [33]; for Hintikka's views (and Niniluoto's) , in a Carnapian tradition, the best source may well be, [11] & [12]. A different source for Niniluoto's work on induction is, of course, his early joint monograph with another distinguished Finn, Tuomela, in, [21]. One reference for Hilpinen's work on inductive logic is [9].

2 Extracting Methodological Precepts for *Algorithmic Economics* from Rissanen's Modelling Philosophy

> "Regarding the ultimate model, no algorithmic procedure to find it can exist, as shown in the theory of the *algorithmic complexity*, Solomonoff (1964), Kolmogorov (1965), Chaitin (1973), which also *is the spiritual father of our main notion.*"
>
> Jorma Rissanen, [25], p.224; italics added.

Rissanen's philosophy of stochastic complexity suggests a way of exorcising the search for that traditional 'Will o' the Wisp' in formal modelling exercises: the '*true*' model underpinning observable, empirical data. Secondly, in one of its recent incarnations, the modelling philosophy of stochastic complexity has evolved into *algorithmic statistics*. As defined by the three pioneers, *algorithmic statistics* is the theory of the '*relation* between an individual data sample and an individual model summarizing the information in the data', [8], p. 2443. In this theory the search is for an 'absolute notion' of such a 'relation' in analogy with the way 'Kolmogorov complexity is the accepted absolute measure of information content of an individual finite object' (*ibid*). Thirdly, the concept of universality – either of the Universal Turing Machine in recursion theory, or of the prior in the Solomonoff scheme or of models in Rissanen's recent work on stochastic complexity.

Finally, I want to return to one of the earliest insights and interpretations of 'stochastic complexity' as a computable approximation of the uncomputable Kolmogorov complexity (or, equivalently, of Solomonoff's uncomputable 'universal prior'). The orientation of my own research in algorithmic economics has been almost entirely determined by this particular insight. Therefore, let me, touch on this point, very briefly, before going on to the main sections of the paper.

In his original paper introducing the stochastic complexity approach to statistical inference as inductive inference from finite data sequences, Rissanen acknowledged his indebtedness to Kolmogorov ([18], p. 465). It is generally understood, by scholars who have closely studied the origins and evolution of Rissanen's ideas on stochastic complexity, that this horn of the original motivation – the other being Akaike's *AIC* model[7] – led to the idea of stochastic complexity being a computable approximation to the uncomputable Kolmogorov complexity. In this original paper, Kolmogorov defined the notion that has forever since

[7] Even with some reservations, Rissanen is handsome in his acknowledgement to Akaike, [25], p.224:

> "[W]e are indebted to Akaike's pioneering and innovative work for inspiration in our own efforts."

then been associated with his name in the following way:

$$K_\phi(y|x) = \begin{cases} \min_{\phi(p,x)=y} l(p) \\ \infty \nexists\ p \text{ s.t } \phi(p,x) = y \end{cases} \tag{1}$$

where:

$\phi(p,x) = y$: a partial recursive function – the 'method of programming' – associating a (finite) object y with a *program* p and a (finite) object x.

Kolmogorov went on to observe, crucially, that (ibid, pp.299-300):

> "[T]he function $K_\phi(y|x)$ need not be effectively computable (generally recursive) even if it is *a fortiori* finite for any x and y."

Remark 1 *The proof that $K_\phi(y|x)$ is nonconstructive, freely appealing to tertium non datur. I consider this an infelicity. But since it is not an existence proof, rectifying the infelicity by a constructive proof may not be essential*

To the best of my knowledge most proofs of the uncomputability of $K_\phi(y|x)$ are based on the unsolvability of the Halting problem for Turing Machines[8]. Shortly after Kolmogorov's above paper was published, Zvonkin and Levin, [34], p.92, Theorem $1.5, b$, provided the result and proof that rationalises the basic principle of stochastic complexity providing the computable approximation to the uncomputable $K_\phi(y|x)$ .The significant relevant result is:

Theorem 2 *Zvonkin-Levin*

$\exists$ *a general recursive function $H(t,x)$, monotonically decreasing in t, s.t :*

$$\lim_{t\to\infty} H(t,x) = K_\phi(y|x) \tag{2}$$

Remark 3 *This result guarantees, the existence of 'arbitrarily good upper estimates' for $K_\phi(y|x)$, even although $K_\phi(y|x)$ is uncomputable. I am not sure this is a claim that is constructively substantiable[9]. How can a noncomputable function be approximated? If any one noncomputable function can be approximated uniformly, then by 'reduction' it should be possible, for example, to 'approximate', say, the Busy Beaver function. I suspect an intelligent and operational interpretation of the Zvonkin-Levin theorem requires a broadening of the notion of 'approximation'.*

[8] For example in [6], §7.7, pp.162-8. Incidentally, the section on *Models of Computation* (§7.1, pp.146-7), in this book is quite unreliable and strange, to put it mildly. The presentation of the genesis of the Turing Machine and Church's Thesis are both incorrect to the point of being absurd.

[9] My view on tis further strengthened by some of the remarks in [6], particularly, p.163, where one reads (italics added):

> "The shortest program is not computable, although as more and more programs are shown to produce the string, the estimates from above of the Kolmogorov complexity *converge to the true Kolmogorov complexity*, (the problem, of course, is that *one may have found the shortest program and never know that no shorter program exists*).

These remarks border on the metaphysical! How can one approximate to a true value which cannot be known, by definition?

Universality, (*approximate*) *computability*, *data compression*, the eschewing of '*truth*' (in model selection) – these are, in my reading, the four fundamental building blocks of Rissanen's methodology. They form the methodological building blocks of algorithmic economics, which in earlier writings I called *Computable Economics* (cf. [28]).

3 Re-reading Jevons in the Light of MDL

> "Doubtless there is in nature some invariably acting mechanism, such that from some fixed conditions an invariable result always emerges. But we, with our finite minds and short experience, can never penetrate the mystery of these existences .... . We are in the position of *spectators who witness the production* of a complicated machine, but are not allowed to examine its structure. *We learn what does happen and what does appear,* but if we ask for the reason, the answer would involve an infinite depth of mystery."
>
> [13], p.222; italics added.

William Stanley Jevons, a pioneer of neoclassical economics was implacably opposed to the inductive method. His methodological precepts against the inductive method were cogently presented in his monumental treatise on *The Principles of Science* (*ibid*, henceforth referred to as *TPOS*). However, a close reading of its almost 800 pages, against the backdrop of some knowledge of the principles underpinning the MDL principle has convinced me that the Jevonian opposition to the inductive method is untenable. In this section a sketch of my re-interpretation of *TPOS* as a treatise supporting what I have in earlier writings called *The Modern Theory of Induction* ([28], Chapter 5) is outlined.

3.1 Background

> "What especially characterised Jevons's view of logical method was the prominence he attached to the combination of formal and empirical principles *through the inverse application of the theory of probability*"
>
> [14], p.638; italics added

TPOS, a book of almost 800 dense pages refers to almost every known Western natural philosopher without, however, a single mention of *William of Ockham*, *Occam's Razor* or *Ockham's Principle*[10]! The closest he gets to anything like a (dismissive) mention of *Occam's Razor* is when he rejects Newton's *Rule 1*

[10] The general literature seems to refer to William of Ockham but Occam's Razor; hence I retain this schizophrenia in my own spelling. Furthermore, Ockham's own most often stated version of the principle named after him seems to have been: *'Pluralitas non est ponenda sine necessitae'* – plurality is not be posited without necessity.The more commonly attributed version: *'Entia non sunt multiplicanda sine necessitate'* – entities must not be multiplied without necessity – appears not to have been used by him (cf. [3], p.xxi).

for Natural Philosophy in the *Principia* as irrelevant for any inductive purpose, let alone for acting as an anchor to eliminate *inductive indeterminacy*:

> "It is by false generalisation, again, that the laws of nature have been supposed to possess that perfection which we attribute to *simple* forms and relations. ... Newton seemed to adopt the questionable axiom that nature always proceeds in the *simplest* way; in stating his first rule of philosophising, he adds: 'To this purpose the philosophers say, that nature does nothing in vain, when less will serve; for nature is pleased with *simplicity*, and affects not the pomp of superfluous causes.'..... *Simplicity* is naturally agreeable to a mind of limited powers, but to an infinite mind all things are simple."
>
> *TPOS*, p.625; italics added.

Is Jevons suggesting, in the context of his times, beliefs and traditions, that the omnipotence and omniscience of the architect of the laws of nature – the designer of the 'complicated machine' – are such that we are as likely to witness the 'productions of a complicated machine' as to a simple one[11]. Jevons may have been trying to make the point that Newton's was a metaphysical assumption and that we have no grounds for assuming anything about structure in the absence of empirical evidence to the contrary[12]. However, Jevons, who was almost as obsessed with consistency as he was with deduction[13], did not obey his own precepts when it came to choosing the order and degree of equations to fit observed data. In such an example he argues clearly in favour of choosing the *simplest* hypothesis, at least in the first instance:

> "It is *a general rule* in quantitative investigation that we commence by discovering linear, and *afterwards* proceed to elliptic or more *complicated* laws of variation."
>
> *TPOS*, p.474; italics added.

Perhaps, given the times and context, one can be generous to Jevons – more generous than he was to Newton and more, also, than Marshall was to Jevons – and suggest that he was doubtful about any reliance on Occam's Razor because he did not feel it possible to give a rigorous, invariant, analytical definition of *simplicity*. I think, therefore, it may be reasonable to assume, counterfactually, that Jevons would have accepted the use of Occam's Razor in hypothesis selection and inductive inference had it been possible to demonstrate that it was

[11] I cannot but reflect on Einstein's wise maxim when faced with the Great Scorer's devises, 'Subtle is the Lord, but malicious he is not' – Raffiniert ist der Herrgott aber boshaft ist er nicht – and wish Jevons had shown some humility in the face of Einstein's undisputed predecessor's, i.e., Newton's, own methodological maxims.

[12] The Einsteinian example of the way he reasoned his way towards the general theory of relativity from the special theory is clearly described and discussed by Kemeny, [15]. This example is paradigmatic, of the role of *simplicity* in hypothesis selection and formation, in the logic of scientific practice.

[13] He would, surely, find it uncomfortable to live in a post-Gödelian world where consistency has been dethroned from its crowing place in the deductive enterprise!

possible to define, rigorously, the notion of simplicity. After Solomonoff – not a little influenced by Keynes – and Rissanen, re-reading Jevons and substantiating a rigorous method of inductive inference is not the most difficult task for a philosophy of science. This will be attempted in the last sub-section of this section, after first summarising the Jevonian vision of inductive indeterminacy in the next sub-section.

3.2 The Jevonian Vision on Induction and its Indeterminacy

> "Combining insight and error, he spoilt brilliant suggestions by erratic and atrocious arguments. His application of inverse probability to the inductive problem is crude and fallacious, but the idea which underlies it is substantially good. ... There are few books, so superficial in argument yet suggesting so much as Jevons's *Principles of Science*."
>
> Keynes, [17], p.204

I shall summarise, rather telegraphically and in an inelegant numbered-list format, Jevons's precepts on inductive inference. This will, then, enable me to refer to them conveniently in the next section when a simple case is made to encapsulate the Jevonian vision in the modern inductive fold.

The following twelve points summarise, however audacious the task of encapsulating summarily, a sustained criticism of the inductive method, spread over a discursive book of almost 800 pages (all quotes in this list are from *TPOS*):

1. "The theory of inductive inference stated [in *TPOS*] was suggested by the study of the Inverse Method of Probability." (p.265)
2. Induction is the inverse operation of deduction. (p.121)
3. Induction is *perfect* when an enumeration of all possible instances of the phenomenon under consideration is feasible, at least in principle. (pp.146-7)
4. Induction is *imperfect* in case the 'enumeration', as in (3), is infeasible.
5. The results of imperfect induction are, therefore, never more than probable:

   "Only in proportion as our induction approximates to the character of *perfect induction*, does it approximate to certainty. The amount of uncertainty corresponds to the probability that other objects than those examined may exist and *falsify our inferences*; ...". (p.229; italics added)
6. The number of instances of any inductive phenomenon is, at most, denumerably infinite; and the number of alternative hypotheses that may be entertained to account for any given inductive phenomenon is, at most, denumerably infinite.

7. Inductive processes are those, and only those, that generate general laws such that the hypothesis underlying them 'yield deductive results in accordance with experience.'

8. "That process only can be called induction which gives general laws, and it is by the subsequent employment of deduction that we anticipate particular events. .... I hold that in all cases of inductive inference we must invent hypotheses, until we fall upon some hypotheses which yields deductive results in accordance with experience." (p.226-8)

9. The extraction of general laws, from a denumerably infinite set of plausible hypotheses, proceeds by way of applying the 'inverse method of probability' (i.e., using Bayes's Rule):

   "[I]n all cases ... of inductive inference where we seem to pass from some particular instances to a new instance, we ... form an hypothesis as to the logical conditions under which the given instances might occur; we calculate inversely the probability of that hypothesis and compounding this with the probability that a new instance would proceed from the same conditions, we gain the absolute probability of occurrence of the new instance in virtue of this hypothesis. But as several, or many, or even an infinite number of mutually inconsistent hypothesis may be possible, we must repeat the calculation for each such conceivable hypothesis, and then the complete probability of the future instances will be the sum of the separate probabilities." (p.268)

   This description indicates tat Jevons's inductive method, despite its rhetoric about being simply 'the inverse of deduction', is nothing other than a simple Bayesian procedure.

10. However, there is no rule or uniform principle on the basis of which it is possible to assign priors to implement 'the inverse method of probability' in the mechanical way in which deductive rules can be applied:

    "To assign the antecedent probability of any proposition, may be a matter of difficulty or impossibility, and one with which logic and the theory of probability have little concern." (p.211-2)

11. "All logical inference involves classification [and it] is not really distinct from the process of perfect induction. [But] there will be no royal road to the discovery of *the best system* and it will even be *impossible to lay down the rules of procedure* to assist those who are in search of good arrangement." (pp.673-90; italics added)

12. The Ramean Tree (pp.702-3), is an encapsulation of the exhaustive method of classification.

3.3 Disciplining Jevonian Inductive Indeterminacies in a Post-Solomonoff MDL World

> "[T]he *most probable* cause of an event which has happened is that which would *most probably* lead to the even supposing the cause to exist."
>
> *TPOS*, p.243; italics added.

I claim that '*most probable*', in the above Jevonian sense of being encapsulated within the inverse probability framework, is equivalent to the precise *recursion theoretic inductive inference* concept of *simplest* and it removes, effectively, the much vaunted *indeterminacy of induction*. The fundamental notion of the modern theory or recursion theoretic induction can be stated as the following proposition:

Proposition 4 *An event with the highest probability of occurring is also that which has the simplest description*

Let me give a brief and elementary sketch of the kind of analysis that makes such an equivalence possible – i.e., to be able to use Occam's Razor to eliminate the indeterminacy in the 'inverse probability' method, correctly identified by Jevons. Consider a standard version of Bayes's rule:

$$P\left(H_i|E\right) = \frac{P\left(E|H_i\right)P\left(H_i\right)}{\sum_i P\left(E|H_i\right)P\left(H_i\right)} \tag{3}$$

Where, apart from absolutely standard, textbook interpretations of all variables and notations, the only implicit novelty – for a Jevonian vision – is the assumption of a denumerable infinity of hypotheses (i.e., above, §3.2:(6)-(7)). This, in a standard inverse probability exercise, E, the class of 'observed' events, and $P(H_i)$ are given; Jevons's inductive inference problem is, then, to find the 'most probable' H_i that would 'most probably' lead to the observed event of relevance. To get the perspective I want, rewrite (3) as:

$$-\log P\left(H_i|E\right) = -\log P\left(E|H_i\right) - \log P\left(H_i\right) + \log P(E) \tag{4}$$

where the last term on the r.h.s of (4) is a shorthand expression for the denominator in (3) which, in turn, is the normalising factor in such inverse probability exercises.

Now, finding the Jevonian 'most probable hypothesis' is equivalent to determining that H_i, w.r.t which (4) is *minimised*. However, in (4), $\log P(E)$ is invariant w.r.t H_i; hence the problem is to minimise (w.r.t., H_i) :

$$-\log P\left(E|H_i\right) - \log P\left(H_i\right) \tag{5}$$

However, it is clear that the problem of indeterminacy remains so long as we do not have a principle on the basis of which the *prior* cannot be assigned *universally*.

Recall, now, that the Jevonian inductive enterprise is supposed to interpret a class of observations, events, data, etc., – 'the production of a complicated machine' – in terms of a denumerable infinity of hypotheses, in such a way that a general law is formalised from which, by deductive processes, the outcomes with which one began are generated (cf. above, §3.2, (2), (7)-(9)). These entities are formalised – in pre-set theoretic days – in terms of logical and mathematical formulas. As far as the requirements of the logic of the inductive method recommended in *TPOS* is concerned, we need only formalise, at most, a denumerable infinity of outcomes in an observation space, and there is a similar quantitative upper bound for the number of hypotheses. Thus the space of computable numbers is sufficient for this formalisation exercise.

Suppose, now, that every element in the outcome space and every potential hypothesis – being denumerably infinite – is associated with a positive integer, perhaps ordered lexicographically. In *TPOS* every outcome and every hypothesis is framed as a logical proposition. Every such proposition can, therefore, be assigned one of the computable numbers and they, in turn, can be processed, say, by a Turing Machine. Next, the *binary codes* for the assigned computable numbers can be constructed, and thereby they can also be given a precise quantitative measure in terms of their counts in *bits*. Thus the basic result of modern recursion theoretic inductive inference, summarised in the above proposition, results from the following *Rissanen Rule of MDL Inductive Inference*:

Proposition 5 *Rule of Induction*[14]

The 'best theory' is that which minimizes the sum of:

(a). The length, in bits, of the number theoretic representation of the denumerable infinity of hypothesis;

(b). The length, in bits, of the elements of the space of outcomes (also, by assumption, at most, denumerably infinite);

The conceptual justification for this 'rule' as the underpinning for Proposition 4 is something like the following reasoning> if the elements of the observation space (E) have any *patterns* or *regularities*, then they can be encapsulated in a *law*, on the basis of some *hypothesis*. The idea that the *best law* is that which can extract and summarise the *maximum amount of regularities* or patterns in E and represent them *most concisely* captures the workings of Occam's razor in an inductive exercise. In homely terms: if two hypotheses can encapsulate the patterns in the data, then choose the more concise one.

The final link in this inductive saga is a universal formula for the prior in the inverse probability exercise.

Proposition 6 $\exists$ *a probability measure* $m(.)$ *that is universal (in the sense of being invariant except for an inessential additive constant) such that:*

$$\log_2 m(.) \approx K(.) \tag{6}$$

[14] The problem of summing an infinite sum has to be resolved by some kind of standard normalization procedure in the case, as here, of denumerable infinity of hypotheses. I shall ignore this detail here.

where, $K(.)$: the Kolmogorov complexity of the best theory generated in the implementation of the rule of induction.

I think this closes the circle consistently with the aims set forth in *TPOS* for an inductive exercise. Thus, I rest my case for Jevons, after Solomonoff-Rissanen, as an inductivist.

4 Complexity of an Undecidable Inference in a Dynamical System

> "[T]he question of the decidability of the Mandelbrot set has another justification. It can partly answer and give insight to the question: can one *decide* if a differential equation is chaotic?"
>
> [2], p.5; italics added.

I have had to tackle formal undecidabilities in economic dynamics. One of the formal proposition I have derived in economic dynamics relates to the non-effectivity of policy in a complex dynamic economy. In trying to resolve some dissatisfaction with this result, I have been influenced by some of Rissanen's methodological precepts. An outline of a result, the framework and some conjectures are given in this section.

One of the keys to Rissanen's inference methodology lies in eschewing the search for 'true' models that give rise to observable phenomena which have to be explained. Taking a cue from such a methodology I want to pose the following problem: given the observables of a dynamical system, is it possible to *infer* interesting properties that characterise its basins of attraction? In view of Rice's theorem in classical recursion theory – or, alternatively, due to the ubiquity of the unsolvability of the Halting Problem for Turing Machines – it is often impossible to infer whether observable data is sufficient to decide membership in a set, unless the set is characterised trivially.

Let me first provide the formal background in a general way.

I shall have to assume familiarity with the formal definition of a dynamical system (cf. for example, the obvious and accessible classic, [10] or the more modern, [4]), the necessary associated concepts from dynamical systems theory and all the necessary notions from classical computability theory (for which the reader can, with profit and enjoyment, go to a classic like [26] or, at the frontiers, to [5]). Just for ease of reference the bare bones of relevant definitions for dynamical systems are given below in the usual telegraphic form[15]. An intuitive understanding of the definition of a 'basin of attraction' is probably sufficient for a complete comprehension of the result that is of interest here - provided there is reasonable familiarity with the definition and properties of Turing Machines (or

[15] In the definition of a dynamical system given below I am not striving to present the most general version. The basic aim is to lead to an intuitive understanding of the definition of a basin of attraction so that the main theorem is made reasonably transparent. Moreover, the definiton given below is for scalar ODEs, easily generalizable to the vector case.

partial recursive functions or equivalent formalisms encapsulated by Church's Thesis).

Definition 7 *The Initial Value Problem (**IVP**) for an Ordinary Differential Equation (**ODE**) and **Flows**. Consider a differential equation:*

$$\dot{x} = f(x) \tag{7}$$

where x is an unknown function of $t \in I$ (say, t : time and I an open interval of THE REAL LINE*) and f is a given function of x. Then, a function x is a solution of (7) on the* OPEN INTERVAL *I if:*

$$\dot{x}(t) = f(x(t)), \forall t \in I \tag{8}$$

The initial value problem (ivp) for (7) is, then, stated as:

$$\dot{x} = f(x), \qquad x(t_0) = x_0 \tag{9}$$

and a solution $x(t)$ for (9) is referred to as a solution through x_0 at t_0. Denote $x(t)$ and x_0, respectively, as:

$$\varphi(t, x_0) \equiv x(t), \text{ and } \varphi(0, x_0) \equiv x_0 \tag{10}$$

*where $\varphi(t, x_0)$ is called the **flow** of $\dot{x} = f(x)$.*

Definition 8 ***Dynamical System***

*If f is a C^1 function (i.e., the set of all differentiable functions with continuous first derivatives), then the **flow** $\varphi(t, x_0), \forall t$, induces a **map** of $U \sqsubset \mathbb{R}$ into itself, called a C^1 **dynamical system on** $\mathbb{R}$:*

$$x_0 \longmapsto \varphi(t, x_0) \tag{11}$$

if it satisfies the following (one-parameter group) properties:

1. $\varphi(0, x_0) = x_0$
2. $\varphi(t + s, x_0) = \varphi(t, \varphi(s, x_0)), \forall t \ \& \ s$, *whenever both the l.h and r.h side maps are defined;*
3. $\forall t, \varphi(t, x_0)$ *is a C^1 map with a C^1 inverse given by:* $\varphi(-t, x_0)$;

Remark 9 *A geometric way to think of the connection between a **flow** and the induced **dynamical system** is to say that the flow of an **ODE** gives rise to a dynamical system on $\mathbb{R}$.*

Remark 10 *It is important to remember that the **map** of $U \sqsubset \mathbb{R}$ into itself may **not** be defined on all of $\mathbb{R}$. In this context, it might be useful to recall the distinction between partial recursive functions and total functions in classical recursion theory.*

Definition 11 ***Invariant set***
A set (usually compact) $S \sqsubset U$ is ***invariant*** *under the flow $\varphi(.,.)$ whenever $\forall t \in \mathbb{R}$, $\varphi(.,.) \sqsubset S$.*

Definition 12 ***Attracting set***
A closed invariant set $A \sqsubset U$ is referred to as the ***attracting set*** *of the* ***flow*** *$\varphi(t,x)$ if $\exists$ some neighbourhood V of A, s.t $\forall x \in V$ & $\forall t \geq 0$, $\varphi(t,x) \in V$ and:*

$$\varphi(t,x) \rightarrow A \text{ as } t \rightarrow \infty \tag{12}$$

Remark 13 *It is important to remember that in dynamical systems theory contexts the attracting sets are considered the* ***observable*** *states of the dynamical system and its flow.*

Definition 14 *The basin of attraction of the attracting set A of a flow, denoted, say, by Θ_A, is defined to be the following set:*

$$\Theta_A = \cup_{t \leq 0} \varphi_t(V) \tag{13}$$

where: $\varphi_t(.)$ denotes the flow $\varphi(.,.), \forall t$.

Remark 15 *Intuitively, the basin of attraction of a flow is the* ***set of initial conditions*** *that eventually leads to its attracting set - i.e., to its limit set (limit points, limit cycles, strange attractors, etc). Anyone familiar with the definition of a Turing Machine and the famous Halting problem for such machines – or, alternatively, Rice's theorem – would immediately recognise the connection with the definition of basin of attraction and suspect that my main result is obvious.*

Definition 16 ***Dynamical Systems capable of Computation Universality****:*
A dynamical system capable of computation universality is one whose defining initial conditions can be used to program and simulate the actions of any arbitrary Turing Machine, in particular that of a Universal Turing Machine.

Proposition 17 *Dynamical systems characterizable in terms of limit points, limit cycles or 'chaotic' attractors, called 'elementary attractors', are not capable of universal computation.*

Proposition 18 *Only dynamical systems whose basins of attraction are poised on the boundaries of elementary attractors are capable of universal computation.*

Theorem 19 *There is no effective procedure to decide whether a given observable trajectory is in the basin of attraction of a dynamical system capable of computation universality*

Proof. The first step in the proof is to show that the basin of attraction *of a dynamical system capable of universal computation is recursively enumerable but*

not recursive. The second step, then, is to apply Rice's theorem to the problem of membership decidability in such a set.

First of all, note that the basin of attraction of a dynamical system capable of universal computation is recursively enumerable. This is so since trajectories belonging to such a dynamical system can be effectively listed simply by trying out, systematically, sets of appropriate initial conditions.

On the other hand, such a basin of attraction is not recursive. For, suppose a basin of attraction of a dynamical system capable of universal computation is recursive. Then, given arbitrary initial conditions, the Turing Machine corresponding to the dynamical system capable of universal computation would be able to answer whether (or not) it will halt at the particular configuration characterising the relevant observed trajectory. This contradicts the unsolvability of the Halting problem for Turing Machines.

Therefore, by Rice's theorem, *there is no effective procedure to decided whether any given arbitrary observed trajectory is in the basin of attraction of such recursively enumerable but not recursive basin of attraction.* ■

Remark 20 *There is a 'monumental' mathematical 'fudge' in my proof of the recursive enumerability of the basin of attraction: how can one try out, 'systematically', the set of uncountable initial conditions lying in the appropriate subset of* $\Re$*? Of course, this cannot be done and the theorem is given just to give an idea of the problem that I want to consider.*

Keeping the framework and the questions in mind, one way to proceed would be to constructivise the basic IVP problem for ODEs and then the theorem can be applied consistently. It will require too much space and time to do so within the scope of this paper. Instead, I shall adopt a slightly devious method.

Consider the following Generalized Shift (GS) map ([19],[20]):

$$\Phi : \wp \rightarrow \sigma^{F(\wp)} \left[\wp \oplus G\left(\wp\right)\right] \tag{14}$$

Where:

$\wp$: (bi-infinite) symbol sequence;

F : mapping from a finite subset of $\wp$ to the integers;

G: mapping from a finite subset of $\wp$ into $\wp$;

σ: a shift operator;

The given 'finite subset of $\wp$', on which F and G operate is called the *domain of dependence* (*DOD*).

Let the given symbol sequence be, for example:

$$\wp \equiv \{....p_{-1}pp_{=1}....\} \tag{15}$$

Then:

$\wp \oplus G\left(\wp\right) \Rightarrow$ replace DOD by $G\left(\wp\right)$.

$\sigma^{F(\wp)} \Rightarrow shift$ the sequence left or right by the amount $F\left(\wp\right)$

Remark 21 *In practice, a GS is implemented by denoting a distinct position on the initially given symbol sequence as, say,* p_0 *and placing a 'reading head' over it. It must also be noted that* $p_i \in \wp, \forall i = 1, 2, ...$ *could, for example, denote whole words from an alphabet, etc., although in practice it will be* $0, 1$ *and* $\circ$ *('dot'). The 'dot' will be signify that the 'reading head' will be placed on the symbol to the right of it.*

The following results about Generalized Shift maps are relevant for my discussion:

Proposition 22 *Any GS is a nonlinear (in fact, piecewise linear) dynamical system capable of universal computation; hence they are universal dynamical systems and are equivalent to some constructible Universal Turing Machine.*

Thus the GS is capable of universal computation and it is minimal in a precisely definable sense (see [19] and [20] for full details). It is also possible to construct, for each such generalized shift dynamical system[16], an equivalent UTM that can simulate its dynamics, for sets of initial conditions. Now consider the observable set of the dynamical system, $y \in A$; given the UTM, say $\mathbb{U}$, corresponding to $\wp$, the question is: for what set of initial conditions, say x, is y the halting state of $\mathbb{U}$. Naturally, by the theorem of the unsolvability of the Halting problem, this is an undecidable question. This is the theorem used in demonstrating the uncomputability of $K_\phi(y|x)$. However, by the above Zvonkin-Levin theorem, we know that the *existence* of 'arbitrarily good upper estimates' for $K_\phi(y|x)$, even although $K_\phi(y|x)$ is uncomputable.

Now, taking a cue from Rissanen's methodological point about the irrelevance of 'true' models, but only of models that can explain the data minimally, let me consider the above (minimal) universal dynamical system as canonical for any question about membership in attracting sets, A. What is the complexity of $K_U(p|x)$? By definition it should be:

$$K_{\mathbb{U}}(y|x) = \{ \begin{matrix} \min_{\mathbb{U}(p,x)=y} l(p) \\ \infty \nexists\, p \text{ s.t } \phi(p,x) = y \end{matrix}$$

The meaning, of course, is: the minimum over all programs, p, implemented on $\mathbb{U}$, with the given initial condition, x, which will stop at the halting configuration, y. The above theorem formalizes the notion that there is no general algorithmic procedure to decide any such membership.

Remark 23 *Why is it important to show the existence of the minimal program? Because, is the observed* y *corresponds to the minimal program of the dynamical system, i.e., of* $\mathbb{U}$*, then it is capable of computation universality; if there is no minimal program, the dynamical system is not interesting! A monotone decreasing set of programs that can be shown to converge to the minimal program is analogous to a series of increasingly complex finite automata converging to a*

[16] They can also encapsulate smooth dynamical systems in a precise sense. I have described the procedure, summarising a part of Chris Moore's approach, in [28], Chapter 4.

TM. What we have to show is that there are programs converging to the minimal program from above and below, to the border between two basins of attractions.

Thus, behind every undecidable proposition – at least in principle – there is an inference principle which may or may not suggest an 'approximation' strategy to 'decide the undecidable'. After all, Gödel himself thought that the undecidable may become decidable by going 'upwards', so to speak, in strengthening the axiom systems; surely, there must be a practical way of going in the opposite direction to locate the borders of the decidable as approximation to the undecidable, too. Naturally, I expect these highly speculative conjectures to apply, *pari passu*, to the computable-uncomputable divide, too.

5 Concluding Thoughts

> "Inductive processes have formed, of course, at all times a vital, habitual part of the mind's machinery. Whenever we learn by experience, we are using them. But in the logic of the schools they have taken their proper place slowly."
>
> John Maynard Keynes, [16], p.241.

It is to Jorma Rissanen's lasting credit that he has, almost single-handedly developed a scientific method to make this 'habitual part of the mind's machinery' entirely and rigorously algorithmic. Thus, he belongs to the modern scientific movement towards an algorithmic approach to statistics, randomness and information. As an economist, I have strived to develop an analogous field of algorithmic economics, where stochastic complexity and the MDL principle are as central as algorithmic randomness, computability theory and computational complexity theory. Learning and induction – indeed, learning as induction – is a central topic at the frontiers of economics. The frontier researchers remain blissfully ignorant of the algorithmic approach to learning and inductive inference, randomness and information. This is strange in a subject which prides itself on placing the role of scarce information and its husbanding in its citadel.

Economics is singularly free of an algorithmic vision. The mathematics of economic theory is dominated by Bourbakian thinking. The methodology of statistical inference in economics is equally stone-aged.

The success of Jorma Rissanen's single-handed, even single-minded, efforts to inject a new algorithmic vision into statistical methodology, particularly in inference, estimation and prediction theories, is heartening for those of us who find ourselves at the fringes of mathematical economics in view of our algorithmic vision.

I believe Jorma Rissanen's work contributes a missing link to the great Finnish tradition of work in inductive logic, one which was most cogently stated by Hilary Putnam in that period of an interregnum between the growing sytematization of the philosophy of inductive logic and the emergence of the recursion

theoretic inductive movement[17] ([22], p.297):

> "[W]e may think of a system of inductive logic as a design for a 'learning machine': that is to say, a design for a computing machine that can extrapolate certain kinds of empirical regularities from the data with which it is supplied."

Jorma Rissanen, together with Ray Solomonoff, have pioneered and 'patented' not only the design for a 'learning machine'; they have actually built it.

[17] In particular, it must be remembered that Solomonoff's work straddles the two traditions and his two path-breaking contributions appeared almost before the proverbial ink was dry on Putnam's seminal contribution.

References

[1] Barron, A. R, Jorma Rissanen & B. Yu, (1998), *The Minimum Description Length Principle in Coding and Modelling*, **IEEE Transactions on Information Theory**, Vol. 44, #6, pp. 2743-60.

[2] Blum, Lenore, Felipe Cucker, Michael Shub and Steve Smale (1998), **Complexity and Real Computation**, Springer Verlag, New York.

[3] Boehner, P, (1990), **Okham – Philosophical Writings**, Hackett, Indianapolis.

[4] Brin, Michael and Garrett Stuck (2002): **Introduction to Dynamical Systems**, Cambridge University Press, Cambridge.

[5] Cooper, S Barry (2004): **Computability Theory**, Chapman & Hall/CRC, Boca Raton and London

[6] Cover, Thomas. M & Joy A. Thomas, (1991), **Elements of Information Theory**, John Wiley & Sons, Inc., New York & Chichester.

[7] Dordick, Rowan, (1988), *Understanding the 'go' of it*, **IBM Research Magazine**, Winter.

[8] Gács, Péter, John T. Tromp and Paul M. B. Vitányi, (2001), *Algorithmic Statistics*, **IEEE Transactions on Information Theory**, Vol.47, #6, September, pp.2443-63

[9] Hilpinen, Risto, (1968), **Rules of Acceptance and Inductive Logic**, Acta Philosophica Fennica, 22, North-Holland, Amsterdam.

[10] Hirsch, Morris. W, Stephen Smale & Robert L. Devaney, (2004), **Differential Equations, Dynamical Systems & An Introduction to Chaos**, Academic Press, an Imprint of Elsevier, San Diego & London.

[11] Hintikka, Jaakko & Ilkka Niniluoto, (1966), *A Two-Dimensional Continuum of Inductive Methods*, pp. 113-32, in: **Aspects of Inductive Logic**, edited by Jaako Hintikka & Patrick Suppes, North-Holland, Amsterdam.

[12] Hintikka, Jaakko & Ilkka Niiniluoto, (1980), *An Axiomatic Foundation for the Logic of Inductive Generalization*, Chapter 7, pp.157-81, in: **Studies in Inductive Logic and Probability, Volume II**, edited by Richard C. Jeffrey, University of California Press, Berkeley and Los Angeles.

[13] Jevons, W. Stanley, (1877), **The Principles of Science: A Treatise on Logic and Scientific Method**, Second Edition, Revised, Macmillan & Co., London & New York.

[14] Johnson, W. E, (1895), *Logic and Political Economy*, pp.637-8, in: **Dictionary of Political Economy**, Volume 2, edited by R.H. Inglis Palgrave, Macmillan, London.

[15] Kemeny, J.G, (1953), *The Use of Simplicity in Induction*, **The Philosophical review**, Vol.62, pp.391-408.

[16] Keynes, John Maynard, (1921), **A Treatise on Probability,** Reprinted as: Volume VIII, *The Collected Writings of John Maynard Keynes*, Macmillan, Cambridge University Press, for the Royal Economic Society, London.

[17] Keynes, John Maynard, (1973, [1936]), *William Stanley Jevons*, Reprinted in: Volume X, **The Collected Writings of John Maynard Keynes**, Macmillan, Cambridge University Press, for the Royal Economic Society, London.

[18] Kolmogorov, A.N, (1968), *Three Approaches to the Definition of the Concept of the "Amount of Information"*, pp. 293-302, **Selected Translations in Mathematical Statistics and Probability**, Volume 7, American Mathematical Society, Providence, Rhode Island.

[19] Moore, Christopher, (1990), *Unpredictability and Undecidability in Dynamical Systems*, **Physical Review Letters**, Vol.64, #4, 14 May, pp.2354-7.

[20] Moore, Christopher, (1991), *Generalized Shifts: Unpredictability and Undecidability in Dynamical Systems*, **Nonlinearity**, Vol.4, pp. 199-230.

[21] Niniluoto, Ilkka & Raimo Tuomela, (1973), **Theoretical Concepts and Hypothetico-Inductive Inference**, Dordrecht, Boston.

[22] Putnam, Hilary, (1975 [1963]), *Probability and Confirmation*, Reprinted in: **Mathematics, Matter and method – Philosophical Papers**, Vol. 1, Cambridge University Press, Cambridge.

[23] Rissanen, Jorma, (1978), *Modelling by Shortest Data Description*, **Automatica**, Vol.14, # 5, September, pp.465-71.

[24] Rissanen, Jorma, (1986), *Stochastic Complexity and Modeling*, **The Annals of Statistics**, Vol.14, #3, pp.1080-1100.

[25] Rissanen, Jorma, (1987), *Stochastic Complexity*, **Journal of the Royal Statistical Society, Series B**, Vol.49, #3,pp.223-39 & pp.252-65.

[26] Rogers, Hartley, Jr., (1967): **Theory of Recursive Functions and Effective Computability**, McGraw-Hill, New York.

[27] Rustem, Berc, Kumaraswamy Velupillai & John H. Westcott, (1978), *Respecifying the Weighting Matrix of a Quadratic Objective Function*, **Automatica**, Vol.14,# 6, November, pp. 567-82.

[28] Velupillai, Kumaraswamy, (2000), **Computable Economics**, Oxford University Press, Oxford.

[29] Velupillai, K. Vela, (2007), *The Impossibility of an Effective Theory of Policy in a Complex Economy*, in: **Complexity Hints for Policy**, edited by David Colander and Massimo Salzano, Springer-Verlag, Berlin and New York.

[30] Velupillai, K. Vela, (2007), *Demystifying Induction and Falsification: Trans-Popperian Suggestions*, in: **Induction, Metaphysics and Economic Methodology: Reflection on Popperian Themes** edited by Tom Boylan and Pascal O'Gorman, Routledge, London.

[31] Vitanyi, Paul M.B & Ming Li, (2000), *Minimum Description length Induction, Bayesianism, and Kolmogorov Complexity*, **IEEE Transactions on Information Theory**, Vol.46, # 2, March, pp.446-64.

[32] von Wright, Georg Henrik, (1951), **A Treatise on Induction and Probability**, Routledge & Kegan Paul, London.

[33] von Wright, Georg Henrik, (1957), **The Logical Problem of Induction**, Basil Blackwell, Oxford.

[34] Zvonkin, A.K & L.A. Levin, ( ), *The Complexity of Finite Objects and the Development of the Concepts of Information and Randomness by Means of the Theory of Algorithms*, **Russian Mathematical Surveys**, Vol.25, # 6, pp. 83-124.

Compression-based methods for nonparametric on-line prediction, regression, classification and density estimation of time series *

Boris Ryabko

Siberian State University of Telecommunications and Informatics,
Institute of Computational Technologies of Siberian Branch of Russian Academy of Sciences,
Novosibirsk, Russia. e-mail: boris@ryabko.net

Abstract

Jorma Rissanen has discovered some deep connections between universal coding (or universal data compression) and mathematical statistics. In particular, the MDL principle has been one of the most powerful methods of modern mathematical statistics. In this paper we apply Rissanen's approach and ideas to some statistical problems concerned with time series. We address the problem of nonparametric estimation of characteristics of stationary and ergodic time series. We consider finite-alphabet as well as real-valued time series and the following four problems: i) estimation of the (limiting) probability $P(u_0...u_s)$ for every s and each sequence $u_0 \dots u_s$ of letters over the process alphabet (or estimation of the density $p(x_0, ..., x_s)$ for real-valued time series), ii) so-called on-line prediction, where the conditional probability $P(x_{t+1}|x_1x_2...x_t)$ (or the conditional density $p(x_{t+1}|x_1x_2...x_t)$) should be estimated (when $x_1x_2 \dots x_t$ is known), iii) regression and iv) classification (or so-called problems with side information). We show that so-called archivers (or data compressors) can be used as a tool for solving these problems. In particular, it is proven that any universal code (or universal data compressor) can be used as a basis for constructing asymptotically optimal methods for the above problems. (By definition, a universal code can "compress" any sequence generated by a stationary and ergodic source asymptotically to the Shannon entropy of the source.)

AMS subject classification: 60G10, 60J10, 62G07, 62G08, 62M20, 94A29.

keywords: time series, nonparametric estimation, prediction, universal coding, data compression, on-line prediction, Shannon entropy, stationary and ergodic process, regression.

1 Introduction

We consider a stationary and ergodic source which generates sequences $x_1x_2 \dots$ of elements (letters) from some set (alphabet) A, which is either finite or real-valued. It is supposed that the probability distribution (or distribution of limiting probabilities) $P(x_1 = a_{i_1}, x_2 = a_{i_2}, ..., x_t = a_{i_t})$ (or the density $p(x_1, x_2, ..., x_t)$) is unknown, but we are given either one sample $x_1...x_t$ or several (r) independent samples $x^1 = x^1_1...x^1_{t_1}, ..., x^r = x^r_1...x^r_{t_r}$ generated by the source. (Generally speaking, they cannot be combined into one sample for a stationary and ergodic source as it can be done for an i.i.d. one.) Of course, if someone knows the probability distribution (or the density) he has all information about the source and can solve all problems in the best

*Research was supported by Russian Foundation for Basic Research (grant no. 06-07-89025.)

way. Hence, precise estimations of the probability distribution and the density can be used for prediction, regression estimation, etc. In this paper we use the following scheme: we consider the problems of estimation of the probability distribution or density. Then we show how the solution can be applied to other problems, paying the main attention to the problem of prediction, because of its practical applications and importance for probability theory, information theory, statistics and other theoretical sciences, see [1, 16, 17, 20, 28, 29, 31, 34, 35, 36, 38, 41, 46]. We show that universal codes (or data compressors) can be applied directly to the problems of estimation, prediction, regression and classification. It is not surprising because for any stationary and ergodic source P generating letters from a finite alphabet and any universal code U the following equality is valid with probability 1:

$$\lim_{t\to\infty} t^{-1}(-\log P(x_1\dots x_t) - |U(x_1\dots x_t)|) = 0,$$

where $x_1\dots x_t$ is generated by P. (Here and below $\log = \log_2, |v|$ is the length of v, if v is a word, and the number of elements of v if v is a set.) So, in fact, the length of the universal code $(|U(x_1\dots x_t)|)$ can be used as an estimate of the logarithm of the unknown probability and, obviously, $2^{-|U(x_1\dots x_t)|}$ can be considered as the estimation of $P(x_1\dots x_t)$. In fact, a universal code can be viewed as a non-parametrical estimation of (limiting) probabilities for stationary and ergodic sources. This was recognized shortly after the discovery of universal codes (for the set of stationary and ergodic processes with finite alphabets [40]) and universal codes were applied for solving prediction problem [35, 41].

We would like to emphasize that, on the one hand, all results are obtained in the framework of classical probability theory and mathematical statistics and, on the other hand, everyday methods of data compression (or archivers) can be used as a tool for density estimation, prediction and other problems, because they are practical realizations of universal codes. It is worth noting that modern data compressors are based on deep theoretical results of source coding theory (see, e.g., [11, 17, 21, 24, 33, 34, 35, 37, 48]) and have demonstrated high efficiency in practice as compressors of texts (zip, arj, rar, etc.), DNA sequences [24] and many other types of real data. In fact, archivers can find many kinds of latent regularities, that is why they look like a promising tool for prediction and other problems. Moreover, recently universal codes and archivers were effectively applied to some problems which are very far from data compression: first, their applications created a new and rapidly growing line of investigations in clustering and classification (see [4, 5] and references therein) and, second, universal codes were used as a basis for non-parametric tests for the main statistical hypotheses concerned with stationary and ergodic time series [44, 45]. The outline of the paper is as follows. Section 2 contains description of the Laplace predictor and its generalizations, a review of known results and description of one universal code. Sections 3 and 4 are devoted to processes with finite and real-valued alphabets, correspondingly.

2 Predictors and universal data compressors

2.1 The Laplace measure and on-line prediction for i.i.d. processes

We consider a source with unknown statistics which generates sequences $x_1x_2\cdots$ of letters from some set (or alphabet) A. It will be convenient at first to describe briefly the prediction problem. Let the source generate a message $x_1\dots x_{t-1}x_t,\ x_i \in A$ for all i, and the next letter x_{t+1} needs to be predicted. This can be traced back to Laplace who considered the problem how to estimate the probability that the sun will rise tomorrow, given that it has risen every day since Creation

(see [12]). In our notation the alphabet A contains two letters: 0 ("the sun rises") and 1 ("the sun does not rise"), t is the number of days since Creation, $x_1 \dots x_{t-1}x_t = 00 \dots 0$.

Laplace suggested the following predictor:

$$L_0(a|x_1 \cdots x_t) = (\nu_{x_1 \cdots x_t}(a) + 1)/(t + |A|), \tag{1}$$

where $\nu_{x_1 \cdots x_t}(a)$ denotes the count of letter a occurring in the word $x_1 \dots x_{t-1}x_t$. For example, if $A = \{0, 1\}$, $x_1 ... x_5 = 01010$, then the Laplace prediction is as follows: $L_0(x_6 = 0|01010) = (3+1)/(5+2) = 4/7, L_0(x_6 = 1|01010) = (2+1)/(5+2) = 3/7$. In other words, $3/7$ and $4/7$ are estimates of the unknown probabilities $P(x_{t+1} = 0|x_1 \dots x_t = 01010)$ and $P(x_{t+1} = 1|x_1 \dots x_t = 01010)$. (It is worth noting that the estimated probability to encounter zero after observing a binary string that contains only zeros is not one.)

We can see that Laplace considered prediction as a set of estimations of unknown (conditional) probabilities. This approach to the problem of prediction was developed in [41] and now is often called on-line prediction or universal prediction [1, 20, 28, 31]. As we mentioned above, it seems natural to consider conditional probabilities to be the best prediction, because they contain all information about the future behavior of the stochastic process. Moreover, this approach is deeply connected with game-theoretical interpretation of prediction (see [18, 43]) and, in fact, all obtained results can be easily transferred from one model to the other.

Any predictor γ defines a measure by the following equation

$$\gamma(x_1 ... x_t) = \prod_{i=1}^{t} \gamma(x_i|x_1 ... x_{i-1}). \tag{2}$$

For example, $L_0(0101) = \frac{1}{2}\frac{1}{3}\frac{1}{2}\frac{2}{5} = \frac{1}{30}$. And, vice versa, any measure γ (or estimate of the measure) defines a predictor: $\gamma(x_i|x_1 ... x_{i-1}) = \gamma(x_1 ... x_{i-1}x_i)/\ \gamma(x_1 ... x_{i-1})$. The same is true for a density (and its estimate): a predictor is defined by conditional density and, vice versa, the density is equal to the product of conditional densities:

$$p(x_i|x_1 \dots x_{i-1}) = \frac{p(x_1 \dots x_{i-1}x_i)}{p(x_1 \dots x_{i-1})},\ p(x_1 \dots x_t) = \prod_{i=1}^{t} p(x_i|x_1 \dots x_{i-1}).$$

The next natural question is how to estimate the precision of a prediction and a probability estimation. Mainly we will estimate the error of prediction by the Kullback-Leibler (KL) divergence between a distribution P and its estimate as follows:

$$\rho_{\gamma,P}(x_1 \cdots x_t) = \sum_{a \in A} P(a|x_1 \cdots x_t) \log \frac{P(a|x_1 \cdots x_t)}{\gamma(a|x_1 \cdots x_t)}, \tag{3}$$

where γ is the estimate of an unknown conditional probability. It is well-known that for any distributions P and γ the KL divergence is nonnegative and equals 0 if and only if $P(a) = \gamma(a)$ for all a, see, e.g., [15]. The following inequality (Pinsker's inequality)

$$\sum_{a \in A} P(a) \log \frac{P(a)}{Q(a)} \geq \frac{\log e}{2} ||P - Q||^2 \tag{4}$$

connects the KL divergence with the so-called variation distance

$$||P - Q|| = \sum_{a \in A} |P(a) - Q(a)|,$$

where P and Q are distributions over A, see [7]. For fixed t, $\rho_{\gamma,P}(\,)$ is a random variable, because $x_1, x_2, \cdots, x_t$ are random variables. We define the average error at time t by

$$\rho^t(P\|\gamma) = E\ (\rho_{\gamma,P}(\cdot)) = \sum_{x_1\cdots x_t \in A^t} P(x_1\cdots x_t)\ \rho_{\gamma,P}(x_1\cdots x_t). \tag{5}$$

It is shown in [42] that the error of Laplace predictor L_0 goes to 0 for any i.i.d. source P. More precisely, it is proven that

$$\rho^t(P\|L_0) \leq (|A|-1)\log e/(t+1) \tag{6}$$

for any source P, [42], see also [46]. So, we can see from this inequality that the average error of the Laplace predictor L_0 (estimated either by the KL divergence or the variation distance) goes to zero for any unknown i.i.d. source, when the sample size t grows. Moreover, it can be easily shown that the error (3) (and the corresponding variation distance) goes to zero with probability 1, when t goes to infinity. Obviously, such a property is very desirable for any predictor and for larger classes of sources, like Markov, stationary and ergodic, etc. However, it is proven in [41] (see also [1]) that such predictors do not exist for the class of all stationary and ergodic sources (generating letters from a given finite alphabet). More precisely, for any predictor γ there exists a source P and $\delta > 0$ such that with probability 1 $\rho_{\gamma,P}(x_1\cdots x_t) \geq \delta$ infinitely often when $t \to \infty$. So, the error of any predictor may not go to 0, if the predictor is applied to an arbitrary stationary and ergodic source, that is why it is difficult to use (3) and (5) to compare different predictors.

On the other hand, it is shown in [41], that there exists a predictor R, such that the following Cesaro average $t^{-1}\sum_{i=1}^{t}\rho_{R,P}(x_1\cdots x_t)$ goes to 0 (with probability 1) for any stationary and ergodic source P, where t goes to infinity. That is why we will focus our attention on such averages and by analogy with (5) we define

$$\bar{\rho}_{\gamma,P}(x_1...x_t) = t^{-1}\,(\log(P(x_1...x_t)/\gamma(x_1...x_t)) \tag{7}$$

and

$$\bar{\rho}_t(\gamma,P) = t^{-1}\sum_{x_1...x_t\in A^t} P(x_1...x_t)\log(P(x_1...x_t)/\gamma(x_1...x_t)), \tag{8}$$

where, as before, $\gamma(x_1...x_t) = \prod_{i=1}^{t}\gamma(x_i|x_1...x_{i-1})$.

From these definitions and (6) we obtain the following estimation of the error of the Laplace predictor L_0 for any i.i.d. source:

$$\bar{\rho}_t(L_0,P) < ((|A|-1)\ \log t + c)/t, \tag{9}$$

where c is a certain constant. So, we can see that the average error of the Laplace predictor goes to zero for any i.i.d. source (which generates letters from a known finite alphabet). As a matter of fact, the Laplace probability $L_0(x_1\ldots x_t)$ is a consistent estimate of the unknown probability $P(x_1\ldots x_t)$.

A natural problem is to find a predictor whose error is minimal (for i.i.d. sources). This problem was considered and solved by Krichevsky in [25], see also [26]. He suggested the following predictor:

$$K_0(a|x_1\cdots x_t) = (\nu_{x_1\cdots x_t}(a) + 1/2)/(t + |A|/2), \tag{10}$$

where, as before, $\nu_{x_1\cdots x_t}(a)$ is the count of letter a occurring in the word $x_1\ldots x_t$. We can see that the Krichevsky predictor is quite close to Laplace's one (1). For example, if $A = \{0,1\}$, $x_1...x_5 = 01010$, then $K_0(x_6 = 0|01010) = (3{+}1/2)/(5{+}1) = 7/12$, $K_0(x_6 = 1|01010) = (2{+}1/2)/(5{+}1) = 5/12$ and $K_0(01010) = \frac{1}{2}\frac{1}{4}\frac{1}{2}\frac{3}{8}\frac{1}{2} = \frac{3}{256}$.

The Krichevsky measure K_0 can be presented as follows:

$$K_0(x_1...x_t) = \prod_{i=1}^{t} \frac{\nu_{x_1...x_{i-1}}(x_i) + 1/2}{i - 1 + |A|/2} = \frac{\prod_{a\in A}(\prod_{j=1}^{\nu_{x_1...x_t}(a)}(j - 1/2))}{\prod_{i=0}^{t-1}(i + |A|/2)}. \tag{11}$$

It is known that

$$(r + 1/2)((r + 1) + 1/2)...(s - 1/2) = \frac{\Gamma(s + 1/2)}{\Gamma(r + 1/2)}, \tag{12}$$

where $\Gamma(\ )$ is the gamma function (see, e.g., [22] for definition). So, (11) can be presented as follows:

$$K_0(x_1...x_t) = \frac{\prod_{a\in A}(\Gamma(\nu_{x_1...x_t}(a) + 1/2)\,/\Gamma(1/2)\,)}{\Gamma(t + |A|/2)\,/\Gamma(|A|/2)}. \tag{13}$$

For this predictor

$$\bar{\rho}_t(K_0, P) < ((|A| - 1)\,\log t + c)/(2t), \tag{14}$$

where c is a constant, and, moreover, in a certain sense this average error is minimal: for any predictor γ there exists such a source p^* that

$$\bar{\rho}_t(\gamma, p^*) \geq ((|A| - 1)\,\log t + c)/(2t),$$

see [25, 26].

2.2 Consistent estimations and on-line predictors for Markov and ergodic processes

Now we briefly describe consistent estimations of unknown probabilities and efficient on-line predictors for general stochastic processes (or sources of information). Denote by A^t and A^* the set of all words of length t over A and the set of all finite words over A correspondingly ($A^* = \bigcup_{i=1}^{\infty} A^i$). By $M_\infty(A)$ we denote the set of all stationary and ergodic sources which generate letters from A and let $M_0(A) \subset M_\infty(A)$ be the set of all i.i.d. processes. Let $M_m(A) \subset M_\infty(A)$ be the set of Markov sources of order (or with memory, or connectivity) not larger than m, $m \geq 0$. Let $M^*(A) = \bigcup_{i=0}^{\infty} M_i(A)$ be the set of all finite-memory sources.

The Laplace and Krichevsky predictors can be extended to general Markov processes. The trick is to view a Markov source $P \in M_m(A)$ as resulting from $|A|^m$ i.i.d. sources. We illustrate this idea by an example from [46]. So assume that $A = \{O, I\}$, $m = 2$ and assume that the source $P \in M_2(A)$ has generated the sequence

$$OOIOIIOOIIIOIO.$$

We represent this sequence by the following four subsequences:

$$**I*****I*****,$$

$$***O*I***I***O,$$

$$****I**O****I*,$$

$$******O***IO**.$$

These four subsequences contain letters which follow OO, OI, IO and II, respectively. By definition, $P \in M_m(A)$ if $P(a|x_1 \cdots x_t) = P(a|x_{t-m+1} \cdots x_t)$, for all $0 < m \leq t$, all $a \in A$ and all $x_1 \cdots x_t \in A^t$. Therefore, each of the four generated subsequences may be considered as

generated by a Bernoulli source. Further, it is possible to reconstruct the original sequence if we know the four ($= |A|^m$) subsequences and the two ($= m$) first letters of the original sequence.

Any predictor γ for i.i.d. sources can be applied to Markov sources. Indeed, in order to predict, it is enough to store in the memory $|A|^m$ sequences, one corresponding to each word in A^m. Thus, in the example, the letter x_3 which follows OO is predicted based on the Bernoulli method γ corresponding to the x_1x_2- subsequence ($= OO$), then x_4 is predicted based on the Bernoulli method corresponding to x_2x_3, i.e. to the OI- subsequence, and so forth. When this scheme is applied along with either L_0 or K_0 we denote the obtained predictors as L_m and K_m, correspondingly, and define the probabilities for the first m letters as follows: $L_m(x_1) = L_m(x_2) = \ldots = L_m(x_m) = 1/|A|$, $K_m(x_1) = K_m(x_2) = \ldots = K_m(x_m) = 1/|A|$. For example, having taken into account (13), we can present the Krichevsky predictors for $M_m(A)$ as follows:

$$K_m(x_1...x_t) = \begin{cases} \frac{1}{|A|^t}, & \text{if } t \le m\,, \\ \frac{1}{|A|^m} \prod_{v\in A^m} \frac{\prod_{a\in A} ((\Gamma(\nu_x(va)+1/2)\,/\,\Gamma(1/2))}{(\Gamma(\bar\nu_x(v)+|A|/2)\,/\,\Gamma(|A|/2))}, & \text{if } t > m \quad , \end{cases} \tag{15}$$

where $\bar\nu_x(v) = \sum_{a\in A} \nu_x(va)$, $x = x_1...x_t$; see [25] and references therein. It is worth noting that representation (12) can be more convenient for carrying out calculations. Let us consider an example. For the word $OOIOIIOOIIIOIO$ considered in the previous example, we obtain $K_2(OOIOIIOOIIIOIO) = 2^{-2}\ \frac{1}{2}\frac{3}{4}\ \frac{1}{2}\frac{1}{4}\frac{1}{2}\frac{3}{8}\ \frac{1}{2}\frac{1}{4}\frac{1}{2}\ \frac{1}{2}\frac{1}{4}\frac{1}{2}$.

Let us define the measure R, which is a consistent estimator of probabilities for the class of all stationary and ergodic processes with a finite alphabet. First we define a probability distribution $\{\omega = \omega_1, \omega_2, ...\}$ on integers $\{1, 2, ...\}$ by

$$\omega_1 = 1 - 1/\log 3,\ ...\ ,\ \omega_i = 1/\log(i+1) - 1/\log(i+2),\ ...\ . \tag{16}$$

(In what follows we will use this distribution, but results described below are obviously true for any distribution with nonzero probabilities.) The measure R is defined as follows:

$$R(x_1...x_t) = \sum_{i=0}^{\infty} \omega_{i+1}\, K_i(x_1...x_t). \tag{17}$$

It is worth noting that this construction can be applied to the Laplace measure (if we use L_i instead of K_i) and any other family of measures.

The main properties of the measure R are connected with the Shannon entropy, which is defined as follows

$$H(P) = \lim_{m\to\infty} -\frac{1}{m} \sum_{v\in A^m} P(v) \log P(v). \tag{18}$$

Theorem 1 ([41]). *For any stationary and ergodic source P the following equalities are valid:*

$$i)\ \lim_{t\to\infty} \frac{1}{t} \log(1/R(x_1 \cdots x_t)) \ = \ H(P)$$

with probability 1,

$$ii)\ \lim_{t\to\infty} \frac{1}{t} \sum_{u\in A^t} P(u) \log(1/R(u)) \ = \ H(P).$$

2.3 Nonparametric estimations and data compression

One of the goals of the paper is to show how practically used data compressors can be employed as a tool for nonparametric estimation, prediction and other problems. That is why a short description of universal data compressors (or universal codes) will be given here.

A data compression method (or code) φ is defined as a set of mappings φ_n such that $\varphi_n : A^n \to \{0,1\}^*$, $n = 1, 2, \ldots$ and for each pair of different words $x, y \in A^n$ $\varphi_n(x) \neq \varphi_n(y)$. It is also required that each sequence $\varphi_n(u_1)\varphi_n(u_2)...\varphi_n(u_r), r \geq 1$, of encoded words from the set $A^n, n \geq 1$, could be uniquely decoded into $u_1u_2...u_r$. Such codes are called uniquely decodable. For example, let $A = \{a, b\}$, the code $\psi_1(a) = 0, \psi_1(b) = 00$, obviously, is not uniquely decodable. It is well known that if a code φ is uniquely decodable then the lengths of the codewords satisfy the following inequality (Kraft's inequality): $\sum_{u\in A^n} 2^{-|\varphi_n(u)|} \leq 1$, see, e.g., [15]. It will be convenient to reformulate this property as follows:

Claim 1. *Let φ be a uniquely decodable code over an alphabet A. Then for any integer n there exists a measure μ_φ on A^n such that*

$$-\log \mu_\varphi(u) \ \leq \ |\varphi(u)| \tag{19}$$

for any u from A^n .

(Obviously, Claim 1 is true for the measure $\mu_\varphi(u) = 2^{-|\varphi(u)|}/\Sigma_{u\in A^n} 2^{-|\varphi(u)|}$). In what follows we call uniquely decodable codes just "codes".

It is worth noting that, in fact, any measure μ defines a code for which the length of the codeword associated with a word u is (close to) $-\log \mu(u)$.

Now we consider universal codes. By definition, a code U is universal if for any stationary and ergodic source P the following equalities are valid:

$$\lim_{t\to\infty} |U(x_1 \ldots x_t)|/t = H(P) \tag{20}$$

with probability 1, and

$$\lim_{t\to\infty} E(|U(x_1 \ldots x_t)|)/t = H(P), \tag{21}$$

where $H(P)$ is the Shannon entropy of P, $E(f)$ is a mean value of f. In fact, (21) and (20) are valid for known universal codes, but there exist codes for which only one equality is valid.

3 Finite-alphabet processes

3.1 The estimation of (limiting) probabilities

The following theorem shows how universal codes can be applied for probability estimations.

Theorem 2. *Let U be a universal code and*

$$\mu_U(u) = 2^{-|U(u)|}/\Sigma_{v\in A^{|u|}} 2^{-|U(v)|}. \tag{22}$$

Then, for any stationary and ergodic source P the following equalities are valid:

$$i) \ \lim_{t\to\infty} \frac{1}{t}(-\log P(x_1 \cdots x_t) - (-\log \mu_U(x_1 \cdots x_t))) \ = \ 0$$

with probability 1,

$$ii) \ \lim_{t\to\infty} \frac{1}{t} \sum_{u\in A^t} P(u) \log(P(u)/\mu_U(u)) \ = \ 0,$$

Proof. The proof is based on Shannon-MacMillan-Breiman Theorem which states that for any stationary and ergodic source P

$$\lim_{t\to\infty} -\log P(x_1 \dots x_t)/t = H(P)$$

with probability 1, see [3, 15]. From this equality and (20) we obtain the statement i). The second statement follows from the definition of Shannon entropy (18) and (21). □

So, we can see that, in a certain sense, the measure μ_U is a consistent (nonparametric) estimate of the (unknown) measure P.

Nowadays there are many efficient universal codes (and universal predictors connected with them), see [11, 21, 26, 34, 35, 40, 48], which can be applied to estimation. For example, the above described measure R is based on the code from [40, 41] and can be applied for probability estimation. More precisely, Theorem 2 (and the following theorems) are true for R, if we replace μ_U by R.

It is important to note that the measure R has some additional properties, which can be useful for applications. The following theorem describes these properties (whereas all other theorems are valid for all universal codes and corresponding measures, including the measure R).

Theorem 3. *For any Markov process P with memory k*

i) the error of the probability estimator, which is based on the measure R, is upper-bounded as follows:

$$\frac{1}{t}\sum_{u\in A^t} P(u)\log(P(u)/R(u)) \le \frac{(|A|-1)|A|^k \log t}{2\,t} + O(\frac{1}{t}),$$

ii) in a certain sense the error of R is asymptotically minimal: for any measure μ there exists a $k-$memory Markov process p_μ such that

$$\frac{1}{t}\sum_{u\in A^t} p_\mu(u)\log(p_\mu(u)/\mu(u)) \ge \frac{(|A|-1)|A|^k \log t}{2\,t} + O(\frac{1}{t}),$$

iii) Let Θ be a set of stationary and ergodic processes such that there exists a measure μ_Θ for which the estimation error of the probability goes to 0 uniformly:

$$\lim_{t\to\infty} \sup_{P\in\Theta} \left(\frac{1}{t}\sum_{u\in A^t} P(u)\log(P(u)/\mu_\Theta(u))\right) = 0.$$

Then the error of estimator, which is based on the measure R, goes to 0 uniformly too:

$$\lim_{t\to\infty} \sup_{P\in\Theta} \left(\frac{1}{t}\sum_{u\in A^t} P(u)\log(P(u)/R(u))\right) = 0.$$

The proof can be found in [40, 41].

3.2 Prediction

As we mentioned above, any universal code U can be applied for prediction. Namely, the measure μ_U (22) can be used for prediction as the following conditional probability:

$$\mu_U(x_{t+1}|x_1...x_t) = \mu_U(x_1...x_tx_{t+1})/\mu_U(x_1...x_t). \tag{23}$$

Theorem 4. *Let U be a universal code and P be any stationary and ergodic process. Then*

$$i)\ \lim_{t\to\infty}\frac{1}{t}\ \{E(\log\frac{P(x_1)}{\mu_U(x_1)}) + E(\log\frac{P(x_2|x_1)}{\mu_U(x_2|x_1)}) + \ldots + E(\log\frac{P(x_t|x_1...x_{t-1})}{\mu_U(x_t|x_1...x_{t-1})})\}\ =\ 0,$$

$$ii)\ \lim_{t\to\infty} E(\frac{1}{t}\ \sum_{i=0}^{t-1}(P(x_{i+1}|x_1...x_i) - \mu_U(x_{i+1}|x_1...x_i))^2)\ =\ 0\,,$$

and

$$iii)\ \lim_{t\to\infty} E(\frac{1}{t}\ \sum_{i=0}^{t-1}|P(x_{i+1}|x_1...x_i) - \mu_U(x_{i+1}|x_1...x_i)|)\ =\ 0\,.$$

Proof. i) immediately follows from the second statement of the previous theorem and properties of log. The statement ii) can be proven as follows:

$$\lim_{t\to\infty} E(\frac{1}{t}\ \sum_{i=0}^{t-1}(P(x_{i+1}|x_1\ldots x_i) - \mu_U(x_{i+1}|x_1\ldots x_i))^2) =$$

$$\lim_{t\to\infty}\frac{1}{t}\sum_{i=0}^{t-1}\sum_{x_1...x_i\in A^i} P(x_1\ldots x_i)(\sum_{a\in A}|P(a|x_1\ldots x_i) - \mu_U(a|x_1\ldots x_i)|)^2 \le$$

$$\lim_{t\to\infty}\frac{const}{t}\sum_{i=0}^{t-1}\sum_{x_1...x_i\in A^i} P(x_1\ldots x_i)\sum_{a\in A}P(a|x_1\ldots x_i)\log\frac{P(a|x_1\ldots x_i)}{\mu_U(a|x_1\ldots x_i)} =$$

$$\lim_{t\to\infty}(\frac{const}{t}\sum_{x_1...x_t\in A^t} P(x_1\ldots x_t)\log(P(x_1\ldots x_t)/\mu(x_1\ldots x_t))).$$

Here the first inequality is obvious, the second follows from the Pinsker's inequality (4), the others from properties of expectation and log. iii) can be derived from ii) and the Jensen inequality for the function x^2. □

Comment 1. The measure R described above has one additional property if it is used for prediction. Namely, for any Markov process P ($P \in M^*(A)$) the following is true:

$$\lim_{t\to\infty}\ \log\frac{P(x_{t+1}|x_1...x_t)}{R(x_{t+1}|x_1...x_t)} = 0$$

with probability 1, where $R(x_{t+1}|x_1...x_t) = R(x_1...x_tx_{t+1})/R(x_1...x_t)$; see [42].

Comment 2. In fact, the statements ii) and iii) are equivalent, because one of them follows from the other. For details see Lemma 2 in [47].

3.3 Problems with side information

Now we consider so-called problems with side information, which are described as follows: there is a stationary and ergodic source, whose alphabet A is presented as a product $A = X \times Y$. We are given a sequence $(x_1, y_1), \ldots, (x_{t-1}, y_{t-1})$ and so-called side information y_t. The goal is to predict, or estimate, x_t. This problem arises in statistical decision theory, pattern recognition, and machine learning, see [29]. Obviously, if someone knows the conditional probabilities $P(x_t|(x_1, y_1), \ldots, (x_{t-1}, y_{t-1}), y_t)$ for all $x_t \in X$, he has all information about x_t, available before x_t is known. That is why we will look for the best (or, at least, good) estimators for this conditional probabilities. Our solution will be based on results obtained in the parts 3.1 and 3.2. More precisely, for any universal code U and the corresponding measure μ_U (22) we define the following estimate for the problem with side information:

$$\mu_U(x_t|(x_1, y_1), \ldots, (x_{t-1}, y_{t-1}), y_t) = \frac{\mu_U((x_1, y_1), \ldots, (x_{t-1}, y_{t-1}), (x_t, y_t))}{\sum_{x_t \in X} \mu_U((x_1, y_1), \ldots, (x_{t-1}, y_{t-1}), (x_t, y_t))}.$$

Theorem 5. *Let U be a universal code and P any stationary and ergodic process. Then*

$$i) \lim_{t\to\infty} \frac{1}{t} \{E(\log \frac{P(x_1|y_1)}{\mu_U(x_1|y_1)}) + E(\log \frac{P(x_2|(x_1, y_1), y_2)}{\mu_U(x_2|(x_1, y_1), y_2)}) + \ldots$$
$$+E(\log \frac{P(x_t|(x_1, y_1), ..., (x_{t-1}, y_{t-1}), y_t)}{\mu_U(x_t|(x_1, y_1), ..., (x_{t-1}, y_{t-1}), y_t)})\} = 0,$$
$$ii) \lim_{t\to\infty} E(\frac{1}{t} \sum_{i=0}^{t-1} (P(x_{i+1}|(x_1, y_1), ..., (x_i, y_i), y_{i+1})) -$$
$$\mu_U(x_{i+1}|(x_1, y_1), ..., (x_i, y_i), y_{i+1}))^2) = 0\,,$$

and

$$iii) \lim_{t\to\infty} E(\frac{1}{t} \sum_{i=0}^{t-1} |P(x_{i+1}|(x_1, y_1), ..., (x_i, y_i), y_{i+1})) -$$
$$\mu_U(x_{i+1}|(x_1, y_1), ..., (x_i, y_i), y_{i+1})|) = 0\,.$$

Proof. The following inequality follows from the nonnegativity of the KL divergency (see (4)), whereas equality is obvious.

$$E(\log \frac{P(x_1|y_1)}{\mu_U(x_1|y_1)}) + E(\log \frac{P(x_2|(x_1, y_1), y_2)}{\mu_U(x_2|(x_1, y_1), y_2)}) + \ldots \le E(\log \frac{P(y_1)}{\mu_U(y_1)})$$
$$+E(\log \frac{P(x_1|y_1)}{\mu_U(x_1|y_1)}) + E(\log \frac{P(y_2|(x_1, y_1)}{\mu_U(y_2|(x_1, y_1)}) + E(\log \frac{P(x_2|(x_1, y_1), y_2)}{\mu_U(x_2|(x_1, y_1), y_2)}) + \ldots$$
$$= E(\log \frac{P(x_1, y_1)}{\mu_U(x_1, y_1)}) + E(\log \frac{P((x_2, y_2)|(x_1, y_1))}{\mu_U((x_2, y_2)|(x_1, y_1))}) + ...$$

Now we can apply the first statement of Theorem 4 to the last sum as follows:

$$\lim_{t\to\infty} \frac{1}{t} E(\log \frac{P(x_1, y_1)}{\mu_U(x_1, y_1)}) + E(\log \frac{P((x_2, y_2)|(x_1, y_1))}{\mu_U((x_2, y_2)|(x_1, y_1))}) + ...$$
$$E(\log \frac{P((x_t, y_t)|(x_1, y_1) \ldots (x_{t-1}, y_{t-1}))}{\mu_U((x_t, y_t)|(x_1, y_1) \ldots (x_{t-1}, y_{t-1}))}) = 0.$$

From this equality and last inequality we obtain the proof of i). The proof of the second statement can be obtained from the similar representation for ii) and the second statement of Theorem 4. iii) can be derived from ii) and the Jensen inequality for the function x^2. □

3.4 The case of several independent samples

Now we extend our consideration to the case where the sample is presented as several independent samples $x^1 = x^1_1 \dots x^1_{t_1}$, $x^2 = x^2_1 \dots x^2_{t_2}$, ..., $x^r = x^r_1 \dots x^r_{t_r}$ generated by a source. More precisely, we will suppose that all sequences were independently created by one stationary and ergodic source. (As it was mentioned above, it is impossible just to combine all samples into one, if the source is not i.i.d.) We denote this sample by $x^1 \diamond x^2 \diamond \dots \diamond x^r$ and define $\nu_{x^1 \diamond x^2 \diamond \dots \diamond x^r}(v) = \sum_{i=1}^r \nu_{x^i}(v)$. For example, if $x^1 = 0010, x^2 = 011$, then $\nu_{x^1 \diamond x^2}(00) = 1$. The definition of K_m and R can be extended to this case:

$$K_m(x^1 \diamond x^2 \diamond \dots \diamond x^r) = \tag{24}$$

$$(\prod_{i=1}^r |A|^{-\min\{m,t_i\}}) \prod_{v \in A^m} \frac{\prod_{a \in A} ((\Gamma(\nu_{x^1 \diamond x^2 \diamond \dots \diamond x^r}(va) + 1/2) / \Gamma(1/2))}{(\Gamma(\bar{\nu}_{x^1 \diamond x^2 \diamond \dots \diamond x^r}(v) + |A|/2) / \Gamma(|A|/2))},$$

whereas the definition of R is the same (see (17)). (Here, as before, $\bar{\nu}_{x^1 \diamond x^2 \diamond \dots \diamond x^r}(v) = \sum_{a \in A} \nu_{x^1 \diamond x^2 \diamond \dots \diamond x^r}(va)$. Note, that $\bar{\nu}_{x^1 \diamond x^2 \diamond \dots \diamond x^r}(\,) = \sum_{i=1}^r t_i$ if $m = 0$.)

The following example is intended to show the difference between the case of many samples and one. Let there be two independent samples $y = y_1 \dots y_4 = 0101$ and $x = x_1 \dots x_3 = 101$, generated by a stationary and ergodic source with the alphabet $\{0, 1\}$. One wants to estimate the (limiting) probabilities $P(z_1 z_2), z_1, z_2 \in \{0, 1\}$ (here $z_1 z_2 \dots$ can be considered as an independent sequence, generated by the source) and predict $x_4 x_5$ (i.e. estimate conditional probability $P(x_4 x_5 | x_1 \dots x_3 = 101, y_1 \dots y_4 = 0101)$. For solving both problems we will use the measure R (see (17)). First we consider the case where $P(z_1 z_2)$ is to be estimated without knowledge of sequences x and y. From (11) and (15) we obtain:

$$K_0(00) = K_0(11) = \frac{1/2}{1}\,\frac{3/2}{1+1} = 3/8, \;\; K_0(01) = K_0(10) = \frac{1/2}{1+0}\,\frac{1/2}{1+1} = 1/8,$$

$$K_i(00) = K_i(01) = K_i(10) = K_i(11) = 1/4; \;, \; i \geq 1.$$

Having taken into account the definitions of ω_i (16) and the measure R (17), we can calculate $R(z_1 z_2)$ as follows:

$$R(00) = \omega_1 K_0(00) + \omega_2 K_1(00) + \dots = (1 - 1/\log 3)\, 3/8 + (1/\log 3 - 1/\log 4)\, 1/4 +$$

$$(1/\log 4 - 1/\log 5)\, 1/4 + \;\dots = (1 - 1/\log 3)\;\, 3/8 + (1/\log 3)\;\, 1/4 \approx 0.296.$$

Analogously, $R(01) = R(10) \approx 0.204, R(11) \approx 0.296$.

Let us now estimate the probability $P(z_1 z_2)$ taking into account that there are two independent samples $y = y_1 \dots y_4 = 0101$ and $x = x_1 \dots x_3 = 101$. First of all we note that such estimates are based on the formula for conditional probabilities:

$$R(z | x \diamond y) = R(x \diamond y \diamond z) / R(x \diamond y).$$

First we estimate the frequencies :

$$\nu_{0101 \diamond 101}(0) = 3, \nu_{0101 \diamond 101}(1) = 4, \nu_{0101 \diamond 101}(00) = \nu_{0101 \diamond 101}(11) = 0, \nu_{0101 \diamond 101}(01) = 3,$$

$$\nu_{0101 \diamond 101}(10) = 2, \nu_{0101 \diamond 101}(010) = 1, \nu_{0101 \diamond 101}(101) = 2, \nu_{0101 \diamond 101}(0101) = 1,$$

whereas frequencies of all other three-letters and four-letters words are 0. Then we calculate :

$$K_0(\,0101 \diamond 101) = \frac{1}{2}\frac{3}{4}\frac{5}{6}\frac{7}{8}\frac{1}{10}\frac{3}{12}\frac{5}{14} \approx 0.00244, K_1(\,0101 \diamond 101) = (2^{-1})^2\, \frac{1}{2}\frac{3}{4}\frac{5}{6}\;\, 1\, \frac{1}{2}\frac{3}{4}\;\, 1$$

$$\approx 0.0293, \quad K_2(\,0101 \diamond 101) \approx 0.01172, \quad K_i(\,0101 \diamond 101) = 2^{-7},\ i \geq 3,$$

$$R(\,0101 \diamond 101) = \omega_1 K_0(\,0101 \diamond 101) + \omega_2 K_1(\,0101 \diamond 101) + \ \ldots \approx$$

$$0.369\,0.00244 \ + \ 0.131\,0.0293 \ + 0.06932\,0.01172 \ + 2^{-7} \ / \ \log 5 \approx 0.0089.$$

In order to avoid repetitions, we estimate only one probability $P(z_1 z_2 = 01)$. Carrying out similar calculations, we obtain

$$R(0101 \diamond 101 \diamond 01) \approx 0.00292,$$

$$R(z_1 z_2 = 01 | y_1 \ldots y_4 = 0101, x_1 \ldots x_3 = 101) =$$

$$R(0101 \diamond 101 \diamond 01) / R(\,0101 \diamond 101) \approx 0.32812.$$

If we compare this value and the estimation $R(01) \approx 0.204$, which is not based on the knowledge of samples x and y, we can see that the measure R uses additional information quite naturally (indeed, 01 is quite frequent in $y = y_1 \ldots y_4 = 0101$ and $x = x_1 \ldots x_3 = 101$).

Such generalization can be applied for many universal codes, but, generally speaking, there exist codes U for which $U(x^1 \diamond x^2)$ is not defined and, hence, the measure $\mu_U(x_1 \diamond x_2)$ is not defined. That is why we will describe properties of R, but do not describe properties of universal codes in general. For the measure R all asymptotic properties are the same for the cases of one sample and several samples. More precisely, the following statement is true:

Claim 2. *Let $x^1, x^2, ..., x^r$ be independent sequences generated by a stationary and ergodic source and t be a total length of those sequences ($t = \sum_{i=1}^{r} |x^i|$). Then, if $t \to \infty$, (and r is fixed) the statements of Theorems 1–5 are valid, when applied to $x^1 \diamond x^2 \diamond ... \diamond x^r$ instead of $x_1 \ldots x_t$. (In Theorems 2, 4, 5 μ_U should be changed to R.)*

The proofs are analogous to the proofs of Theorems 1–5.

4 Real-valued time series

Here we will consider problems of the density estimation and prediction for a stationary process with densities. We have seen that Shannon-MacMillan-Breiman theorem played a key role in the case of finite-alphabet processes. In this part we will use its generalization to the processes with densities. This result was proved by Barron [2] and was an extension of the $L1$ convergence obtained in [19, 30, 32]. First we describe considered processes with some properties needed for a fulfilment of the generalized Shannon-MacMillan-Breiman theorem.

Let (Ω, F, P) be a probability space and let $X_1, X_2, \ldots$ be a stochastic process with each X_t taking values in a standard Borel space. As in [2], suppose that the joint distribution P_n for $(X_1, X_2, \ldots, X_n)$ has a probability density function $p(x_1 x_2 \ldots x_n)$ with respect to a sigma-finite measure M_n. Assume that the sequence of dominating measures M_n is Markovian of order $m \geq 0$ with a stationary transition measure. Familiar cases for M_n are Lebesgue measure and counting measure. Let $p(x_{n+1}|x_1 \ldots x_n)$ denote the conditional density given by the ratio $p(x_1 \ldots x_{n+1})/p(x_1 \ldots x_n)$ for $n > 1$. It is known that for stationary and ergodic processes there exists a so- called relative entropy rate h defined by

$$h = \lim_{n \to \infty} -E(\log p(x_{n+1}|x_1 \ldots x_n)), \tag{25}$$

where E denotes expectation with respect to P. The following generalization of the Shannon-MacMillan-Breiman theorem is obtained by Barron in [2]:

Claim 3. *If $\{X_n\}$ is a stationary ergodic process with density $p(x_1 \dots x_n) = dP_n/dM_n$ and $h_n < 1$ for some $n \geq m$, the sequence of relative entropy densities*

$$-(1/n) \log p(x_1 \dots x_n)$$

converges almost surely to the relative entropy rate, i.e.,

$$\lim_{t\to\infty} \frac{1}{t} \log p(x_1 \dots x_t) = h. \tag{26}$$

with probability 1 (according to P).

Now we return to the estimation problems. Let $\{\Pi_n\}, n \geq 1$, be an increasing sequence of finite partitions of Ω that asymptotically generates the Borel sigma-field on F, and let $x^{[k]}$ denote the element of Π_k that contains the point x. (Informally, if Ω is an interval, $x^{[k]}$ is obtained by quantizing x to k bits of precision). For integers s and n we define the following approximation of the density

$$p^s(x_1, \dots, x_n) = P(x_1^{[s]}, \dots, x_n^{[s]})/M_n(x_1^{[s]} \dots x_n^{[s]}). \tag{27}$$

We also consider

$$h_s = \lim_{n\to\infty} E(\log p^s(x_{n+1}|x_1, \dots, x_n)). \tag{28}$$

Applying Claim 3 to the density $p^s(x_1, \dots, x_t)$, we obtain that a.s.

$$\lim_{t\to\infty} \frac{1}{t} \log p^s(x_1, \dots, x_t) = h_s. \tag{29}$$

Let U be a universal code, which is defined for any finite alphabet. In order to describe the density estimate we will use the distribution ω; see (16). Now we define the corresponding density r_U as follows:

$$r_U(x_1 \dots x_t) = \sum_{i=0}^{\infty} \omega_i \mu_U(x_1^{[i]} \dots x_t^{[i]})|/M_t(x_1^{[i]} \dots x_t^{[i]}) \,, \tag{30}$$

where the measure μ_U is defined by (22). (It is supposed here that the code $U(x_1^{[i]} \dots x_t^{[i]})$ is defined for the alphabet, which contains $|\Pi_i|$ letters.)

It turns out that, in a certain sense, the density $r_U(x_1 \dots x_t)$ estimates the unknown density $p(x_1, \dots, x_t)$.

Theorem 6. *Let X_t be a stationary ergodic process with densities $p(x_1 \dots x_t) = dP_t/dM_t$ such that*

$$\lim_{s\to\infty} h_s = h < \infty, \tag{31}$$

where h and h_s are relative entropy rates, see (25), (28). Then

$$\lim_{t\to\infty} \frac{1}{t} \log \frac{p(x_1 \dots x_t)}{r_U(x_1 \dots x_t)} = 0 \tag{32}$$

with probability 1 and

$$\lim_{t\to\infty} \frac{1}{t} E(\log \frac{p(x_1 \dots x_t)}{r_U(x_1 \dots x_t)}) \;=\; 0 \,. \tag{33}$$

Proof. First we note that for any integer s the following obvious equality is true: $r_U(x_1 \dots x_t) = \omega_s \mu_U(x_1^{[s]} \dots x_t^{[s]})/M_t(x_1^{[s]} \dots x_t^{[s]})\,(1+\delta)$ for some $\delta > 0$. From this equality, (22) and (32) we immediately obtain that a.s.

$$\lim_{t\to\infty} \frac{1}{t} \log \frac{p(x_1 \dots x_t)}{r_U(x_1 \dots x_t)} \;\le \lim_{t\to\infty} \frac{1}{t} \log \frac{p(x_1 \dots x_t)}{2^{-|U(x_1^{[s]} \dots x_t^{[s]})|}/M_t(x_1^{[s]} \dots x_t^{[s]})}\,. \tag{34}$$

The right part can be presented as follows:

$$\lim_{t\to\infty} \frac{1}{t} \log \frac{p(x_1 \dots x_t)}{2^{-|U(x_1^{[s]} \dots x_t^{[s]})|}/M_t(x_1^{[s]} \dots x_t^{[s]})}$$

$$= \lim_{t\to\infty} \frac{1}{t} \log \frac{p^s(x_1 \dots x_t)\, M_t(x_1^{[s]} \dots x_t^{[s]})}{2^{-|U(x_1^{[s]} \dots x_t^{[s]})|}} + \lim_{t\to\infty} \frac{1}{t} \log \frac{p(x_1 \dots x_t)}{p^s(x_1 \dots x_t)}. \tag{35}$$

Having taken into account that U is the universal code, (27) and Theorem 1, we can see that the first term equals to zero. From (26) and (29) we can see that a.s. the second term is equal to $h_s - h$. This equality is valid for any integer s and, according to (31), the second term equals to zero too, and we obtain (33). The first statement is proven.

From (34) and (35) we can can see that

$$E \log \frac{p(x_1 \dots x_t)}{r_U(x_1 \dots x_t)} \le E \log \frac{p_t^s(x_1, \dots, x_t)\, M_t(x_1^{[s]} \dots x_t^{[s]})}{2^{-|U(x_1^{[s]} \dots x_t^{[s]})|}}$$

$$+E \log \frac{p(x_1 \dots x_t)}{p^s(x_1, \dots, x_t)}. \tag{36}$$

The first term is the average redundancy of the universal code for a finite- alphabet source, hence, according to Theorem 1, it tends to 0. The second term tends to $h_s - h$ for any s and from (31) we can see that it is equal to zero. The second statement is proven. □

We have seen that the requirement (31) plays an important role in the proof. A natural question is whether there exist processes for which (31) is valid. The answer is positive. For example, let Ω be an interval $[-1, 1]$, M_n be Lebesgue measure and a considered process is Markovian with conditional density

$$p(x|y) = \begin{cases} 1/2 + \alpha \;\; sign(y), & \text{if } x < 0 \\ 1/2 - \alpha \;\; sign(y), & \text{if } x \ge 0 \end{cases},$$

where $\alpha \in (0, 1)$ is a parameter and

$$sign(y) = \begin{cases} -1, & \text{if } y < 0, \\ 1, & \text{if } y \ge 0 \end{cases}.$$

It is easy to see that (31) is true for any $\alpha \in (0, 1/2)$.

The following theorem describes properties of conditional probabilities $r_U(x|x_1 ... x_m) = r_U(x_1 ... x_m x)/r_U(x_1 ... x_m)$ which, in turn, is connected with the prediction problem. We will see that the conditional density $r_U(x|x_1 ... x_m)$ is a reasonable estimation of $p(x|x_1 ... x_m)$.

Theorem 7. *Let f be an integrable function whose absolute value is bounded by a certain constant $\bar{M}$. Then the following equalities are valid:*

$$i) \lim_{t\to\infty} \frac{1}{t} E\Big(\sum_{m=0}^{t-1} \Big(\int f(x) p(x|x_1 ... x_m) dM_m - \int f(x) r_U(x|x_1 ... x_m) dM_m\Big)^2\Big) = 0, \tag{37}$$

$$ii)\ \lim_{t\to\infty}\frac{1}{t}E(\sum_{m=0}^{t-1}|\int f(x)p(x|x_1...x_m)dM_m-\int f(x)r_U(x|x_1...x_m)dM_m|)=0.$$

Proof. The last inequality of the following chain follows from the Pinsker's one, whereas all others are obvious.

$$(\int f(x)p(x|x_1...x_m)dM_m-\int f(x)r_U(x|x_1...x_m)dM_m)^2=$$

$$(\int f(x)(p(x|x_1...x_m)-r_U(x|x_1...x_m))dM_m)^2$$

$$\le \bar{M}^2(\int(p(x|x_1...x_m)-r_U(x|x_1...x_m))dM_m)^2$$

$$\le \bar{M}^2(\int|p(x|x_1...x_m)-r_U(x|x_1...x_m)|dM_m)^2\le$$

$$const\int p(x|x_1...x_m)\log(p(x|x_1...x_m)/r_U(x|x_1...x_m)dM_m.$$

From these inequalities we obtain:

$$\sum_{m=0}^{t-1}E(\int f(x)p(x|x_1...x_m)dM_m-\int f(x)r_U(x|x_1...x_m)dM_m)^2)\le \tag{38}$$

$$\sum_{m=0}^{t-1}const\,E(\int p(x|x_1...x_m)\log(p(x|x_1...x_m)/r_U(x|x_1...x_m))dM_m.$$

The last term can be presented as follows:

$$\sum_{m=0}^{t-1}E(\int p(x|x_1...x_m)\log(p(x|x_1...x_m)/r_U(x|x_1...x_m))dM_m)=$$

$$\sum_{m=0}^{t-1}\int p(x_1...x_m)\int p(x|x_1...x_m)\log(p(x|x_1...x_m)/r_U(x|x_1...x_m))dM_1\,dM_m)=$$

$$\int p(x_1...x_t)\log(p(x_1...x_t)/r_U(x_1...x_t))dM_t.$$

From this equality, (38) and (33) we obtain (37). ii) can be derived from (38) and the Jensen inequality for x^2. □

References

[1] P.Algoet, "Universal Schemes for Learning the Best Nonlinear Predictor Given the Infinite Past and Side Information", *IEEE Trans. Inform. Theory,* vol. 45, pp. 1165–1185, 1999.

[2] A.R. Barron, "The strong ergodic theorem for dencities: generalized Shannon-McMillan-Breiman theorem", *The annals of Probability*, vol. 13, pp.1292–1303, 1985.

[3] P. Billingsley, *Ergodic theory and information,* John Wiley & Sons, 1965.

[4] R. Cilibrasi and P.M.B. Vitanyi, "Clustering by Compression," *IEEE Transactions on Information Theory,* vol. 51, 2005.

[5] R. Cilibrasi, R. de Wolf and P.M.B. Vitanyi, "Algorithmic Clustering of Music," *Computer Music Journal,* vol. 28, pp.49–67, 2004.

[6] T.M. Cover and J.A. Thomas, *Elements of Information Theory,* John Wiley and sons, 1991.

[7] I. Csiszár, J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems,* Budapesht, Akadémiai Kiadó, 1981.

[8] I. Csiszár, P. Shields, "The consistency of the BIC Markov order estimation", *Annals of Statistics,* vol. 6, pp. 1601–1619, 2000.

[9] G.A. Darbellay, I. Vajda, "Entropy expressions for multivariate continuous distributions," in. Research Report no 1920, UTIA, Academy of Science, Prague (library@utia.cas.cz), 1998.

[10] G.A. Darbellay, I. Vajda, "Estimation of the mutual information with data-dependent partitions," *IEEE Trans. Inform. Theory,* vol. 48, pp. 1061–1081, 1999.

[11] M. Effros, K. Visweswariah, S.R. Kulkarni and S. Verdu, "Universal lossless source coding with the Burrows Wheeler transform," *IEEE Trans. Inform. Theory,* vol. 45, pp. 1315–1321, 1999.

[12] W. Feller, *An Introduction to Probabability Theory and Its Applications, vol.1*, John Wiley & Sons, New York, 1970.

[13] L. Finesso, Chuang-Chun Liu, P. Narayan, "The optimal error exponent for Markov order estimation," *IEEE Trans. on Information Theory,* vol. 42, pp. 1488–1497, 1996.

[14] B.M. Fitingof, "Optimal encoding for unknown and changing statistica of messages," *Problems of Information Transmission*, vol.2, n. 2, pp. 3–11, 1966.

[15] R.G. Gallager, *Information Theory and Reliable Communication,* John Wiley & Sons, New York, 1968.

[16] L. Gyorfi, G. Morvai and S.J. Yakowitz, "Limits to consistent on-line forecasting for ergodic time series," *IEEE Transactions on nformation Theory,* vol. 44, pp.886– 892, 1998.

[17] P. Jacquet, W. Szpankowski and L. Apostol, "Universal predictor based on pattern matching," *IEEE Trans. Inform. Theory,* vol.48, pp. 1462–1472, 2002.

[18] J.L. Kelly, "A new interpretation of information rate," *Bell System Tech. J.*, vol. 35, pp. 917–926, 1956.

[19] J. Kieffer, "A simple proof of the Moy-Perez generalization of the Shannon- MacMillan theorem," *Pacific J. Math.*, vol.51, pp. 203–206, 1974.

[20] J. Kieffer, *Prediction and Information Theory,* Preprint, 1998. (available at ftp://oz.ee.umn.edu/users/kieffer/papers/prediction.pdf/)

[21] J.C. Kieffer and En-Hui Yang, "Grammar-based codes: a new class of universal lossless source codes," *IEEE Transactions on Information Theory,* vol.46, pp.737 – 754, 2000.

[22] D.E. Knuth, *The art of computer programming,* Vol.2. Addison Wesley, 1981.

[23] A.N. Kolmogorov, "Three approaches to the quantitative definition of information," *Problems of Inform. Transmission,* vol. 1, pp.3–11, 1965.

[24] G. Korodi, I. Tabus, J. Rissanen and J. Astola, "DNA sequence compression - based on the normalized maximum likelihood model," *IEEE Signal Processing Magazine,* vol. 24, pp.47 – 53, 2007.

[25] R. Krichevsky, "A relation between the plausibility of information about a source and encoding redundancy," *Problems Inform. Transmission,* vol. 4, n.3, pp. 48–57, 1968.

[26] R. Krichevsky, *Universal Compression and Retrival.* Kluver Academic Publishers, 1993.

[27] S. Kullback, *Information Theory and Statistics*, Wiley, New York, 1959.

[28] D.S. Modha and E. Masry, "Memory-universal prediction of stationary random processes," *IEEE Trans. Inform. Theory,* vol. 44, pp. 117–133, 1998.

[29] G. Morvai, S.J. Yakowitz and P.H. Algoet, "Weakly convergent nonparametric forecasting of stationary time series," *IEEE Trans. Inform. Theory,* vol. 43, pp.483 – 498, 1997.

[30] S.C. Moy, "Generalisations of Shannon-MacMillan theorem," *Pacific J. Math.*, vol. 11, pp.705–714, 1961.

[31] A.B. Nobel, "On optimal sequential prediction," *IEEE Trans. Inform. Theory,* vol. 49, pp. 83–98, 2003.

[32] A. Perez, "Extensions of Shannon-MacMillan's limit theorem to more general stohastic processes," in *Trans. Third Prague Conference on Information Theory, Statistical Decision Functions, and Random Processes,* . Czehoslovak Academy of Sciences, Prague, 1964, pp. 545–574.

[33] J. Rissanen, "Generalized Kraft inequality and arithmetic coding," *IBM J. Res. Dev.*, vol. 20, n. 5, pp. 198–203, 1976.

[34] J. Rissanen, "Modeling by shortest data description," *Automatica,* vol.14, pp. 465–471, 1978.

[35] J. Rissanen, "Universal coding, information, prediction, and estimation", *IEEE Trans. Inform. Theory,* vol. 30, pp. 629–636, 1984.

[36] J. Rissanen, *Stochastic Complexity in Statistical Inquiry,* World Scientific Publishing Co., Singapore, 1989.

[37] J. Rissanen, "Fisher information and stochastic complexity," *IEEE Trans. Inform. Theory*, vol.42, n.1, pp. 40–47 1996.

[38] J. Rissanen, *Information and complexity in statistical modeling,* Springer Verlag, 2007.

[39] A. Rukhin and others. *A statistical test suite for random and pseudorandom number generators for cryptographic applications,* NIST Special Publication 800- 22 (with revision dated May,15,2001). http://csrc.nist.gov/rng/SP800-22b.pdf

[40] B.Ya. Ryabko, "Twice-universal coding," *Problems of Information Transmission,* vol.20, n.3, pp. 173–177, 1984.

[41] B.Ya. Ryabko, "Prediction of random sequences and universal coding," *Problems of Inform. Transmission,* vol. 24, n.2, pp. 87–96, 1988.

[42] B.Ya. Ryabko, "A fast adaptive coding algorithm," *Problems of Inform. Transmission,* vol. 26, pp. 305–317, 1990.

[43] B. Ya. Ryabko, "The complexity and effectiveness of prediction algorithms," *J. Complexity*, vol. 10, n.3, pp. 281–295, 1994.

[44] B. Ryabko, J. Astola and A. Gammerman, "Application of Kolmogorov complexity and universal codes to identity testing and nonparametric testing of serial independence for time series," *Theoretical Computer Science,* vol. 359, pp.440–448, 2006.

[45] B. Ya. Ryabko and V.A. Monarev, "Using information theory approach to randomness testing," *Journal of Statistical Planning and Inference,* vol. 133, n.1, pp. 95–110, 2005.

[46] B. Ryabko and F. Topsoe, "On Asymptotically Optimal Methods of Prediction and Adaptive Coding for Markov Sources," *Journal of Complexity,* vol. 18, n.1, pp. 224–241, 2002.

[47] D. Ryabko and M. Hutter, "Sequence prediction for non-stationary processes." In Proceedings IEEE International Symposium on Information Theory, 2006. pp. 2346-2350. (see also http://arxiv.org/pdf/cs.LG/0606077)

[48] S. A. Savari, "A probabilistic approach to some asymptotics in noiseless communication," *IEEE Transactions on Information Theory,* vol. 46, pp. 1246-1262, 2000.

[49] C. E. Shannon, "A mathematical theory of communication," *Bell Sys. Tech. J.,* vol. 27, pp. 379–423 and pp.623–656, 1948.

[50] C.E. Shannon, "Communication theory of secrecy systems," *Bell Sys. Tech. J.,* vol. 28, pp. 656–715, 1948.

[51] P.C. Shields, "The interactions between ergodic theory and information theory," *IEEE Transactions on Information Theory,* vol. 44, pp.2079–2093, 1998.

[52] W. Szpankowsky, *Average case analysis of algorithms on sequences,* John Wiley and Sons, New York, 2001.

[53] P.M.B. Vitanyi and M. Li, "Minimum description length induction, Bayesianism, and Kolmogorov complexity," *IEEE Trans. Inform. Theory,* vol.46, pp. 446–464, 2000.

That Simple Device Already Used by Gauss

Peter Grünwald
CWI Amsterdam
the Netherlands
www.cwi.nl/~pdg

Abstract

From November 1998 until September 1999, Jorma Rissanen and I met on a regular basis. Here I recall some of our stimulating conversations and some of the work that we did together. This work, based almost exclusively on a single page of [12], was left unfinished and has never been published, but it has indirectly had a profound impact on my career.

1 Meet Jorma Rissanen

I first met Jorma in November 1998. I had just obtained my Ph.D. in Amsterdam and started a postdoc at Stanford University. These were exciting times: it was at the height of the dot-com boom, and Stanford was right in the middle of it. Since my thesis was all about the MDL Principle, I had suggested that Jorma and I could meet in person during my stay in California. Jorma replied that he would like to. I was delighted, honored but also a bit worried, since I had been forewarned that Jorma was not your "usual" kind of scientist...

San Francisco, Category Theory and Statistics I found that, while Jorma was not one for polite small talk, he did like having lots of beer with a small circle of academic friends (for most scientists it seems to be the other way around). He didn't talk much, but what he said was invariably to the point, direct and frank. During our first meeting, when I told him that I lived in San Francisco since it seemed to me so much nicer than Palo Alto, his immediate reply was "I don't like San Francisco". Later that day, we talked about science in general, and I was quite surprised to learn that, in his first years at IBM, Jorma had been a serious student of category theory – he even published on it while he was professor in Sweden [15]. He went on to say that he found category theory to be *much* easier than statistics, the field to which he had made such major contributions. When I asked him why, he said *"because much of statistics is nonsense. It is exceedingly hard to teach yourself nonsense!"*. Vintage Jorma: blunt and sharp at the same time. Some fear him for this, others, like me, find Jorma's conversation delightfully refreshing.

> Jorma is, in fact, notorious for his strong opinions about sub-fields of mathematics – when a Ph.D. student once told him that he had spent a lot of time studying complex function theory, Jorma exclaimed (not entirely seriously, I believe) *"you've wasted your youth!"*

Jorma could be as harsh about his own work as he could be about others work – I vividly recall him saying "how could I have been so stupid" when I suggested that the central proof in [13] was much more difficult than was needed. When I told him that I did not quite understand another proof of his, in [11], he told me that he himself only understands it some of the time, "when I have one of my better days". I am referring to the proof of what is perhaps his most well-known theorem: the central result of [11], which is an extension of the information inequality, but, from a statistical point of view, can also be interpreted as — as Jorma put it, now less modestly but entirely correctly — "a grand Cramér-Rao theorem."

Soccer and Impounded Cars Our first meeting went quite well, and I was honored that Jorma asked me to visit him again. In the end, I visited him every 3-4 weeks during my year at Stanford - at the time, Jorma was still at IBM Almaden, in San Jose, just 40 minutes further down the highway. I usually arrived at 10 in the morning and stayed for the whole day. Between 12 and 2 however, Jorma would often leave to play soccer with a group of friends, something he did three times a week. At the time he was 67, but still very good at it: in his youth Jorma had seriously considered becoming a professional soccer player. In the end, he had decided to pursue his Ph.D. instead – as Jorma told me during one of my visits, he still doesn't know whether he has made the right choice. When Jorma went out for soccer, I used to have lunch at IBM's luxurious cafeteria, paid for by Jorma's card. On one occasion, my car had broken down, and Jorma told me that Nemo, one of his soccer friends, could probably arrange a good new car for me for as little as $ 150: Nemo was a car dealer who took over impounded cars from the police if they hadn't been picked up for more than a year. I was fascinated: a brilliant scientist who always speaks his mind, played soccer at world-class level and counts impounded car dealers among his friends.

During my visits, Jorma and I also did some work together. Rather than actually working on a joint publication, we were both pursuing related but distinct ideas, always discussing our latest progress during our meetings. So what did we work on?

2 Prediction is Coding

I learned about the MDL Principle from the monograph "Stochastic Complexity in Statistical Inquiry" [12], the "little green book," in which Jorma so eloquently puts forward the main ideas underlying the MDL Principle. In Chapter 2, Jorma notes that different research communities have a different understanding of the concept of a "model". In statistics, a model is usually a family of probability distributions, for example, the Gaussian or normal family. In other fields such as pattern recognition and machine learning, a model is usually a family of deterministic hypotheses or *predictors*. For example, we may try to find a relationship between a variable X, taking values in some set $\mathcal{X}$, and a variable Y, taking values in $\mathcal{Y}$, by considering a "model" $\mathcal{F}$, which is really a family of deterministic functions $f : \mathcal{X} \to \mathcal{Y}$. f is then

fit to the data $(x_1, y_1), \ldots, (x_n, y_n)$, using, for example, the least squares criterion; concrete examples are given below. Jorma claims that, from an MDL perspective, there is no real distinction between the probabilistic and the deterministic type of model: they may both be viewed as defining *codes* or equivalently, description methods for the data. Statistical inference should then proceed by selecting the model (set of description methods) that leads to the shortest codelength of the data.

A Simple Device Already Used by Gauss How, then, should we associate models with codes? For probabilistic models this is obvious: the Kraft inequality tells us that for every probability distribution P, there exists a uniquely decodable code such that, for all outcomes x, the codelength of x is (essentially) equal to $-\log p(x)$, p being the mass function of P. The information inequality indicates that this particular code is the *only* reasonable code that one may want to associate with P [3, Chapter 3]. But what about deterministic predictors f?

According to Jorma, we should map them to codes as follows. We first map each f to a conditional distribution p_f, defined in such a way that $-\log p_f(y \mid x)$ is an affine (linear with constant offset) function of the loss $L(y, f(x))$ that f makes when predicting y given x. We should then use the code with lengths $-\log p_f$. His remarks are worth quoting in full [[12], page 18; mathematical notation slightly adjusted, material between square brackets and emphasis added by myself]:

> ...The two views however can be reconciled by the *simple device used already by Gauss* for the distributions bearing his name. In fact, for any desired distance measure $L(y_i, \hat{y}_i)$, [and any predictor f under consideration] define a density function
>
> $$p_f(y_i \mid x_i) = Ke^{-L(y_i, f(x_i))}, \tag{1}$$
>
> where K is so chosen that p_f becomes a proper density function over the range of y_i. Taking the product of these, $p_f(y^n \mid x^n) = \prod_{i=1}^n p_f(y_i \mid x_i)$, over the observed data set, gives the desired conditional density function for sequences [...]
>
> [For example] let the data consist of a binary sequence $y^n = y_1, \ldots, y_n$. With some predictor $\hat{y}_i$ as a function of the past observations, let $L(y_i, \hat{y}_i) = 0$ if the prediction is correct, i.e., if $y_i = \hat{y}_i$, else let $L(y_i, \hat{y}_i) = 1$. The desired criterion for the goodness of the predictors is the number of mispredictions in the observed sequence. Picking in (1) the base of the exponential as 2, we get $P(1 \mid \hat{y}_i) = K$ or $K/2$, depending on what the predicted symbol is. In either case, $P(0 \mid \hat{y}_i) = 1 - P(1 \mid \hat{y}_i)$, which makes $K = 2/3$. With this the number of mistaken predictions made differs only by a constant from the quantity $-\sum_i \log p(y_i \mid \hat{y}_i)$, which is seen to be an expression in terms of probabilistic model [and corresponds to the codelength of the data according to a particular uniquely decodable code]
>
> ...As another example, with $L(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2$, (1) defines a normal density function with mean $\hat{y}_i$ and variance $1/2$. The induced normal density function for the data $p_f(y^n \mid x^n)$, as defined by its negative

logarithm, is

$$-\log p_f(y^n \mid x^n) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \frac{n}{2}\ln \pi.$$

Again we see that the sum of the squared errors differs from the negative logarithm of a density function only by a constant.

This single page constitutes *all* Jorma writes about the unification of different types of models. When I first read it (in 1995) I was highly intrigued: one is promised here a new, fully general notion of "model", encapsulating all previous ones, in which a model really becomes *a language that allows one to express particular properties of the data* [3]. This claim is bold, exciting, but only treated in the most sketchy of fashions. The examples that Jorma gave raise all kinds of questions. As will become clear, trying to answer these, and validating Jorma's claim, has, to a large extent, shaped my own career.

An immediate question that comes to mind is: why use logarithm to the base 2 in the 0/1-loss example? And, relatedly, why use variance 1/2 in the Gaussian (squared error) example? These seem to be arbitrary choices. They may be justifiable if we do have some additional *probabilistic* knowledge about the situation we are trying to model, e.g. that the errors are Gaussian with known variance 1/2. But we often want to use predictors with squared error in cases where we hardly have any probabilistic knowledge; in particular, we usually do not even want to assume normality. Does the approach still work in such cases?

Optimality To phrase this question more precisely, for a fixed class of predictors $\mathcal{F}$, a fixed loss function of interest $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ and a fixed $\beta > 0$, for each $f \in \mathcal{F}$, define an associated conditional probability distribution $P_{f,\beta}$ identified by its mass function $p_{f,\beta}(Y|X)$, as follows:

$$p_{f,\beta}(y \mid x) \equiv \frac{1}{Z(\beta)} e^{-\beta L(y;f(x))} \tag{2}$$

where $Z(\beta) \equiv \sum_{y \in \mathcal{Y}} e^{-\beta L(y;f(x))}$ is a normalization factor. (2) is extended to several outcomes by taking product distributions. For the 0/1-loss, $Z(\beta) = 1 + e^{-\beta}$; for other loss functions $Z(\beta)$ may depend on f and x; if $\mathcal{Y}$ is not countable, $p_{f,\beta}$ becomes a density with respect to some fixed underlying measure and the summation in the definition of $Z(\beta)$ becomes integration. For example, for the squared loss, with the variable substitution $\sigma^2 = 1/(2\beta)$, (2) becomes a conditional normal density and $Z(\beta) = 1/\sqrt{2\pi\sigma^2} = \sqrt{\beta/2\pi}$. We see that for every choice of $\beta > 0$, the codelength obtained by coding with p_f is an increasing affine function of the loss induced by f: for all $x^n \in \mathcal{X}^n$, all $y^n \in \mathcal{Y}^n$, we have

$$-\log p_{f,\beta}(y^n \mid x^n) = \beta \sum_{i=1}^{n} L(y_i, f(x_i)) + n \ln Z(\beta). \tag{3}$$

Thus, for each given sequence of data, and any two predictors f_1 and f_2, f_1 better fits the data in terms of L if and only if $p_{f_1,\beta}$ better fits the data than $p_{f_2,\beta}$ in terms of likelihood. For fixed β, the likelihood will be maximized for $p_{\hat{f},\beta}$, where $\hat{f}$ is the f

that minimizes, among all $f \in \mathcal{F}$, the *empirical risk* $\sum_{i=1}^{n} L(y_i, f(x_i))$. In my thesis I called this the *optimality* property of the mapping (2): the optimal (best-fitting in terms of L) $\hat{f}$ gets mapped to the optimal (best-fitting in terms of likelihood) $\hat{p}$. This is one of the main reasons why the property (3) is desirable for our mapping from deterministic f to probabilistic p.

Reliability The question that was asked above can now be rephrased as: is there a natural choice for β? After two years, in 1997, I discovered that such a natural choice exists: we should simply *learn* β from the data, using some likelihood-based method such as maximum likelihood, MDL or Bayesian inference. The resulting β has a particularly useful property which may be called *reliability*. To explain this further, assume, for example, that we use two-part code MDL [3]. If one has only vague or no prior knowledge about what β should be, the straightforward thing is to first encode some $f \in \mathcal{F}$, then encode some $\beta \geq 0$, and then encode the data with the encoded $p_{f,\beta}$, using the combination (f, β) which minimizes the total two-part code-length. In this way, we can usually gain some additional compression of the data, which indicates that β captures some interesting property of the data. This is indeed the case, as is now shown. Note first that under any reasonable method for coding β, for large n, the encoded β will be close to the ML estimator $\hat{\beta}_f$ achieving $\max_\beta p_{f,\beta}(y^n \mid x^n)$, where f is the previously encoded predictor f. This ML estimator has a very special property, and this property is what makes learning β the natural thing to do. Namely, letting $E_{f,\beta}[\cdot]$ denote expectation under $p_{f,\beta}$, we find that, no matter what data (x^n, y^n) is observed, for any fixed f, we have

$$E_{f,\hat{\beta}_f}[L(Y,X)] = \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i)). \tag{4}$$

To see this, simply differentiate the minus log-likelihood (3) with respect to β. The minimum is achieved if we set the derivative to 0, and this gives, for arbitrary y',

$$\sum_{i=1}^{n} L(y_i, f(x_i)) + n\frac{-\sum_y L(y,y')e^{-\beta L(y,y')}}{Z(\beta)} = 0, \tag{5}$$

from which (4) follows. Note that the $p_{f,\beta}$ are really conditional distributions for Y^n given X^n, so they only induce a conditional expectation of $L(Y,X)$, i.e. a function which maps each value for X to a corresponding expectation. However, this function is a constant: the expectation does not depend on X, and we may treat it like a single number, as in (4). In my thesis, I called (4) the *reliability* property of the mapping (2), since, for each $f \in \mathcal{F}$, the best-fitting $\hat{\beta}_f$ gives a "reliable" (unbiased in a very strong sense) indication of the performance of f on future data.

Entropification The reliability property indicates that it may be useful not only to learn f, but also to learn β from the data, and this suggests mapping $\mathcal{F}$ to the set of distributions $\mathcal{P} = \{P_{f,\beta} \mid f \in \mathcal{F}, \beta \geq 0\}$, where $P_{f,\beta}$ is given by (2). In my thesis and in [6], I called this mapping the *entropification* of $\mathcal{F}$, a name which has hardly caught on (but see [7]). As long as we restrict β to be nonnegative, the likelihood will be jointly maximized for a distribution $p_{\hat{f},\hat{\beta}_{\hat{f}}}$ such that $\hat{f}$ is optimal in terms of empirical risk, and $\hat{\beta}_{\hat{f}}$ is a reliable estimate of the performance of $\hat{f}$. There is a

slight problem in that for some loss functions, $\hat{\beta}$ may become negative, and then (3) indicates that the maximum likelihood will be achieved for some $p_{f,\beta}$ such that f has the *largest* rather than the smallest empirical risk within the set $\mathcal{F}$. The problem can be avoided by restricting the model $\mathcal{P}$ to $\beta \geq 0$. In fact, it is not clear whether this is a "problem" at all, because negative $\hat{\beta}$ has a clear interpretation. In the classification case, if $\hat{\beta}_f < 0$, this means that one obtains smaller loss by predicting a 0 whenever f predicts a 1 and vice versa. Thus, f makes a mistaken prediction on more than half of the sample points, which means that it performs worse than random guessing.

One may doubt the practical relevance of the optimality and reliability properties (3) and (4), since, if $\mathcal{F}$ is large, the ML estimator $p_{\hat{f},\hat{\beta}_{\hat{f}}}$ may of course be prone to terrible overfitting and one might prefer other estimators instead. However, (3) and (4) immediately imply "expected" versions of the two properties, and these make the notions relevant for *any* likelihood-based inference method, including Bayes and MDL. Namely, fix any joint distribution P^* on $\mathcal{X} \times \mathcal{Y}$. Let $\tilde{f}$ be the unique best predictor relative to P^*, i.e.

$$\tilde{f} := \arg\min_{f \in \mathcal{F}} E_{X,Y \sim P^*}[L(Y, f(X))],$$

and let $\tilde{P}$ be the conditional distribution of the form (2) that minimizes conditional KL-divergence to P^*, i.e.

$$\tilde{P} := \arg\min_{P \in \mathcal{P}} D(P^* \| P),$$

where $D(P^* \| P) := E_{X \sim P^*}[D(P^*_{Y|X} \| P)]$ is the conditional KL divergence [5]. For simplicity we assume here that $\tilde{f}$ and $\tilde{P}$ exist and are unique; otherwise, we make no assumptions about P^*; in particular, we do not assume that $P^* \in \mathcal{P}$. Then, as shown in my thesis, we have

$$\tilde{P} = P_{\tilde{f},\tilde{\beta}} \qquad \textbf{(optimality)},$$

for some $\tilde{\beta} \geq 0$, and, if $\tilde{\beta} > 0$, then

$$E_{\tilde{P}}[L(Y, \tilde{f}(X))] = E_{P^*}[L(Y, \tilde{f}(X))] \ \textbf{(reliability)} \tag{6}$$

For discussion about the case $\tilde{\beta} = 0$, see [6].

Now, as the sample size increases, *if a likelihood-based estimation method for $\mathcal{P}$ converges at all*, it will converge to the $\tilde{P}$ minimizing KL divergence to P^* [3]. Thus, the expectation-versions of the reliability and optimality-properties indicate that, if predictors are mapped to distributions using the "entropification" method (2), then, whenever estimation methods such as two-part or predictive MDL converge at all, they will (a) converge to the $\tilde{P}$ that leads to the best predictions in terms of the user-supplied loss function of interest; and (b) give a consistent (asymptotically unbiased) estimate of how good the predictions of $\tilde{P}$ really are. This seems exactly what is wanted from a mapping from deterministic predictors to description methods. I also found that earlier, [9] had explored a one-part code for the 0/1-loss based on essentially the same idea (averaging out β rather than encoding it). Finally, a seemingly disturbing aspect of Rissanen's approach (fixed β), when applied to the 0/1-loss, was its apparent difference from an earlier approach by Quinlan and

Rivest [10], who also tried to apply MDL in a deterministic classification context. I found that, if we allowed β to be determined by the data, then a simple application of Stirling's approximation showed that the Quinlan and Rivest approach in fact became *equivalent* to Rissanen's method after all. All this lead me to believe that "entropification" (i.e. extending the Gauss-Rissanen idea by allowing β to be learned from the data) was the right way to go.

Simple vs. Nonsimple Loss Functions Yet all was not well: as discussed in my Ph.D. thesis, entropification is surrounded by a myriad of slings and arrows. Here I concentrate on the most important one: the whole idea only works if the loss function is such that $Z(\beta)$ does not depend on f and x. In my thesis, I called such loss functions "simple". Both the 0/1- and the square loss functions are "simple", but most other loss functions of practical interest are not "simple". On our very first meeting, while praising the entropification idea in general, Jorma expressed doubts that it could be extended to such more general loss functions. For example, in a classification context, we may deal with an *asymmetric loss function*, which applies when predicting a 0 while the outcome should have been a 1 is much worse than vice versa. For example, $L_{\mathtt{as}}(0,0) = L_{\mathtt{as}}(1,1) = 0, L_{\mathtt{as}}(0,1) = 1, L_{\mathtt{as}}(1,0) = 10^6$. Loss functions such as this one are very important in practical applications such as deciding whether or not a certain drug should be administered to a patient. $L_{\mathtt{as}}$ is *not* simple, since $\sum_y e^{-\beta L(y,y')}$ depends on y'. In that case, the mapping (2) leads to values of $Z(\beta)$ depending on f and x, and the optimality property (3) is destroyed. We could try to save it by defining $Z'(\beta) = \sup_{y' \in \mathcal{Y}} \sum_{y \in \mathcal{Y}} e^{-\beta L(y,y')}$, and using (2) with $Z(\beta)$ replaced by $Z'(\beta)$. The distributions in $\mathcal{P}$ would then become defective (summing to less than one). Note that we will interpret our distributions as codes, and from such a coding perspective, there is nothing wrong with defective distributions in principle: via Kraft's inequality, they still correspond to codes. However, the approach is still flawed, since when using defective distributions in this way, the optimality property is restored but the reliability property is lost! How can one generalize the idea so that both the optimality and the reliability property continue to hold? Jorma was doubtful that this could be done, and that his own bold claim "prediction is coding" could still be maintained for such nonsimple loss functions (interestingly, Phil Dawid, during my thesis defense two months earlier, had raised exactly the same doubts).

Feeling challenged by Rissanen and Dawid, I spent many an afternoon in the Stanford or San Francisco sun, trying to figure out a way to make "entropification" work for a larger class of loss functions. I felt that somehow it was possible. Each time I drove over to Jorma, I discussed my newest ideas on the topic, and gradually, I convinced him that my approach was feasible. Usually he would just listen, make some encouraging but general remarks, and then one day later send me an email, invariably starting with "Peter:", followed either by a counterexample to my latest approach or some other profound issue.

> **The Importance of Being Brief** Just as in conversation, Jorma's emails are invariably brief and to the point. Here is an example dated February 1999: "Peter: The thing that's missing in your lemma is only to show the convergence $\hat{\theta} \to \theta^*$, which permits you to replace the almost sure convergence of the sum simply by the entropy. Incidentally, if you

have time next week we should meet. Jorma."

My responses were always long. Similarly, Jorma's MDL books are short, mine is very long. Jorma sticks to his own principle. I would love to do the same, but find that I lack the time: as Blaise Pascal has said "I have only made this letter longer because I have not had the time to make it shorter."

In the end, with Jorma's help, I found a partial and surprisingly simple solution to the nonsimple loss problem, which I'll now describe. For sake of generality, we drop the restriction that the set of possible predictions coincides with the set of outcomes $\mathcal{Y}$. Thus, let $\mathcal{A}$ be the set of possible predictions ($\mathcal{A}$ stands for "acts" or "actions", as decision theorists would prefer to call them), and let $\mathcal{Y}$ be the set of possible outcomes to be predicted. Then a predictor f is a function $f : \mathcal{X} \to \mathcal{A}$, and a loss function is a function $L : \mathcal{Y} \times \mathcal{A} \to \mathbb{R}$. With each loss function we can associate its range

$$\mathcal{L} := \{l \in \mathbb{R} : l = L(y, a) \text{ for some } \ y \in \mathcal{Y} \text{ and } a \in \mathcal{A}\}.$$

For simplicity, we restrict ourselves to cases where $\mathcal{L}$ is finite. For example, with $L = L_{\mathtt{as}}$, the asymmetric loss defined before, we have $\mathcal{L} = \{0, 1, 10^6\}$.

Coding the Loss rather than the Data The central idea to make entropification work again is *to code the losses rather than the y-values.* Thus, we associate each f and β with a code for describing, for each i, the size of the *loss* $l_i := L(y_i, f(x_i))$, obtained when predicting y_i based on $f(x_i)$. Thus, rather than coding y_i given x_i using a code not depending on f (as in the original entropification-approach), we now code l_i using a code which does depend on f.

To this end, for each $\beta \geq 0$, we define a mass function p_β on $\mathcal{L}$, simply by setting

$$p_\beta(l) := \frac{1}{Z(\beta)} e^{-\beta l} \tag{7}$$

where $Z(\beta)$ is given by

$$Z(\beta) \equiv \sum_{l \in \mathcal{L}} e^{-\beta l} \tag{8}$$

p_β is extended to sequences by independence, $p_\beta(l_1, \ldots, l_n) = \prod_{i=1}^n p_\beta(l_i)$.

For given data x^n, y^n, given $f \in \mathcal{F}$ and $\beta \geq 0$, we now code the corresponding losses $l_1, l_2, \ldots, l_n$ using the code with lengths $-\log p_\beta$. Thus, for each x^n, y^n, we get codelength

$$-\log p_\beta(l^n) = \beta \sum_{i=1}^n L(y_i, f(x_i)) + n \ln Z(\beta). \tag{9}$$

(note again that p_β does not depend on f, but l^n does). (9) corresponds to (3) and shows that our mapping satisfies *optimality.* How about reliability? Let's fix some f and compute the derivative of (9) with respect to β, i.e. $(d/d\beta)(-\log p_\beta(l^n))$. A straightforward calculation analogous to (5) shows that *if* the derivative is 0 for some $\beta > 0$, then the likelihood achieves a maximum (the minus log likelihood is minimized) for this β, and for this β, we have

$$E_\beta[L] = \frac{1}{n} \sum_{i=1}^n L(y_i, f(x_i)).$$

But this is just the reliability property (4) again. Further analysis reveals that an analogue of the expectation-versions of reliability and optimality also holds. Thus, by directly coding losses rather than outcomes we have recovered the essential properties of entropification for simple loss functions!

This approach does the trick, but it also raises a lot of questions: (a) what if the maximum likelihood is achieved at some $\beta < 0$? (b) can we still think of inference based on coding of the losses rather than the data as a form of the MDL principle? Philosophically, can there be some rationale for coding losses rather than data? Or should we somehow further adjust the approach such that, from what is encoded, the data $y_1, \ldots, y_n$ can always be recovered? More generally, the idea to code losses rather than data, which implies that the objects one actually wants to encode depend on the code one chooses to encode them, seems highly unusual! (c) How does the approach extend to continuous loss functions and loss functions with infinite domains? (d) does there exist a general formulation which subsumes both the "simple" approach (2) and the nonsimple approach described above? (e) How does the approach compare to the "aggregating algorithm" designed by Vovk and others [16, 8, 17, 1]? The aggregating algorithm can also be re-interpreted as mapping predictors/loss functions to probability distributions using (2), but rather than being learned from the data, β is chosen as a function of the loss function, and sometimes the sample size, in order to achieve good worst-case performance. Intriguingly, it turns out that Vovk's approach can only work for nonsimple loss functions – when Vovk deals with the squared error loss function, he restricts the range to $[-1, 1]$, and then $Z(\beta)$ becomes dependent on $f(x)$.

3 An Unfinished Tale

I found the solution sketched above in the final weeks of my stay in Stanford. A few months later I discovered how one can avoid the problems with $\beta < 0$, and I also discovered a more general approach subsuming the original entropification method, the method described above, and a third method which works for some continuous asymmetric loss functions. Still, the important question (b) remained open, and I thought that I should only publish this work after having given more thought to it. Yet, until 2003, there were always more urgent things to work on, and I could not find the time for these thoughts. Then, in 2003, John Langford and I discovered that even with the simple 0/1-loss, applying MDL or Bayesian inference to an "entropified" model class $\mathcal{P}$ can be inconsistent: there exist sets of 0/1-predictors $\mathcal{F}$ such that two-part MDL or Bayesian inference based on the associated $\mathcal{P}$ never converges [4, 5]. Note that the optimality and reliability properties indicate, as written below (6), that *if MDL estimation converges at all, it converges to the "right" $\tilde{P}$*. The problem that Langford and I discovered is that in some cases, MDL estimation does not converge at all! Even though the best classifier $\tilde{f} \in \mathcal{F}$ has a small description length, as the sample size increases, MDL keeps selecting ever more complex classifiers, all of which are of much worse quality than the simple $\tilde{f}$. This seemed to be such a setback for the whole entropification idea, that I further postponed sorting out the details. Thus, my paper "Prediction *is* Coding, " about entropification for general loss functions, has been left unfinished and has never been published. I do hope to finish it someday soon! (but I've been saying that for

years). Of course, we could have made it into a joint project with Jorma, but his interests shifted as well – soon he was all into the Kolmogorov minimum statistic. He did publish, in 2003, a paper on normalized maximum likelihood for 'simple' nonlogarithmic loss functions, which was inspired by our many conversations at Almaden [14].

The Impact of Entropification It should be clear, that, although, in the end, we have not jointly published about it, Jorma's thoughts about "the device already used by Gauss", while fitting on a single page, have had a tremendous influence on my career. It was the basis for a large part of my Ph.D. thesis, of my first COLT paper [6], of the Bayes/MDL-inconsistency papers [4, 2, 5] – the latter having caused quite a stir among some Bayesian statisticians. In 2004, I was awarded a prestigious VIDI-grant by NWO, the Dutch science foundation. This award has made it possible to start what is rapidly becoming my own research group. The grant proposal was, in fact, all about entropification. I should really like to thank Jorma for that single page in his book!

References

[1] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning and Games*. Cambridge University Press, Cambridge, UK, 2006.

[2] P. D. Grünwald. Bayesian inconsistency under misspecification, 2006. Presentation at the Valencia 8 ISBA Conference on Bayesian Statistics.

[3] P. D. Grünwald. Prediction is coding, 2007. Manuscript in preparation.

[4] P. D. Grünwald and J. Langford. Suboptimality of MDL and Bayes in classification under misspecification. In *Proceedings of the Seventeenth Conference on Learning Theory (COLT' 04)*, New York, 2004. Springer-Verlag.

[5] P. D. Grünwald and J. Langford. Suboptimal behavior of Bayes and MDL in classification under misspecification. *Machine Learning*, 2007. To appear.

[6] P.D. Grünwald. Viewing all models as 'probabilistic'. In *Proceedings of the Twelfth Annual Workshop on Computational Learning Theory (COLT '99)*, 1999.

[7] M.D. Lee. Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin and Review*, 15(1):1–15, 2008.

[8] N. Littlestone and M. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994.

[9] R. Meir and N. Merhav. On the stochastic complexity of learning realizable and unrealizable rules. *Machine Learning*, 19:241–261, 1995.

[10] J. Quinlan and R. Rivest. Inferring decision trees using the minimum description length principle. *Information and Computation*, 80:227–248, 1989.

[11] J. Rissanen. Stochastic complexity and modeling. *The Annals of Statistics*, 14:1080–1100, 1986.

[12] J. Rissanen. *Stochastic Complexity in Statistical Inquiry*. World Scientific Publishing Company, 1989.

[13] J. Rissanen. Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, 42(1):40–47, 1996.

[14] J. Rissanen. Complexity of simple nonlogarithmic loss functions. *IEEE Transactions on Information Theory*, 49(2):476–484, 2003.

[15] J. Rissanen and Bostwick F. Wyman. Duals of input/output maps. In E.G. Manes, editor, *Category Theory Applied to Computation and Control, Proceedings of the First International Symposium*, volume 25 of *Lecture Notes in Computer Science*, pages 204–208. Springer, 1975.

[16] V.G. Vovk. Aggregating strategies. In *Proceedings of the Third Annual Workshop on Computational Learning Theory (COLT' 90)*, pages 371–383, 1990.

[17] K. Yamanishi. A decision-theoretic extension of stochastic complexity and its applications to learning. *IEEE Transactions on Information Theory*, 44(4):1424–1439, 1998.

A Great Mind

Paul Vitányi
CWI and University of Amsterdam

April 7, 2008

1 First Contact

I may have met Jorma Rissanen in the flesh for the first time in Japan on Mount Fuji, in the Fuji Center for Training and Education, a marvelous facility for small conferences on the slope of Fuji-San, where an IEEE Workshop was organized in 1993. Jorma turned out to be a dapper, ramrod straight, wiry and muscular, little man with sand-colored hair that stood straight up on his head. Talking with others, he stood straighter than ever and defiantly looked up, and either got bored or said "gosh, that may well be true; I may have to look into that." His size is a source of worry to him, not in the last place because it prevented his dream come true. As he told me many times over the years, in his youth he was an intrepid soccer player. But he became a scientist instead. Spending most of his career at IBM Almaden Research Center, he has been the main stay of the IBM soccer team, practicing almost daily, until his retirement at the age of seventy.

Lucky for Jorma, he has been supported and guided by his lovely wife Riitta. "Riitta is infinitely wise. Sometimes I go against her counsel, but she is always right and I live to be sorry." Jorma gave evidence: "Let me tell you a story. There was a meeting in Khwarizm, south of the Aral Sea in central Asia, where Al-Khwarizmi came from. Riitta told me that I was stupid to go there, and why should I want to go there anyway? Nonetheless I went. The first evening I had something to eat that didn't agree with me. The remainder of the meeting I spent in the bathroom, and couldn't talk with anyone. Riitta was right again." Jorma told about Linkoping University "they offered me a professorship at the University. Riitta said 'don't go, what do you want there?' I went nonetheless, no doubt driven by ambition. But I felt miserable there. After a year I quit. Riitta was right." If I quote Jorma, it is from memory. Nothing I can write will do justice to his inimitable style, lucid, brief, to the point, and disconcertingly honest, both in writing and in the spoken word. Thinking about Jorma I have many fond memories; to group them may be easiest by tying them to various beverages. My present circumstances where the consumption of alcohol is a hazard, excuses this greediness, much like Evelyn Waugh states in the Preface to 'Brideshead Revisited:'

> "It was a bleak period of present privation and threatening disaster [...] and in consequence the book is infused with a kind of gluttony, for food and wine, for the splendors of the recent past, and for rhetorical and ornamental language which now, with a full stomach, I find distasteful."

2 Sake

At the meeting in the Fuji Center for Training and Education, the limited hotel facilities in the Center itself were not sufficient to house all participants. So many were staying at the excellent Sun Green Fuji Hotel in nearby Hakone, Japan and Tokio's rural countryside for hiking and going to the 'onsen', the open air thermal baths. Among those so favored were Jorma and me. This meant that one had to get from the Center to the Hotel and back. After a memorable evening in the Center where with the rolling thunder of a Taiko drum group, led by a supreme female Taiko drummer which is unusual, we were transported back to the Sun Green Fuji by a small bus. Jorma, however, was invited by a group of distinguished Japanese scientist in their limousine. Being in conversation, Jorma invited me too. I, however, suggested it was more cozy to take the minibus. "I cannot do that," Jorma said, "because it does them honor to drive me home in the limousine."

Back in the Sun Green Fuji we looked for a bar to discuss things, but there was no bar. So back in the hotel room with Joe Suzuki and some other Japanese whose names have escaped me, we called room service and asked for sake. It turned out that sake was only served together with a meal. "How much sake with 150 dollars of sushi?" Turned out to be a lot of sake and exciting conversation till deep in the night.

In Melbourne, Australia, at a meeting organized by the indefeatigable David Dowe, and possibly David Wallace and Kevin Korb, called by the improbable name of ISIS'96, we met again. Jorma held his own against characters like Marvin Minsky and Ray Solomonoff. At one of the first nights I went to dinner with Jorma and a small Finnish contingent led by Henry Tirri. We ended up in a Japanese restaurant. Conversation was amiable and heated, but being jet-lagged I cannot remember details. Anyway, the sushi was sprinkled with sake; and more sake as the evening progressed. Next day I had cause to regret this. Around midnight the place wanted to close; we were the last guests. On the way out the entire Japanese staff and waitresses formed a queue to the exit, and bowed us reverently farewell. "So much sake" murmered the waitresses softly.

In the DIMACS Workshop on Complexity and Inference in 2003 we had occasion to visit a Japanese restaurant again; sushi and sake galore. Ray Solomonoff and his wife Grace, who are very cunning in these matters, had figured out that if one became a member of the Honors Club of the Hilton Hotel were we all stayed, one had the right to a beverage of his choice in the bar of the hotel. The bar was usually deserted when we came in, and the bartender got very happy by the unusual choices we made, Daquiri, Tropical Sunrise, and so on. He consulted his cocktail manual, and provided ever better concoctions, to Jorma's delight.

3 Beer

In Kopenhagen at a tutorial meeting organized by Peter Johansen at the Datalogisk Institut of Kopenhagen's Universitet (if I spell it right) Jorma was accompanied by Riitta. As was often the case, Jorma looked for a friendly trusted face. We three spend most evenings together, and, as Jorma told me "Riitta likes you. That is all I need." I liked Riitta too. She often has a marvelously malicious sense of humor which is rare to come by. We drank several beers together, in or close by Tivoli. In Jorma's tutorial, in an ancient lecture hall, he held forth over the philosophy and the sublime qualities of the minimum description length principle (invented by him) for doing statistical inference. One of the members of the audience objected "but the statisticians say ..." This was grist on Jorma's mill. He shouted triumphantly, with brightened eyes, one of his favorite homelies, thus silencing the opposition. I recall that the outing and dinner

of this meeting were to a remarkably pretty old-fashioned wooden house-restaurant probably near a lake. There, in the old-fashioned very light dining room we had an excellent dinner with Riitta, Jorma, and me at a table for three.

In Barcelona in 1995 I invited Jorma to give a keynote lecture. He graciously accepted and brought Riitta. We drank beer and wine and at the conference dinner, in some dark setting if I recall correctly, Riitta said to me with a dangerous gleam in her eye "I don't like your friend, ..."

In November 2002 Ursula Gather organized a small workshop in Statistics at the University of Dortmund. "Apparently the statistical community is finally starting to appreciate these ideas," Jorma wrote. The meeting took place in a small hotel with meeting rooms so that the participants could mingle and meet in the bar and at dinner. Jorma had his heyday. Most of the participants were German scientists of leftish persuasion. Gleefully he said "Riitta and I voted for George W. Bush." After an appreciable silence "And before that for Ronald Reagan." After a rather deeper silence he explained "The terrorists of nine-eleven have to be treated harshly. The French and Germans were really cowards not to support the attack on Afghanistan. Especially the French are bad. Bush renamed 'french' fries to 'freedom' fries, and 'french' toast to 'freedom' toast, to show the French what was what. Riitta and I only talk about 'freedom' fries and 'freedom' toast. There was an embarrassed silence among the German scientists, which endured while they were struggling with the conflicting emotions of admiration for the great scientist Jorma Rissanen and the surprising political insights he had just offered.

4 Wine

Dagstuhl Castle, the German facility of the state Saarland to foster Computer Science by facilitating the organization of small live-in week-long seminars, has, apart from excellent served meals also excellent drinks. Situated in the forests near the Saarland village of Wadern, the castle invites nature walks and great conversations. There is also a fitness room and exceptionally large sauna facilities. Experience has it that although some groups don't use the latter at all, some groups use it intensively every day. The beer is of three types, with the Bitburger "bitte ein bit" one of the favorites. But here I want to talk about the wines. By far the best, and capable of competition with the finest restaurants, is the 'Chante Alouette' Saint Emillion Grand Cru. I was introduced to it by Steve Smale, and have drunk no other wine at Dagstuhl since. In the spring of 2003, at the Centennial Kolmogorov Seminar, in honor of the 100th birthday of that great Russian mathematician, I met Jorma again. Also present was a contingent of Russian mathematicians and computer scientists, primarily from Moscow, and constituting a group that is known as the 'Kolmogorov school', even though the namegiver has long passed away. One other information theorist of Russian extraction was Boris Ryabko, a good friend of Jorma from Novosibirsk. In the evenings there was sauna, enthusiastically taken by the Fins like Jorma, Henry Tirri and Petri Myllymaki, myself, and the odd Russian like my co-author Kolya Vereshchagin and Boris. Jorma watched approvingly the antics of Kolya, who, after a sauna session, plunged with a great splash in the man's height tall wooden tub of ice-cold water. Taking a sip of his beer and reclining on the relaxing after-sauna beds, Jorma asked me "what was this all about, the talk this fellow, making the big splash, gave this afternoon?" "Well," I said "actually it was about that paper of him and mine which you told me you refereed." "Gosh, I would never have guessed." "But", Jorma said "it gets increasingly hard to understand other peoples work, especially if they are young and eager." He added "This was also remarked by Stanislav Ulam in his autobiography 'Adventures of a Mathematician' where he said 'I feel like an old boxer; I can still dish it out but I can't take it anymore.' "

Suddenly, the lights went out and came flickeringly on again. Slightly later a fireman in a lot of clothes, contrasting with our nudity, came in telling that a sudden freak-tornado had blown away the roofs of the annex and the library, and for security reasons they were shutting down the sauna. It appeared, that Kolmogorov on the centenary of his birthday had called forth the winds. A few participants—not us—had to leave to a nearby hotel since their bedrooms were now open-air. Later in the dining room, enjoying a glass of Chante Alouette, Jorma turned to Kolya and asked "What was your name again? I understand you are Russian and belong to that group over there" pointing at the Kolmogorov school sitting at another table. He explained, with a straightforwardness most of us, alas, tend to loose over the years "I have noticed that this group, maybe from Moscow, doesn't want to interact with the other Russian, my friend Boris. Why is this so? Do they feel themselves too good, being from Moscow? But I can tell you that none of them are anything compared to Boris or his work. They are not worth to tie his shoe-laces. Can you explain to me why this group behaves so?" Kolya's explanation didn't satisfy Jorma, but he had made his point and returned to the glass of Chante Alouette.

5 Wodka

The Finnish group in Helsinki regularly invited me to give a week-long seminar or short course for credit for the students of Helsinki University and Helsinki Institute of Technology. Indeed, the university conferred to me the inscribed Medal of Helsinki University for services delivered. I usually made it a condition to have the timing coincide with Jorma's short lecture courses at the same university, so that I would have congenial company. Indeed, they took care that both of us resided in the magnificent Scandic hotel in the center of Helsinki. It turned out that the Scandic was owned by the Hilton group, and since both Jorma and me were members of the Honors Club, we both merited a number of drinks per diem. However, this being Finland, no cocktails but bottles of beer. In fact, so many bottles that we couldn't drink them. Unofficially, our hosts were the COSCO group, led by Henry Tirri, now Nokia Research Fellow at Nokia. This group resides in an uncommonly beautiful location, a new office building made almost completely of glass in a restored section of harbor buildings. In fact, Henry's office had completely glass walls, being 3 or 4 meters high, and one could use them as blackboard using a marker pen. There I explained some work on Kolmogorov's structure function to Jorma and Henry, writing on the glass outside wall. This was a possible new approach to the foundation of MDL used by Jorma in his new book published by Springer in 2007.

Generally, after hours, we were taken to a nearby upscale pub in this see-sun-clouds restored harbor quarter. Henry tossed a credit card to the barman at the end to pay for our desires. An important part of the visit was dinner at a Russian restaurant. In Helsinki this is the epitome of luxury, and the pinnacle of the Russian restaurants is the Shashlik restaurant. There, we had many a memorable evening. The Shashlik has at least twenty types of wodka of which I remember the lemon wodka, the pepper wodka, and the cranberry wodka. The menu had several non-correct items of which I vividly remember the bear-meat. One started with a platter of titbits on which the bear sausage stood out. The conversation, of which I remember little because of the wodka, was mainly about university politics in Finland, and how to deal with them. Also how to organize payment of the incredibly expensive meal.

6 Margaritha

After visiting and working with Ming Li in Santa Barbara I decided to drive up to San Francisco and visit Jorma on the way. But first I was going to Big Sur and stay there in the pretty Riverside Campground an Cabins in one of their red-painted cabins among the redwood trees, bordering the softly murmuring Big Sur river. This is in the middle of the purest stretch of California coastline, along the picturesque meandering Route 1. There are only a few isolated restaurants and motels in the wild nature; every city or village is hundreds of miles away. This makes for small numbers of people; if you go to Pfeiffer's Beach, featuring in many a movie, you will not find much company except the birds and the sea dwellers. There is Arthur Miller's house, now a small museum among the rocks and the redwoods. One restaurant is the famous Nepenthe, perched on the side of a rock above the Pacific Ocean. Customers waiting for dinner to be ready sit outside near a giant open-air fire, or perch along the railing overlooking the falling rock and the Pacific. Opossums sneak by and try to make off with the odd food-scrap. Perching along the railing, sipping a Gold Margaritha, and reminiscencing with other customers, makes life feel as good as it gets. So I called Jorma and told him I was delayed. Next day I called again, being delayed once more. Jorma tried to entice me to leave Big Sur and come to San Jose referring to Riitta and making everything in his home in San Jose sound extra good. To no avail; and in the end my time was up, I was still in Big Sur, and had to make haste to catch my plane in San Francisco. So I called Jorma and told I had to take a rain check. Rain is water. Little did I know.

7 Water

In Spring 2006 I was working with Ming Li at the University of Waterloo, Waterloo, Ontario, Canada. One evening I didn't feel well. Thinking it was a mild food-poisoning I went to bed early, only to wake up in the middle of the night being paralyzed. I was, with difficulty, able to reach the phone and call Ming telling him I needed an ambulance. I was admitted in the small and nice local 'Grand River Hospital' in between the Amish people. "Sir, you are having a stroke" the doctor on call in the admission ward told me. Later many people send me cards, flowers, and best wishes, to go with the thickened water I was allowed to drink. No get-well gift was bigger that the giant fruit-and-food basked I got from the faraway COSCO group and Jorma. No email was more complimentary and fortifying than Jorma's.

My encounters with MDL

Terry Speed

University of California, Berkeley
Department of Statistics

August 26, 2007

Abstract

In this short contribution, I present a personal view of MDL.

I first met Jorma Rissanen in the mid-1980s in Canberra, Australia. At that time he gave a talk about his MDL, and I was instantly captivated by it. I talked to him immediately afterwards to learn more about it, and have talked to him about it many times since then. Though I was aware of Kolmogorov complexity, it was through Jorma I learned of its broader significance, and of the closely related work of Chaitin and Solomonoff. I will be forever grateful to him for introducing me to the beautiful ideas and results that he and these other individuals explored, connecting mathematics, logic, philosophy, information, probability and statistical inference. It has greatly enriched my appreciation of what I do, and of the broader mathematical world around me. I would find it hard not to be fascinated by ideas which connect Gödel's incompleteness theorem, data compression, gambling, and statistical inference. Jorma's world does this, and much more. It has been a pleasure to learn from him, a privilege to know him, and an honor to be invited to contribute to this volume. I'm just sorry that I couldn't offer some original research in his field.

I'll now try to explain how his research relates to my own day-to-day work, and why I don't use MDL all that much. As is well known, the MDL principle is something we use to select models from classes of models. I'm sure we'd all agree that while many of the things we describe as statistical are doing just that, there are plenty of statistical activities not of this kind. That is certainly the case for me, as I'll explain shortly. But even when I could use MDL in my daily work, I have not done so more than once, because it did not seem right to do so in my context. As I write this, I am aware that almost everything I say can also be applied to the Bayes/non-Bayes distinction. A committed Bayesian is going to use Bayesian methods in contexts where I don't, and feel that he has good reasons for doing so. This suggests that to a larger extent than we might be happy to admit, the particular tools and techniques we use in any given statistical analysis are only partly chosen by rational processes. We'd probably like to think we seek out the most appropriate thing to do in a particular situation, reach a conclusion by objective means, and then follow the course we've decided is best. Perhaps we do, and the differences between the way different statisticians approach the same problem are all explainable by their different levels of knowledge and experience, but I suspect it's not that simple.

For most of the time since I met Jorma, I've been engaged in the statistics of genetics and molecular biology. The raw data can be genotypes at hundreds of markers on scores of

mice, tens of thousands of summarized fluorescence intensities from each of tens to hundreds of microarrays, millions of ion intensities from the interface of a gas chromatograph with a mass spectrometer, or billions of nucleotides comprising the DNA sequence of a genome. Over this period, the amount of data coming my way has grown dramatically, and new types of data keep arriving. My co-workers and I spend much of our time doing what I think of as "taming" the data, that is, getting it into a form in which we can visualize, explore, summarize, and no longer feel intimidated by it. This is necessary to allow us to start appreciating the features of the data that are relevant to the questions people want to address with it. A lot of time has to be spent doing our best to understand the technology which gave rise to the data, and the scientific context in which the data are thought to be informative. When someone gives us a new set of data and asks is it good enough to address the questions of interest to them, it will often take weeks or months of work to provide them an answer to that simple first question, and longer still to help answer their substantive questions, when that is possible. The point here is the following: it's hard to define a model class if you don't understand what the data mean, or if you don't know how to tell signal from noise in that data. But I hope this "pre-model" analysis is seen to be statistics.

Now let's suppose that we've "tamed" the data, that we have a half-reasonable grasp of its context, and are ready to begin addressing the questions of interest; aren't we going to define some models now, and won't model selection follow soon afterwards? Well, yes and no. For me, a common first question is this: which entities (genotypes, gene expression levels, ion intensities, DNA sequences) are different between two specified classes of samples? In order of complexity (no joke intended), we might calculate a mean difference (typically on a log scale), a t-like statistic, an empirical Bayes t-like statistic, a p-value, or a false discovery rate, one each for perhaps tens of thousands of our entities. We do this marginally, though at times permutation testing or bootstrapping or other modeling might take us beyond marginal summaries. Although Bayesians will routinely describe joint distributions for everything under discussion, I have so far not attempted that in the contexts I'm discussing. To date I have felt that the joint distributions for the entities I am analyzing are beyond me. This means I am reluctant to think of a description of it all, preferring to pick away at the margins, in a simple and what I hope is a robust way. So here are two features which make it hard for me to envisage using MDL in my context: most of the time there is no joint distribution for my data, and I seek robustness when I do approximate parametric inference on marginals. Fitting robust methods into a nice model-based framework seems to me to be hard, as I am usually unable to employ mixture models. My preference is for M-estimators, typically ones which do not correspond to MLEs for any distribution.

Of course none of what I have said would stop an enthusiastic MDL-er from using MDL, any more than it stops committed Bayesians from using Bayes theorem, but there would clearly be some serious work needed were I to try. Well, you might argue, if the benefits are likely to be great enough, do that work. I'd reply that it's already a lot of work to get where we've got, and my collaborators are waiting for answers. I would need to be devoid of useful ideas without MDL, and/or be very sure that the benefits would be worth the wait. On one occasion a few years ago, this was the case. I did have a clear model selection problem, and stochastic complexity based on the NML density function gave a very satisfactory solution. I was glad I knew about it, as the alternatives seemed much more complicated. This attitude puts the MDL alongside a host of other techniques that statisticians draw upon when it seems right, and does not give it any central or special status. For me, this is the case, and here's the main reason why. I like to keep the methods I develop for the kinds of data I'm describing simple and reasonably transparent. If at all possible, I try to base my methods on ideas which should or could be accessible to numerate biologists. I do so as I recognize that most such data will not

be analyzed by statisticians, but by the people who collect the data, with the aid of commercial packages. There will be no-one to help them, and in my view it's dangerous to present them with black boxes, and invite them to "trust us". Most will have some intuition about basic statistical notions, and be able to navigate themselves based on their understanding of these notions. Rightly or wrongly, I want what the methods I recommend to be comprehensible to most if not all of these non-statistician potential users, and this discourages me from departing from simple, familiar approaches. In due course notions of coding, the MDL principle, and a host of techniques derived from it may well become common knowledge, and be viewed as simple by the thousands of non-statisticians who analyse their data along the lines laid out by professionals such as us. But we are not there yet.

Summarizing, I have a great affection for the circle of ideas in which MDL is embedded, and I'm glad I have learned what I have about it. I will draw upon that knowledge from time to time. But most of my statistical effort is at the "pre-model" stage, and most of the data I meet is not helpfully modeled in its totality. Hence, I don't use MDL every day in my professional life.

Terry Speed

Jorma's unintentional contributions to the source coding research in Eindhoven.

Tjalling Tjalkens and Frans Willems
Eindhoven University of Technology
P.O.Box 513
5600 MB Eindhoven
The Netherlands

April 8, 2008

1 Introduction

Around 1980, the Information Theory group in Eindhoven was mainly interested in multi-user information theory. The head of the group, Piet Schalkwijk, had previously published papers on the application of enumerative techniques for source coding but the only compression activity at that time was image coding related. In 1982 Frans Willems joined the group as an assistant professor and Tjalling Tjalkens started his Masters thesis work. Tjalling was given a preprint of Jorma's 1983 paper, [1], on the algorithm "context". Frans was given the task to assist the students in their homework assignments for the Information Theory course. This course still contained a significant part of enumerative source coding and Frans worked on an assignment related to the Tunstall algorithm, see [2]. During his work on the Master thesis, Tjalling became interested in the Lempel-Ziv algorithm, [3, 4], and its interpretation as a variable-to-fixed length code. Together Frans and Tjalling started looking into combinatorial approaches to variable-to-fixed length codes.

An observation of R. Petry, see [5], started the research that led to an enumerative implementation of the Tunstall algorithm and finally resulted in multiplication-free arithmetic codes. The results will be summarized in Section 2. We learned about universal coding, partly from Schalkwijk's enumerative scheme, [6], and partly from Jorma's and Glen Langdon paper of 1981, [7]. We wondered if it was possible to come up with an enumerative variable-to-fixed length code that is also universal (at least over the class of memoryless sources). We were not the first researchers to find such a scheme. Lawrence, [8], had adapted Schalkwijks fixed-to-variable length scheme. Precisely how Jorma contributed to our result will become clear in Section 3 where we explain the scheme and its analysis. Jorma introduced us to the model class FSMX, [9], and that enabled Yuri Shtarkov and ourselves to develop the Context-Tree Weighting method, [10]. In the last section we shall present some old and new results related to this algorithm.

2 Enumerative coding

2.1 The Tunstall code

A variable-to-fixed length source code, or V-F code, maps source sequences of variable length onto fixed length code words. The set $\mathcal{M}$ of source sequences, or messages, that are represented by code words is required to be *proper* and *complete*. A message set, and the corresponding variable-to-fixed length

source code, is *optimal* if there exist no other message set of the same size that has a lower rate, or equivalently, that has a larger expected message length. An algorithm by Tunstall produces a sequence of optimal V-F codes of increasing size for a given source.

Tunstall algorithm:

Init. The initial message set $\mathcal{M}_0$ contains all single letters from the source alphabet. Let $i = 0$.

Extend. Let $\boldsymbol{u} \in \mathcal{M}_i$ be a message of maximal probability in the current message set. Remove $\boldsymbol{u}$ from the set and insert all sequences created by extending $\boldsymbol{u}$ by a single letter from the source alphabet $\mathcal{U}$. Thus we created the next set $\mathcal{M}_{i+1}$.

Repeat. Increment i and repeat from step "Extend" until the message set has the required size.

2.2 An enumerative scheme

With a Fibonacci type of recursion one is able to define variable-to-fixed length codes, see [5, 11, 12]. Given a binary memoryless source and two well chosen integers a and b the recursion was

$$c(n) = c(n-a) + c(n-b). \tag{1}$$

With this recursion we can apply Cover's enumeration [13] to compute the *lexicographical index* of a sequence in a set. Consider any binary and memoryless source that produces zeros and ones with probability $1-p$ and p respectively. We can approximate the optimal (Tunstall) code by selecting a and b such that $\lambda^{-b} \approx p$, where λ is the (largest real) solution of

$$f(x) = x^{-a} + x^{-b} = 1. \tag{2}$$

2.3 Simplifying the scheme

The storage complexity of this scheme is determined by the table of $c(i)$ values for $i = 1, 2, \ldots, n$. The magnitude of $c(n)$ is approximately λ^n. We wish to simplify this enumerative scheme. Instead of the precise values $c(i)$ from the recursion (1) we can work with values $\tilde{c}(n)$ for $n = 1, 2, \ldots$ that satisfy

$$\tilde{c}(n) \geq \tilde{c}(n-a) + \tilde{c}(n-b). \tag{3}$$

For $n \leq 0$ we set $\tilde{c}(n) = 1$. We will be able to select values $\tilde{c}(n)$ that can be described in fewer bits than $c(n)$. We also wish to limit the number of elements $\tilde{c}(n)$ that we must store in memory. Because $c(i) \approx \lambda^i$ we can find a constant k (approximately equal to λ^m) such that we can extend the numbers $\tilde{c}(n)$ for $n > m$ as follows. Let i and α be the unique integers such that $n = i + \alpha m$ where $i \in \{1, 2, \ldots, m\}$ and $\alpha \geq 1$, then

$$\tilde{c}(i + \alpha m) = k^{\alpha} \tilde{c}(i). \tag{4}$$

One can check, see [12], that with a proper choice of k expression (3) also holds for $n > m$. Using these numbers we obtain a *decodable* code, although with a somewhat higher redundancy as compared to the original code. Now we have a scheme with a small table of exponential numbers extended indefinitely using a cyclic access and a multiplicative scaling operating.

2.4 Jorma's contribution: (multiplication-free) arithmetic codes

So, where does Jorma contribute to this research. Obviously the "exponential" table is a multiplication table and obviously the approximated enumerative scheme performs similar operations as an arithmetic code. In fact, in [14] Jorma introduced arithmetic codes using an approximated exponential table.

3 A universal variable-to-fixed length code

3.1 A universal setting

From Tunstall's algorithm and common sense it is clear that an optimal variable-to-fixed length code attempts to create almost equally likely messages. In a universal setting, we assume that the source is binary and memoryless, but the actual source probability p is unknown to both encoder and decoder. So it is impossible to determine the message probabilities and thus impossible to design a Tunstall code.

3.2 Another variable-to-fixed length scheme

We assume that p, the source probability of producing a '1', is a random variable, uniformly distributed over the interval [0, 1]. So we end up with a composite source, see [15]. We define the 'composite' probability of a sequence as

$$Q^*(\boldsymbol{u}) = \int_0^1 (1-\theta)^{n_0(\boldsymbol{u})}\theta^{n_1(\boldsymbol{u})}\,d\theta = \frac{1}{n_0(\boldsymbol{u})+n_1(\boldsymbol{u})+1}\binom{n_0(\boldsymbol{u})+n_1(\boldsymbol{u})}{n_1(\boldsymbol{u})}^{-1}. \tag{5}$$

Here $n_u(\boldsymbol{u})$ denotes the number of times the letter u occurs in $\boldsymbol{u}$. In [16] we described an enumerative coding scheme based on a message set $\mathcal{M}$ that was defined using formula (5). A sequence $\boldsymbol{u}$ is in the message set $\mathcal{M}$ if its "probability" $Q^*(\boldsymbol{u})$ drops on or below a given threshold.

3.3 Jorma's contribution

Partly motivated by Jorma's 1984 paper on universal compression [17], and the realization that Schalkwijk's Pascal-triangle' algorithm was universal, we worked out a universal variable-to-fixed length algorithm. Lawrence, a student of Schalkwijk, had already used combinatorial methods to find a universal code. However, our scheme was both better, in terms of compression, and seemed to us to be less ad-hoc. We submitted the paper to the IEEE IT Transactions and received an interesting review. The reviewer stated that the result was interesting but the bounds were not tight enough. The upper bound should be $\frac{1}{2}\log\bar{n}$, where $\bar{n}$ denotes the expected message length, in accordance with the well-known lower bound. Although it was admitted that there was no lower bound for variable-to-fixed length codes yet. We always thought that Jorma was the reviewer. So, as "Jorma" wrote that this task of improving the bounds 'separated the men from the boys', and we definitely wanted to be considered men, we improved the upper bound and came up with a Rissanen like lower bound.

4 Context algorithms

4.1 FSMX models

In [9] Jorma introduced the concept of a FSMX source. As said in the introduction we were also well aware of the algorithm 'Context' from [1] that actually already dealt with FSMX sources and used the idea of "context selection". Thus, when Yuri Shtarkov visited us in 1992, we were ready to discuss alternatives to the context selection mechanism such as weighting models. This resulted in the Context-Tree-Weighting (CTW) method that we presented in [10]. Jorma also introduced us to the idea of the "minimum description length" (MDL) principle, see [17, 18] and we immediately saw that the weighting procedure performed an automatic MDL model selection, although it never explicitly selects a model. In the next sections we will briefly review the CTW algorithm and show how a modified version thereof will perform a MAP model selection.

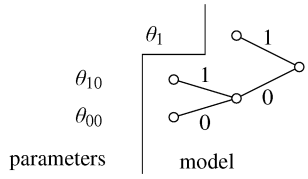


Figure 1: Model (suffix set) and parameters.

4.2 The CTW algorithm

Consider figure 1. For a *tree source* the probability $P_a(U_t = 1|\cdots, u_{t-2}, u_{t-1})$ is determined by starting in the root λ of the tree and moving along the path $u_{t-1}, u_{t-2}, \cdots$ until a leaf of the tree is reached. In this leaf s we find the desired probability (parameter) θ_s. The suffix set or tree $\mathcal{S}$, containing the paths to all leaves, is called the *model* of the source. The source model (tree) $\mathcal{S}$ partitions the source sequence in (conditionally) i.i.d. sub-sequences, one for each leaf $s \in \mathcal{S}$. We can use the KT-estimator, see [19], for each of these sub-sequences.

Suppose that the actual source model $\mathcal{S}$ is unknown, but that its depth is not larger than D. The context-tree weighting (CTW) method efficiently weights the sequence probabilities of all possible tree models. Define the cost of model $\mathcal{S}$ as

$$\Gamma_D(\mathcal{S}) \stackrel{\Delta}{=} 2|\mathcal{S}| - 1 - |\{s \in \mathcal{S}, \text{depth}(s) = D\}|, \tag{6}$$

and let $\mathcal{S}_a$ be the 'actual' model. The individual redundancy $\rho(u_1^T)$ relative to the actual source for sequence u_1^T can be upper bounded by

$$\rho(u_1^T) = L_{\text{CTW}}(u_1^T) - \log_2 \frac{1}{P_a(u_1^T)} < \Gamma_D(\mathcal{S}_a) + \frac{|\mathcal{S}_a|}{2} \log_2 \frac{T}{|\mathcal{S}_a|} + |\mathcal{S}_a| + 2, \tag{7}$$

for $T \geq |\mathcal{S}_a|$. Observe that bound (7) also holds for the redundancy relative to *any other source tree model* $\mathcal{S}$ with depth $\leq D$, i.e. $L_{\text{CTW}}(u_1^T)$ can be seen as a MDL solution to the corresponding modeling problem.

4.3 MAP selection

Suppose we wish to encode the sequence u_1^T in two stages. First we determine the model $\hat{\mathcal{S}}$ that minimizes the total codeword length. We then encode this model $\hat{\mathcal{S}}$ and then the sequence u_1^T given this model. In [20] the Context-Tree Maximizing method was discussed. It was also, and independently, proposed by Nohre in his Ph.D. dissertation [21]. This method recursively computes "maximum probabilities" over nodes given the KT sequence probabilities and results in the maximum a-posteriori probability (MAP) tree model. To reduce the complexity of the CTW algorithm we can store in a node instead of the estimated and weighted block probability a *probability ratio*. This idea was proposed in [22]. It seems odd that we have an efficient method to compute the weighted sequence probability, using the probability ratios, but still need to resort to the direct computation of the KT sequence probability in order to find the MAP model, given the sequence u_1^T. In [23] we presented a maximizing algorithm based on the same probability ratios.

4.4 Jorma's contribution

Clearly Jorma's papers influenced our research in context trees. Through his papers we became aware of FSMX models. The algorithm "Context" and the fact that there were no non-asymptotic performance bounds for this algorithm led us to revisit the idea of weighting. The notion of model redundancy, which we called parameter redundancy, was obviously based on Jorma's results. Finally the close relation between the redundancy bound and the MDL principle turned our attention to the modeling aspects and led indirectly to the efficient MAP model selection algorithm.

References

[1] J. Rissanen, "A Universal Data Compression System," *IEEE Trans. Inform. Theory,* vol. IT-29, pp. 656–664, Sept. 1983.

[2] B.P. Tunstall, *Synthesis of noiseless compression codes,* Ph.D. Dissertation, Georgia Inst. Tech., Atlanta, GA, Sept. 1967.

[3] J. Ziv and A. Lempel, "A universal algorithm for sequential data compression," *IEEE Trans. Inform. Theory*, vol IT-23, no 3, pp. 337–343, 1977.

[4] J. Ziv and A. Lempel, "Compression of individual sequences via variable rate coding," *IEEE Trans. Inform. Theory*, vol IT-24, no 5, pp. 520–536, 1978.

[5] J.P.M. Schalkwijk, "On Petry's extension of a source coding algorithm," in *Proc.* 2^{nd} *Symp. Inform. Theory in the Benelux,* pp. 99–102, 1981.

[6] J.P.M. Schalkwijk, "An algorithm for source coding", *IEEE Trans. Inform. Theory,* vol. IT-18, pp. 395–399, May 1972.

[7] J. Rissanen and G.G. Langdon, Jr., "Universal Modeling and Coding," *IEEE Trans. Inform. Theory,* vol. IT-27, pp. 12-23, Jan. 1981.

[8] J.C. Lawrence, "A new universal coding scheme for the binary memoryless source," *IEEE Trans. Inform. Theory,* vol. IT-23, no. 4, pp. 466–472, 1977.

[9] J. Rissanen, "Complexity of strings in the class of Markov sources," *IEEE Trans. Inform. Theory,* vol. IT-32, pp. 526–532, Jul 1986.

[10] F.M.J. Willems, Y.M. Shtarkov and Tj.J. Tjalkens, "The Context-Tree Weighting Method: Basic Properties," *IEEE Trans. Inform. Theory,* vol. IT-41, pp. 653–664, May 1995.

[11] Tj.J. Tjalkens and F.M.J. Willems, "Variable-to-fixed length codes for Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-33, pp. 246–257, 1987.

[12] Tj.J. Tjalkens, *Efficient and fast data compression codes for discrete sources with memory,* Ph.D. Dissertation, Eindhoven Univ. Tech, The Netherlands, 1987.

[13] T. Cover, "Enumerative source coding," *IEEE Trans. Inform. Theory,* vol. IT-19, pp. 73–76, Jan. 1973.

[14] J.J. Rissanen, "Generalized kraft inenquality and arithmetic coding," *IBM J. Res. Develop.,* vol 20, pp. 198–203, May 1976.

[15] L.D. Davisson, "Universal noiseless coding," *IEEE Trans. Inform. Theory,* vol IT-19, pp. 783–795, Nov 1973.

[16] Tj.J. Tjalkens and F.M.J. Willems, "A universal variable-to-fixed length source code based on Lawrence's algorithm," *IEEE Trans. Inform. Theory,* vol. IT-38, pp. 247–253, Mar 1992.

[17] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Trans. Inform. Theory,* vol. IT-30, pp. 629–636, Jul 1984.

[18] J. Rissanen, "Modeling by shortest data description," *Automatica,* vol. 14, pp. 465–471, 1978.

[19] R.E. Krichevsky and V.K. Trofimov, "The Performance of Universal Encoding," *IEEE Trans. Inform. Theory,* vol. IT-27, pp. 199 - 207, March 1981.

[20] P.A.J. Volf and F.M.J. Willems, "Context Maximizing: Finding MDL Decision Trees," *Proc. 15-th Symp. on Inform. Theory in the Benelux,* pp. 192 - 199, Louvain-la-Neuve, Belgium, May 30 & 31, 1994.

[21] R. Nohre, *Some Topics in Descriptive Complexity,* Ph.D. thesis, Linkoping University, Sweden, 1994.

[22] F.M.J. Willems and Tj.J. Tjalkens, "Reducing complexity of the context-tree weighting method," *Proc. IEEE International Symposium on Information Theory,* p. 347, Cambridge, Mass., August 16 - 21, 1998.

[23] F.M.J. Willems, Tj.J. Tjalkens, and T. Ignatenko, "Context-Tree Weighting and Maximizing: Processing Betas." In *Proc. of Inaugural Workshop ITA (Information Theory and its Applications)*, San Diego, USA, 2006.

Tampereen teknillinen yliopisto
PL 553
33101 Tampere

Tampere University of Technology
Tampere International Center for Signal Processing
P.O.B. 553
FI-33101 Tampere

ISBN 978-952-15-1962-8
ISSN 1456-2774