INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

Estimation of Mixture Models

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
Qiang (Jonathan) Li

Dissertation Director: Andrew R. Barron

May 1999

ABSTRACT

Estimation of Mixture Models

Qiang (Jonathan) Li

Yale University

May 1999

We analyze mixture density approximation and estimation. We form a convex set of density functions by taking the convex hull of a parametric family, e.g. mixtures of the Gaussian location family. A sequence of finite mixture densities is formulated to provide a parsimonious approximation for the target density. If the target density itself is in the convex hull, we show that the approximation error goes to zero with a rate of 1/k, where k is the number of components in the approximation. If the target density is outside of the convex hull, the approximation error is equal to the best achievable error plus a term that goes to zero with a rate of 1/k. A greedy algorithm that introduces one component at each step is shown to achieve such an error rate.

Similarly, a greedy estimation algorithm is provided to find such approximation for data from an arbitrary density. This algorithm estimates one mixture component at one time. We prove that such an algorithm achieves a likelihood nearly as good as the MLE (maximum likelihood estimate) over the whole convex hull. And we identify the difference as being bounded by order O(1/k), where k is the number of components in the estimate.

Risks of such estimators are shown to be bounded by a sum of approximation error and estimation error. The error terms are identified. An optimal choice of k can be derived by minimizing the risk bound. Acting as a similar role as the bandwidth in non-parametric density estimation, k controls two error terms in opposite directions. A large k reduces approximation error and increases estimation error. A MDL (minimum

description length) principle is derived to provide an estimation method for k. And the estimated k is shown to achieve the risk bound as if we know the best k in advance.

A new information projection theory is derived to expand the approximating class to include its information closure. We prove the existence and uniqueness of a f^* in the closure of the convex hull $\mathcal{C}$(in a sense we identify), such that $D(f\|f^*) = \inf_{g\in\mathcal{C}} D(f\|g)$, where $D(f\|g)$ is the Kullback-Leibler divergence. And $\log(f_k) \to \log(f^*)$ in $L_1(f)$ for any sequence f_k in $\mathcal{C}$ with $D(f\|f_k) \to \inf_{g\in\mathcal{C}} D(f\|g)$. Other characterizing properties of f^* are also given.

ACKNOWLEDGEMENT

I deeply thank Professor Andrew Barron for three years of selfless teaching, advising and discussion. I also learned dedication and faith from him. I realize that it takes courage to be an excellent researcher.

I learned a lot from taking classes and having discussions with Professor John Hartigan, David Pollard, Joe Chang, Nick Hengartner and Marten Wegkemp. Their knowledge and enthusiasm help me to build a solid foundation and lead me into the wonderful world of statistics.

I also thank my fellow students at Yale Statisitcs Department. Jay Emerson, Jason Cross and Brendan Murphy are among the ones I talked with most. We shared passion and pain. Yuhong Yang helped me in the beginning of my graduate study.

I also thank Corky Kennedy and Susan Jackson-Mack, who make my life in the United States much easier than it could've been.

At last, but not the least, I want to thank my wonderful girlfriend, Ms. Yuping (Sharon) Hu. We went through ups and downs together in the past two years and a half. Her emotional support has been necessary for me to complete my study.

Contents

Chapter 1

Introduction

1.1 Mixed data sources

We frequently encounter data from mixed sources. For instance, if we measure the heights of all people on the earth and look for a probabilistic model for the data, we could use a mixture of two Gaussians as our first try. We know by *a priori* knowledge that men and women have quite different heights genetically. Therefore, they can be considered as two different sources of height data. It's also possible to consider other factors besides gender. For example, race, nationality and age are all good predictors for heights. The combination of all those factors can represent different sources as well. In this case, we can measure all those factors and label each observations to de-mix the sources. But it's not always possible to do that as we will see in the following example.

Another example involves mixture of Poissons. Large insurance companies receive thousands of claims each year. Some of the claims are made by the same person. In fact, it's valuable for the companies to know how many claims will be made by a particular customer. Studies showed that a mixture of Poisson distributions is appropriate for the distribution of the number of claims made by individuals (see Simar[1976]). Most of

people have very few claims each year. Their claims correspond to a Poisson distribution with a small rate, while a small number of people make a lot of claims, for whom another Poisson distribution with a high rate of claims is more appropriate. When we have a sample consisting of all kinds of customers, it's obvious that a mixture model should be applied. Unlike example on heights data, we can not label customers according to some characteristics of the customers. (It might be theoretically possible after some research were done on human nature of making claims. We assume it's more difficult than estimating mixture models!)

In image processing, it's natural to treat an image as a mixture of densities. An image is a composition of different textures, parts and colors. Image segmentation using mixture models has been a powerful tool for image coding, reconstruction, and classification (see, for example, Liang et al[1992], Sclove [1983], O'Sullivan [1993][1994]).

In fact, we encounter data generated by mixed sources of in many fields. Hartigan[1975] gives a large list of disciplines that have practical concerns about mixed data sources.

1.2 Mixture Models

Depending on the data source, we often are presented with various tasks in mixture modeling. So we need mixture models with different components and characteristics.

We use a particular component family depending on the data and purpose. For example, stock returns tend to have long tails because all too often, stock price have dramatic jumps or drops. A mixture of Gaussian densities has been used to uncover the long tail behavior (Weigend, A. et al [1999]). My view is that a mixture of Cauchy might be a better choice. See table 1.1. for a crude classification of possible applications.

Mixture models with various complexity can be used for different purposes. For in-

Table 1.1: Different Data Types

data type	component family	example
discrete	poisson, binomial	insurance claims
continuous	normal	image and speech
fat tail	cauchy	stock and bond returns

Table 1.2: Different Mixture Models

dimension	number of components	purpose	example
small	small	intepretation	clustering, classification, data mining
small	large	representation approximation	image segmentation, speech modeling data compression
large	small	discovery	knowledge discovery
large	large	unlimited possibilities	unlimited possibilities

stance, clustering analysis uses mixture models with small number of components. Interpretation of each components is important under clustering context. See table 1.2. for other examples. In particular, when dimension is large, the lack of computational methods and the curse of dimensionality make the actual applications of mixture models scarce.

More generally, we can use methodology of mixture models in other contexts. See table 1.3. Neural Networks involve a mixture of conditional logistic densities, for instance.

We can also compare mixture models with a few other related methods to gain more sense of use of mixture models.

- kernal density estimation

  Taking a mixture of some parametric family can provide a flexible family with clear

Table 1.3: Mixture of Different Objects

mixing object	context	examples
marginal density	density estimation	Gaussian mixtures
conditional density	function estimation	neural networks

intepretation. Each component can often be intepreted as a cluster, an agent, or a class depending on the context. A mixture model provides a often concise and parsimonious description for a data source. A kernal estimate essentially needs every data point in its final representation, while a mixture density only needs parameters of each components for a complete representation. Kernal estimates lack intepretability despite all of its nice statistical properties.

- hierachical cluster methods

  There are a lot of clustering techniques based on ad-hoc criteria such as hierarchical clustering (see Hartigan[1975] for a survey). It was shown that many of those criteria can be reproduced using mixture models (see Banfield and Raftery[1993]). Mixture-model-based clustering gains more and more attention because they fit into the framework of the classical statistical inference naturally (see Leroux[1992] for an example). Mixture-model-based clustering gives clustering a sound probabilistic ground. More importantly, it can generate many new methods systematically.

1.3 Estimating Mixture Models

When we estimate a mixture model, we need to consider three factors: approximation, estimation, and computation. The proper trade-off between them under proper contexts determines whether we should choose one method over another.

1.3.1 Classical Approximation-Estimation Trade-off.

Suppose the data are drawn from an underlying density $f(x)$, on which we impose no restrictions. We use a k-component mixture model. If the truth is not a k-component mixture, an approximation error will occur even if we have infinitely many data points.

On the other hand, if the truth is a k-component mixture, an estimation error will occur because we estimate the true density from finite data points. The more components we use, the smaller the approximation error and the bigger the estimation error. In kernel smoothing, such a trade-off is decided by bandwidth.

1.3.2 Curse of Dimensionality

When data goes beyond 1 or 2 dimension, computational aspects of an estimation method become very important. Many methods that work very well for low dimensional data fail in high dimension. Typically, there are two reasons.

- The sample size required to reach certain error bound goes up exponentially with dimensionality. An error bound with a rate such as $(\frac{1}{n})^{(1/d)}$ is unbearable when d is large. Kernel-smoothing methods typically use local averaging. In high dimension, data become so scarcely scattered that it takes an exponentially-growing sample size to reduce the estimation error.

- The search for the global optimum is a typical operation for classical methods such as maximum likelihood or maximum posterior estimation. In high dimension, the task of optimization can be a NP-complete problem unless we have a special structure such as convexity.

In a word, the so-called Curse of Dimensionality dictates that the issue of computation has to be taken into consideration to achieve a reasonable statistical procedure.

In Neural Networks and Projection Pursuit, two popular function estimation methods, the Curse of Dimensionality is attacked by using flexible basis functions. In pioneering work by Andrew R. Barron [1993] and Lee K. Jones [1992], approximation bounds for Artificial Feed-forward Neural Networks were proved to be unrelated to the dimension of

the data. The rate of convergence was shown to be of order $O(1/k)$, where k is the number of nodes in the Neural Networks. In addition, a greedy algorithm that searches for one component at one step was proven to achieve the same approximation error bound.

But we need to be cautious. As Barron [1993] pointed out, "it is not known whether there is a computational algorithm that can be proven to produce accurate estimates in polynomial time as a function of the number of variables for the class of functions studied here." Neural Networks avoided the effects of curse of dimensionality in terms of approximation error but not in terms of computational complexity. Nevertheless, a greedy algorithm cuts the otherwise kd dimensional search to a d dimensional search.

In density estimation, as David Scott[1992, chapter 7.2] put it, kernel smoothing beyond dimension 5 is practically impossible due to the lack of sufficient sample size to reduce both approximation and estimation errors. Procedures have been devised to reduce the dimensionality by projecting a density onto subspaces with lower dimension. Two most prominent examples are Principal Components and Projection Pursuit.

Another important observation in Scott[1992, chapter 7.3] indicates that it's more important to know where to look rather than know the density values for density estimation in high dimension. Density values are practically zero in most of the space anyway. In many multivariate classification schemes, we usually are looking for local clusters. This implies that a mixture-model-based density estimation can be more efficient in high dimension because we can focus on the search for local components.

1.4 Layout of the Dissertation

The thesis is organized as follows.

First, we review some mixture model estimation methods in chapter 2 including the popular MLE-EM algorithm and the Bayesian-MCMC algorithm. We also discuss in

detail the problem of determining number of components k.

In chapter 3, we present the main results. A sequential maximum likelihood estimator is introduced. This greedy-fashion algorithm is shown to achieve a likelihood nearly as good as if we maximize likelihood over $\mathcal{C}$. An approximation result shows that we have a sequence of q_k with $D(f\|q_k)$ that approaches $D(f\|\mathcal{C})$ with a dimension-independent rate of $O(\frac{1}{k})$, where k is the number of components. Again, a greedy algorithm is provided to achieve such a rate. Various risk bounds are also deduced. An MDL principle is derived to do an automatic selection of k. The selected $\hat{k}$ achieves a risk bound almost as good as if we know the best k in advance.

In chapter 4, a new theory of information geometry is established, which runs parallel with Topsoe and Csiszar's Information Projection Theory (see Csiszar[1984]). We show the existence of f^*, the characterizing property of f^*, and a Pythagorean like identity. The results in this chapter play an important role in the next chapter.

The proofs of main results are gathered in chapter 5.

We discuss issues related to the estimator in chapter 6. A brief treatment for an L^2 story is also given.

An introduction of Hilbert space theory in appendix A helps to motivate the information geometry. Indeed, the idea of the information projection is similar to projection onto convex subsets in Hilbert space.

I also gathered some inequalities in appendix B. They are fundamental tools for information theorists and statisticians alike. Many proofs throughout the thesis use these inequalities.

Chapter 2

Estimation of Finite Mixture Models.

2.1 Set-up of Finite Mixture Models.

Let $\mathcal{X}$ be a measurable space. In particular, $\mathcal{X}$ is a subset of R^m equipped with Borel set throughout this thesis. Let Φ_b be probability measures on $\mathcal{X}$ indexed by a parameter b and $b \in \Theta \subset R^d$. We denote this parametric family by

$$G = \{\Phi_b, b \in \Theta \subset R^d\}. \tag{2.1}$$

We also assume that each Φ_b has a density function $\phi_b(x), x \in \mathcal{X}$ with respect to a common dominating measure λ. This density function will also be denoted as $\phi(x, b)$. A finite mixture model with k components is defined as:

$$f_p(x) = \sum_{i=1}^{k} p_i \phi_{b_i}(x), \qquad \sum_{i=1}^{k} p_i = 1. \tag{2.2}$$

There are three sets of parameters we are interested in. They are 1) k, the number

of components, 2) $p_i, i = 1, 2, ..., k$, weight of each component and 3) $b_i, i = 1, 2, ..., k$, parameters of each component.

2.2 Estimation methods with known k

It's usually assumed k is known *a priori* through modeling or determined by some other test procedures. We will discuss the determination of k in the next section.

2.2.1 Maximum Likelihood and EM algorithm.

The maximum likelihood criterion is the most commonly applied criterion in the problem posed above. Let $b = (b_1, ..., b_k), p = (p_1, ..., p_k)$, where the b's are the parameters, and the p's are the weights of each components. Data $x_i, i = 1, 2, ...n$, are i.i.d $\sim f_p(x) = \sum_{i=1}^{k} p_i \phi_{b_i}(x)$,, where $\sum_{i=1}^{k} p_i = 1$. The likelihood of (b, p) for data x^n is

$$\begin{aligned} L(x^n; p, b) &= \prod_{j=1}^{n} f_p(x_j) \\ &= \prod_{j=1}^{n} \sum_{i=1}^{k} \phi(x_j, b_i) p_i. \end{aligned}$$

The log likelihood is:

$$\begin{aligned} l(x^n; p, b) &= \log(L(x^n; p, b)) \\ &= \sum_{j=1}^{n} \log f_p(x_j) \\ &= \sum_{j=1}^{n} \log \sum_{i=1}^{k} \phi(x_j, b_i) p_i. \end{aligned}$$

The maximum Likelihood criterion looks for parameters $(\hat{p}, \hat{b})$ that maximized L or, equivalently, l.

We differentiate the function plus Lagrange multiplier to include the constraint $\sum_{i=1}^{k} p_i = 1$:

$$\mathcal{L} = l(p, b) - \lambda(\sum_{i=1}^{k} p_i - 1).$$

The normal equations are:

$$\frac{\partial \mathcal{L}}{\partial p_i} = \sum_{j=1}^{n} \frac{\phi(x_j; b_i)}{f(x_j)} - \lambda = 0, \ i = 1, 2, ..., k. \tag{2.3}$$

$$\frac{\partial \mathcal{L}}{\partial b_i} = \sum_{j=1}^{n} \frac{p_i \partial \phi(x_j; b_i)/\partial b_i}{f(x_j)} = 0, \ i = 1, 2, ...k. \tag{2.4}$$

Multiply the first normal equation by p_i and sum over all i, we easily get: $\lambda = n$.

We can try to find the global maximum by solving the $2k$ normal equation. A Newton-Raphson algorithm will quickly converge if we are near the global maximum. The EM algorithm introduced in Dempster et al[1977] provides an iterative solution for the problem. The EM algorithm is usually simple to apply, while Newton-Raphson can be complicated with numerical matrix inversion and all. Neither of them can guarantee a finding of global maximum. See Titterington, Smith and Makov [1985], 84-90, for more details.

The scheme can be described as the following:

Define Z_i (i=1,...n) as a Multi-Bernoulli distributed random variable, indicating which component data X_i belongs to, i.e.

$$Z_i = \begin{pmatrix} Z_{i1} \\ ... \\ Z_{ij} \\ ... \\ Z_{iN} \end{pmatrix}$$

where $Z_{ij} = 1$ if X_i belong to jth component, 0 otherwise. Z_i is a Multi-Bernoulli distributed random variable with $P = (p_1, ..., p_N)$, where p_j indicate the membership probability. Now the density of Z_i is

$$p_1^{z_{i1}} ... p_N^{z_{iN}}.$$

The density of $X_i|Z_i$ is

$$\prod_{k=1}^{N} f_k^{Z_{ik}}(x_i).$$

So the joint density of (X_i, Z_i) is :

$$\prod_{k=1}^{N} p_k^{Z_{ik}} f_k^{Z_{ik}}(x_i). \tag{2.5}$$

EM algorithm: We initialize the parameters with $P^{(0)}, \theta_j^{(0)}, j = 1, ...N$. Then the expectation step leads to replace the missing data z by its expectation:

$$Z_{ij}^{(0)} = E(Z_{ij}|X_i, \theta_j^{(0)}, P^{(0)}) = \frac{p_j^{(0)} f_j^{(0)}(x_i)}{\sum_{t=1}^{N} p_t^{(0)} f_t^{(0)}(x_i)}$$

Then we maximize the likelihood of *pseudo-complete data* (X_i, Z_i) , which is easy because equation 2.5 is a product form. It can be shown that this scheme is equivalent to solving the normal equations 2.3 and 2.4. It can also be shown that the iterations will always increase the value of likelihood function. So if likelihood function is bounded, the scheme will converge (see, for instance, Everitt and Hand[1980], Ripley[1995]). The limit will be a point where the gradient of the log-likelihood is zero.

2.2.2 Bayesian Approach: Markov Chain Monte Carlo.

We can also apply Bayesian methodology to the estimation of mixtures. The Bayesian approach provides a *posterior* distribution of the parameter vector b through the Bayes theorem:

$$f(b|X^n) = \frac{f(X^n|b)\pi(b)}{\int f(X^n|b)\pi(b)db}$$

where $\pi(b)$ is the prior probability density of b, and $f(X^n|b)$ is the density of X^n. In our case, $f(X^n|b)$ is a mixture density. Again, it's very difficult to evaluate the integral directly to get the posterior distribution. Instead, a sequence of simulation methods, so-called Monte Carlo Markov Chain, is introduced. We discuss those ideas in a quite general set-up.

To simplify our notation, from now on, we will absorb weights into parameter vector b. We give b corresponding conjugate priors. Then we consider the missing data structure introduced in EM algorithm. Instead of replacing missing data by its expectation, we generate missing data by sampling from its conditional distribution. Instead of maximizing the likelihood of complete data, we sample from posterior distribution of the complete data.

(a) generate $\underline{z}^{(m)} \sim f(\underline{z}|\underline{x}, b^{(m)})$, the conditional distribution of missing data.

(b) generate $b^{(m+1)} \sim \pi(b|\underline{x}, \underline{z}^{(m)})$, the posterior distribution w.r.t. the complete data.

For (a), we can generate $z_1, ..., z_n$ separately since they are conditionaly independent given $x_1, ..., x_n$ and $b^{(m)}$. For (b), if the $b_1, ..., b_k$ are a priori independent, then the result of conditioning on $z_1, ..., z_n$ labeling from (a) is to split the data set into k a posteriori independent sets.

When the $b_1, ..., b_k$ are not a priori independent (p_i for instance), step (b) is not as

clean as the independent case. We could use **Gibbs sampling** to deal with one parameter at one time while conditioning on all other parameters and complete data as following:

(a) generate $z^{(m)} \sim f(z|x, b^{(m)})$, the same procedure.

(b)

1. generate $b_1^{(m+1)} \sim \pi(b_1|x, z^{(m)}, b_2^{(m)}, ..., b_s^{(m)})$ condition on all of other parameters.
2. generate $b_2^{(m+1)} \sim \pi(b_2|x, z^{(m)}, b_1^{(m+1)}, b_3^{(m)}, ..., b_s^{(m)})$
3. ...
4. generate $b_s^{(m+1)} \sim \pi(b_s|x, z^{(m)}, b_1^{(m+1)}, b_2^{(m+1)}, ..., b_{s-1}^{(m+1)})$

2.3 Determining k

It's critical to determine k, the number of the components in the mixture density. In the EM and MCMC algorithms, we usually assume k is known. Richardson and Green [1997] devised a so-called reversible jump Markov chain to accommodate this problem for MCMC algorithm. Still, very little is known about the behavior of such Markov chains.

There are situations in practice that a good *a priori* estimate can be made about k based on observation or experience. For instance, in speech recognition, it's often assumed that k is close to the number of phonemes. It's commented by Hartigan[1975] that such phenomena are quite common: "we are often classifying data which are classified already by people who collected data".

Here we will discuss statistical strategies for choosing k based on data.

2.3.1 Hypotheses testing

Hypothesis testing is a powerful approach for selecting models with wide applications. We can consider a hypothesis testing set up as the following, with $n_1 > n_0$,

$$H_0 \ : \ k = n_0$$

$$H1 \ : \ k = n_1$$

A natural candidate of testing statistic is a likelihood ratio test. The following ratio is formulated:

$$\lambda = L_0/L_1$$

where L_0, L_1 are maximums of likelihood under H_0, H_1 respectively. Under regularity conditions about the density function and a hidden assumption that b_0(the true value) is in the interior of Θ(the parameter space), we know that the sampling distribution of $-2\log_e \lambda$ under null hypothesis H_0 is distributed asymptotically χ^2, where the degrees of freedom of χ^2 is the difference in the number of parameters between H_1 and H_0 . Wolf[1971] suggested that such a test is not valid in the context of testing the number of components of a mixture. Under the null hypothesis, the parameters lie on the boundary of the full parameter space under the alternative hypothesis. Hartigan[1985] and Xu[1993] also observed that mixture model $(1-p)K(0,1)+pK(b,1)$ is not identifiable without constraints on parameters (p, b). Furthermore, Hartigan[1985] showed that in the Gaussian mixture setting, when $n_0 = 1, n_1 = 2$, log likelihood ratio goes to $+\infty$ in probability. Bickel and Chernoff[1992] gave the rate of divergence as $O_p(\log\log n)$.

Some modified tests have been suggested and they were reviewed in Everrit[1980]. However, Everrit concluded as the following: *no really adequate tests are available and this may be a consequence of the problem rather than any lack of ingenuity.* He also

suggested a reason as: *given enough components we can always find a mixture that 'fit' a set of data.*

2.3.2 Model Selection

Model selection approaches have been receiving attention since Akaike[1973] introduced the famous AIC criterion. Some other famous alternatives like Schwartz[1978]'s Bayesian Information Criterion(BIC) and Rissanen[1978,1980]'s Minimum Description Length(MDL) all share similar formula as penalized maximum likelihood criteria. The penalty terms are usually related to the number of parameters in the model and the sample size. As we have noted before, we can always find a mixture that fits a set of data perfectly if we are given an unlimited number of components. An extreme consideration would be to take each different sample as a component. Apparently, there is a danger of over-fitting the data. The rationale of penalized maximum likelihood is to give bigger penalty to models of larger size. Hopefully, the right amount of the penalty can achieve a good model which balances the fitting of the data and the complexity (size) of the model. The question is what amount is the *right* amount. Different criteria have different point of views while answering this question. AIC penalizes each additional parameter by 1. BIC uses $\frac{1}{2}\log(n)$. MDL uses the code length for the description of a model in the model class.

There are some merits that promote the Model Selection approach over Hypothesis Testing. As Akaike[1987] pointed out, test procedures do not penalize for over-parameterization because usually a saturated model is used as reference. In multiple-choice selection, it's easier to apply model selection than to apply a sequence of hypothesis tests, which involves choices of a number of dependent significance levels.

It seems that model selection criteria are readily applied in the clustering context. We just need to compute a penalized maximized likelihood for each K. However, a more

careful inspection of the conditions under which they can be applied shows that they can fail in the context of mixture density estimation.

The derivation of BIC used quadratic approximation of log likelihood around the true value. As we have noted in last section, the true value in mixture density estimation might be on the boundary of the full model parameter space. Therefore, it's not valid to use BIC directly in the context.

The derivation of AIC used the fact that asymptotically, likelihood ratio is distributed as χ^2. By the same reasoning as in Hypothesis Testing, AIC doesn't work in the context of mixture density estimation.

We will show that a MDL (minimum description length) principle can be used to select k in our mixture density estimation algorithm. The selected k achieves risk bound nearly as good as if we know the best k in advance.

Chapter 3

Iterative Estimation of Mixture Models

We want to formulate a mixture model class that includes finite, infinite and continuous mixtures of the parametric family G. This enlarged class gives us notational and conceptual convenience as we will show.

Let $\mathcal{C} = \{Q_P : \int \Phi_b P(db)\}$, where P is a probability measure on Θ, be the class of all convex combinations of elements in G. We denote this convex hull as $\mathcal{C} = CONV(G)$. (It includes the usual convex hull $\mathcal{C}^{Finite}$ of all finite convex combinations of G.)

It's clear that Q_P also has a density with respect to λ. We denote it by $q_P(x) = \int \phi_b(x) P(db)$ for $x \in \mathcal{X}$. If P is taken to be a discrete probability measure with finite support points that have non-zero mass, we are back to the familiar finite mixture model.

In this chapter, we will lay out the structure of a new iterative estimation method for mixture models and show its main properties.

Section 1 describes the algorithm. In section 2, we present an approximation theorem for the case that $f \in \mathcal{C}$. We show that the approximation error is bounded by $\gamma c_f \frac{1}{k}$, where k is the number of mixture components in the approximation. Thus the approximation

rate $1/k$ is dimension independent. The constants c_f and γ are identified and analyzed in some examples. We show that c_f is determined by the clustering patterns of the true density and γ is determined by the magnitude of log densities in G. In section 3, we show that if the true density is outside the approximation family, the approximation error is bounded by $\gamma c_f \frac{1}{k}$, plus the smallest approximation error achievable by this family. In section 4, we show that the likelihood of our estimated k-component mixture density is almost as good as the MLE in the whole family, with a difference of order $O(1/k)$. In section 5, we calculate statistical risks of the iterative maximum likelihood estimator. We also introduce an MDL principle to estimate the optimal number of components k^*, which minimizes the statistical risk bound.

3.1 Iterative Maximum Likelihood Estimation

In this section, we present our iterative algorithm that estimates a mixture density by introducing one component at one time iteratively.

We start the initial step by estimating a single component ϕ_b using maximum likelihood. In the case of a Gaussian location mixture, the first estimate is simply a Gaussian with mean equal to the sample mean. Obviously, this is a very crude approximation to the truth.

To get a more accurate estimate, we use an iterative strategy and increase the number of components in the mixture one at a time. Let $\hat{f}_{k-1}$ be the estimate after step $k-1$. Then we obtain $\hat{f}_k$ by taking a convex combination of $\hat{f}_{k-1}$ and a new component ϕ_b from G:

$$\hat{f}_k = (1-\alpha_k)\hat{f}_{k-1} + \alpha_k \phi_b, \tag{3.1}$$

where α_k and b are chosen to maximize the likelihood:

$$(\hat{b}, \hat{\alpha}_k) = arg \max_{b\in\Theta,\alpha\in(0,1)} \sum_i \log f_k(X_i) \tag{3.2}$$

$$= arg \max_{b\in\Theta,\alpha\in(0,1)} \sum_i \log[(1-\alpha_k)\hat{f}_{k-1} + \alpha_k \phi_b]. \tag{3.3}$$

The iterative procedure is stopped at an optimal $\hat{k}$ according to some criterion we will discuss in section 3.5 [Risk Bound and MDL].

Note that in this algorithm the optimization space is always $\{\Theta \otimes (0,1), \Theta \subset R^d\}$, thus the dimension never exceeds $d+1$. In a way, at each step it's like solving problem of mixture of two, while one of the two components is fixed. We can use EM or Newton-Raphson at each step.

A key result in section 3.4 [Nearly Maximum Likelihood] shows that above optimization procedure does nearly as well as the Maximum Likelihood Estimator over a full k-component mixture parameter space. Furthermore, it does nearly as good as the best density among all mixture densities in $\mathcal{C}$.

In the mixture problems with small d and large k, we can foresee that this type of algorithm offers a great advantage. Even in the problem with small k, this algorithm can provide an excellent starting point for EM and alike. More importantly, in large d case, this method offers a chance of working.

3.2 Approximation Error Bound When Truth Is In $\mathcal{C}$

In the classical setting, the truth is assumed to be a finite mixture density. In our setting, we don't make such an assumption. So an approximation error occurs because we use a finite mixture density to approximate the unknown truth. A natural question arises on how close the finite mixture density is to the truth. We show that finite mixture densities

provide good approximations to infinite mixture and continuous mixtures in the sense that the approximation error goes down with order O(1/k), where k is the number of components.

We use the Kullback-Leibler divergence between f and f_k, $D(f\|f_k)$, as our measure of closeness between two densities.

The K-L divergence is defined as

$$D(f\|f_k) = \int f(x) \log \frac{f(x)}{f_k(x)} \lambda(dx). \tag{3.4}$$

It is a commonly used measure of closeness between densities although it's not a metric. Besides its mathematical properties, it has a close connection with the MLE. Observe that the log-likelihood $\frac{1}{n}\sum \log f_k(x)$ is the sample analogue of $\int f \log f_k$. Consequently, maximum likelihood behavior for large n is dictated in part by corresponding minimization of K-L divergence.

The next theorem establishes an error bound for using finite mixtures to approximate densities in C.

THEOREM 3.1 (Fundamental Theorem of Approximation Using Mixture Models) *Let $G = \{\phi_b, b \in \Theta \subset R^d\}$ and $C = CONV(G)$. Let $f(x) = \int \phi_b(x) P(db) \in C$. There exists f_k, a k-component mixture of ϕ_b such that*

$$D(f\|f_k) \leq \frac{c_f^2 \gamma}{k} \tag{3.5}$$

where c_f^2 is a constant determined by f and γ is a constant determined by C. If f has M components, then $c_f^2 \leq M$.

More specifically, a greedy procedure achieves such an approximation error bound.

THEOREM 3.2 (Iterative Approximation) *Let G, C, and f be the same as before. Initially, choose ϕ_{b_1} in G such that $D(f\|\phi_b)$ is minimized. Let $q_1 = \phi_{b_1}$. Define f_k in an iterative fashion for $k = 2, 3...$,*

$$f_k = (1 - \alpha) f_{k-1} + \alpha \phi_b, \tag{3.6}$$

where α and b are chosen to minimize $D(f\|f_k)$. Then

$$D(f\|f_k) \leq \frac{c_f^2 \gamma}{k}. \tag{3.7}$$

REMARK The rate of convergence, $1/k$, doesn't depend on the dimension d of parameter space. We have a dimension-independent bound on the constant c_f in the case that f has a given number of components. Often γ depends on d linearly. So sequences of finite mixtures of densities from G can approach densities in the convex hull C having finite c_f at a fast dimension-independent rate. Note that such a convex hull C can be a very large family. For instance, consider the convex hull C_σ of Gaussians parameterized by location and scale with a lower bound σ_0 on the scale parameter. In a sense, as $\sigma_0 \to 0$, the class C_σ approaches the collection of all densities. Thus the class C can be made very large. However, as we will see in the following, the larger the family C is, the larger the constant γ on the bound will be. □

Now we specify the constants in the bound. First let's define $c_{x,P}^2$.

DEFINITION 3.1 ($c_{x,P}$) *For $\phi_b(x) \in G$, where $b \in \Theta \subset R^d$ and $x \in \mathcal{X} \subset R^m$, and for $P(db)$ a probability measure on Θ, we define*

$$c_{x,P}^2 = \frac{\int \phi_b^2(x) P(db)}{(\int \phi_b(x) P(db))^2}. \tag{3.8}$$

Note that $c^2_{x,P} - 1$ is the coefficient of variation of $\phi_b(x)$ with respect to the distribution $P(db)$ for a given x. The coefficient of variation is defined as the variance divided by the square of the mean. Now define $c^2_{F,P}$ based on $c^2_{x,P}$.

DEFINITION 3.2 ($c_{f,P}$) *For F a probability measure on $\mathcal{X}$ and P a probability measure on Θ, define*

$$c^2_{F,P} = \int c^2_{x,P} F(dx). \tag{3.9}$$

If F has a density f, we denote it $c^2_{f,P}$. When the true density $f(x)$ is of the form $\int \phi_b(x) P(db)$, we write shorthand c^2_f for $c^2_{f,P}$. For some families there can be more than one mixture representation of a certain function. Each such representation yields a valid bound, so for f in $\mathcal{C}$, define

$$c^2_f = \inf_{P: f = \int \phi_b P(db)} c^2_{f,P}.$$

Obviously,

$$c^2_{f,P} = \int \frac{\int \phi_b^2(x) P(db)}{[\int \phi_b(x) P(db)]^2} f(x) \lambda(dx) \tag{3.10}$$

which reduces to

$$\int \frac{\int \phi_b^2(x) P(db)}{\int \phi_b(x) P(db)} \lambda(dx)$$

when $f(x) = \int \phi_b(x) P(db)$.

The constant c^2_f depends on the true density f. We can establish a rough upper bound for it through the following lemma.

LEMMA 3.1 (Bound for c_f) *Suppose $f = \int \phi_b P(db)$ and P is a discrete distribution with M values, i.e. $f = \sum_{i=1}^{M} p_i \phi_{b_i}$. Then*

$$c^2_f \leq M \tag{3.11}$$

with equality if and only if $\phi_{b_i}(x), i = 1, 2, ..., M$, have disjoint supports.

PROOF Write shorthand $\phi_i = \phi_{b_i}$. Then

$$\begin{aligned} c_f^2 &= \int \frac{\sum_{i=1}^M p_i \phi_i^2(x)}{\sum_{i=1}^M p_i \phi_i(x)} \lambda(dx) \\ &= \int \sum_{i=1}^M \left(\frac{p_i \phi_i(x)}{\sum_{i=1}^M p_i \phi_i(x)}\right) \phi_i(x) \lambda(dx) \\ &\leq \int \sum_{i=1}^M \phi_i = M. \end{aligned}$$

When $p_j \phi_j = \sum_{i=1}^M p_i \phi_i$ for every $1 \leq j \leq M$, the equality sign holds. This is only possible when ϕ_j have disjoint supports. □

This lemma shows that c_f^2 has an upper bound independent of dimensionality of $\mathcal{X}$. In particular c_f^2 is upper bound by the number of disjoint components in f.

Let's turn our attention to the constant γ, which also appears in the bound. We first define a:

DEFINITION 3.3 (a **and** γ) *Define*

$$a = a_{G,\mathcal{X}} = \sup_{\phi_1, \phi_2 \in G, x \in \mathcal{X}} \log\left(\frac{\phi_1(x)}{\phi_2(x)}\right). \tag{3.12}$$

Also define $\gamma = 4[\log(3\sqrt{e}) + a]$.

Thus we essentially have an upper bound on the L^∞ norm of $\log \phi(x)$. We shall look at a specific example to gain some sense on how large this bound can be.

EXAMPLE 1 (Normal Location Mixture With Bounded Support) *Let's first assume all densities are in a bounded support $\mathcal{X}$, a d-dimensional cube with side-length A. Suppose that G is the collection of d-dimensional Gaussian densities $\phi_b(x)$ with mean vector b and covariance matrix $\sigma^2 I$, which are then restricted to the cube $\mathcal{X}$. We assume b is also restricted to the cube x. We have a bound for $a_{G,\mathcal{X}}$ as the following,*

$$a_{G,\mathcal{X}} \leq \frac{dA^2}{2\sigma^2}. \tag{3.13}$$

PROOF

$$\begin{aligned} exp(a) &= \max_x(\max_{\phi_1,\phi_2} \frac{\phi_1(x)}{\phi_2(x)}) \\ &\leq \frac{\max_x \max_\phi \phi(x)}{\min_x \min_x \phi(x)} \\ &= \frac{\exp(-\frac{d}{2}\ln 2\pi\sigma^2)}{\exp(-\frac{d}{2}\ln 2\pi\sigma^2 - \frac{dA^2}{2\sigma^2})} \\ &= \exp(\frac{dA^2}{2\sigma^2}) \end{aligned}$$

So,

$$a \leq \frac{dA^2}{2\sigma^2}. \tag{3.14}$$

□

This constant is linear in dimension d. When σ approaches 0, the constant approaches infinity. This is a result of the trade-off between the size of the approximating family and the speed of the error converging to zero. The smaller σ is, the larger the family $\mathcal{C}$ is. In fact, if σ can be arbitrary non-zero real numbers, $\mathcal{C}$ can approximate any density functions. The identification of γ enables to make a wise choice on how small we can allow σ to be. A is usually taken to be the range of the data coordinates.

REMARK The convex hull $\mathcal{C}$ inherits a bound on its density ratios from the bound on density ratios in G. Indeed for any fixed ϕ in G and $g_P, g_{\bar{P}}$ in $\mathcal{C}$ $g_P(x)/\phi(x) = \int(\phi_b(x)/\phi(x))P(db)$ will be between e^{-a} and e^a and hence $g_P(x)/g_{\bar{P}}(x) = \frac{g_P(x)/\phi(x)}{g_{\bar{P}}(x)/\phi(x)}$ is between e^{-2a} and e^{2a}. □

3.3 Approximation Error Bound For Arbitrary Target Density

In this section, we consider an arbitrary target density, which is not necessarily in C. We still can use a finite mixture model to approximate the target function. We identify the error bound as the sum of the best approximation error achievable by C and an $O(1/k)$ term. The best approximation error is defined as $D(f\|C) = \inf_{q\in C} D(f\|q)$. In the next chapter, we will also show that a f^* achieves such an error and has some desired properties.

First we give a general result as follows.

THEOREM 3.3 (Approximation Error For General Target) *Let G, C, and f be the usual. There exists a f_k, a k-component finite mixture of ϕ_b, such that for all element $q_P = \int \phi_b P(db)$ in C,*

$$D(f\|f_k) \leq D(f\|q_P) + \frac{c_{f,P}^2 \gamma}{k}. \tag{3.15}$$

In addition, the iterative procedure defined in section 3.1 achieves such error bound.

By taking the infimum over all q_P in C, we have

COROLLARY 3.3.1 (Best Approximation Error For General Target)

$$D(f\|f_k) \leq D(f\|C) + \frac{c_{f,*}^2 \gamma}{k} \tag{3.16}$$

where $c_{f,*}^2$ is the limit infimum of $c_{f,P}^2$ for the set of all sequences of P such that $D(f\|q_P) \to D(f\|C)$.

Thus if $c_{f,*}^2$ is finite, $D(f\|f_k)$ approaches $D(f\|C)$ at rate $1/k$.

Moreover, let $C_f = \{q_P : c_{f,P}^2 < \infty\}$. If there exists q_P in C_f with $D(f\|q_P)$ arbitrarily close to $D(f\|C)$, then we have following corollary.

COROLLARY 3.3.2 (Limit of Best Approximation Error) *Suppose for each* $\epsilon > 0$, *there exists* q_P *with* $c_{f,P} < \infty$ *such that* $D(f||q_P) \leq D(f||\mathcal{C}) + \epsilon$ *then*

$$\lim_{k\to\infty} D(f||f_k) = D(f||\mathcal{C}).$$

In the limit, f_k has a K-L divergence that's as good as the infimum. We are also interested in the converging properties of f_k. In the next chapter, we will show that $\log(f_k)$ converges to a $\log(f^*)$ in $L^1(f)$ where $D(f||f^*) = D(f||\mathcal{C})$.

3.4 Nearly Maximum Likelihood

Now we come back to the iterative maximum likelihood algorithm we introduced in section 1. Instead of finding the global maximum likelihood estimate among all densities in $\mathcal{C}$, we find one component at one time, i.e., maximize the likelihood for one component at one iteration step. A surprising result about this procedure is that the resulting likelihood is nearly as good as the best likelihood we can achieve among all densities in $\mathcal{C}$. The difference is of order $O(1/k)$.

THEOREM 3.4 (Nearly Maximum Likelihood) *For every* $q_P \in \mathcal{C}$ *with the form* $q_P(x) = \int \phi_b(x)P(db)$ *and every* $x_1, x_2, ..., x_n$, *if* $\hat{f}_k$ *is the estimated mixture density using the algorithm in Section 3.1. after* k *steps, then we have*

$$\frac{1}{n}\sum_{i=1}^{n} \log \hat{f}_k(x_i) \geq \frac{1}{n}\sum_{i=1}^{n} \log q_P(x_i) - \frac{\gamma c^2_{F_n,P}}{k}, \tag{3.17}$$

where $c^2_{F_n,P} = \frac{1}{n}\sum_{i=1}^{n} c^2_{x_i,P}$ *and* $c^2_{x_i,P}$ *was defined in Definition 3.1.*

In particular, we have

$$\frac{1}{n}\sum_{i=1}^{n}\log \hat{f}_k(x_i) \geq \sup_{q_P \in \mathcal{C}}[\frac{1}{n}\sum_{i=1}^{n}\log q_P(x_i)] - \frac{\gamma c^2_{F_n,*}}{k} \tag{3.18}$$

where $c^2_{F_n,*}$ is the limit infimum of $c^2_{F_n,P}$ among all sequences of q_P that approach the supremum of the likelihood.

3.4.1 Metric Entropy Condition on the Family

In the proof of my risk bounds, we need to impose a smoothness condition on the log densities $\log \phi_b$. It can be stated in the following Lipchitz condition.

$$\sup_x |\log \phi_b(x) - \log \phi_{b'}(x)| \leq B\sum_{j=1}^{d}|b_j - b'_j| \tag{3.19}$$

where B is a bounded constant, b_j is the jth coordinate of b.

Such a condition is satisfied by Gaussian densities on a bounded support.

Under such a condition, we can quantize the parameter space and obtain a parametric family with finite cardinality, while the estimated densities restricted to such a family do as well as the estimates obtained on the continuum of the parameter space.

Without loss of generality, we assume that the parameter space Θ is in a d dimensional cube with side length less than A. We also assume that the width of the grid on each side of the cube is ϵ. After such simplification, the parameter space of a k-component mixture density has finite cardinality. We denote it by $\Theta_{k,\epsilon}$. The cardinality is denoted by $Card(\Theta_{k,\epsilon}) = (A/\epsilon)^{kd}$.

In real computation, we usually work in a bounded and quantized parameter space. Quantization inevitably happens in practice because computer uses finite digits to represent a real number. So it's impossible to search a continuous parameter space anyway.

The following lemma establishes that if we use the quantized parameter space, the estimated density achieves likelihood almost as good as the estimated density obtained on the continuum. Intuitively, a Lipschitz condition guarantees that a quantization error on b will propagate to $\log(\phi_b)$ in a controlled manner.

LEMMA 3.2 (Nearly Maximum Likelihood with Quantization) *Let $q_P, x_1, ..., x_n$ be the usual. Let $\hat{f}_{k,\epsilon}$ be the mixture density estimate obtained on a grid of width ϵ in parameter space. We have a nearly maximum likelihood bound*

$$\frac{1}{n}\sum_{i=1}^{n}\log \hat{f}_{k,\epsilon}(X_i) \geq \frac{1}{n}\sum_{i=1}^{n}\log q_P(X_i) - \frac{\gamma c^2_{F_n,P}}{k} - kdB\epsilon \tag{3.20}$$

REMARK The term $(k+1)dB\epsilon$ upper bounds the error inflicted by the quantization. □

3.5 Risk Bounds and Approximation-adjusted Risk Bounds

Let $\hat{f}_k$ be the estimate at the kth step. We are interested in risk bounds of such an estimate. In addition, we want to estimate k and give risk bounds for $\hat{f}_{\hat{k}}$. We will introduce an MDL estimate of k and give a risk bound for it.

First consider the case of fixed k. Hellinger distance and Kullback-Leibler divergence are two widely-used loss functions between densities. K-L divergence is an upper bound for Hellinger distance. And when the density ratio is bounded away from zero, K-L divergence is upper bounded by a multiple of Hellinger distance. (See Appendix B for details regarding those relationships.)

We introduce approximation-adjusted loss functions because of the existence of non-diminishing approximation error when the target density is outside the model class. The

adjusted loss functions will simplify to the usual loss functions when the target density is in $\mathcal{C}$.

Approximation-adjusted K-L loss is defined as $D(f\|\hat{f}_k) - D(f\|\mathcal{C})$. Observe that when f is in $\mathcal{C}$, the loss is the same as the usual K-L losses.

We will be working on adjusted loss functions because corresponding risks converges to zero when $n \to \infty$.

THEOREM 3.5 (Kullback-Leibler Risk Bound) *Let $\phi_b, b \in \Theta_\epsilon$ be a parametric family of densities. Assume that ϕ_b satisfies smoothness condition as in the last section. Let $\mathcal{C} = \{q_P : \int \phi_b P(db),\ P$ a probablity measure on $\Theta.\}$. Let $\hat{f}_k = \hat{f}_{k,\epsilon}$ be the iterative maximum likelihood estimate after k steps. Let the data $X_1, ..., X_n$ be i.i.d. according to f. We have a bound for Kullback-Leibler risk,*

$$\frac{1}{\gamma}[ED(f\|\hat{f}_k) - D(f\|\mathcal{C})] \leq \frac{\gamma c_{f,*}^2}{k} + \frac{2kd\log(A/\epsilon)}{n} + kdB\epsilon. \tag{3.21}$$

Note that $kd\log(A/\epsilon)$ is equal to the logarithm of $Card(\Theta_{k,\epsilon})$. And the term $kBd\epsilon$ comes from the ϵ-quantization of the parameter space Θ. Choose ϵ to minimize the risk bound. We have

$$\epsilon^* = \frac{2}{nB}.$$

The resulting bound is

$$\frac{1}{\gamma}[ED(f\|\hat{f}_k) - D(f\|\mathcal{C})] \leq \frac{\gamma c_{f,*}^2}{k} + \frac{2kd\log(nABe/2)}{n}. \tag{3.22}$$

We can further choose k to minimize the risk bound. In practice, however, we can not do so because the $c_{f,*}^2$ depends on unknown density f. Instead, an MDL (minimum description length) criterion is used to estimate k. In such a criterion, instead of minimizing

the risk bound, we choose $\hat{k}$ to minimize the length of a code to describe the data. The risk of the resulting $\hat{f}_{\hat{k}}$ is bounded by a penalized risk bound minimized over k.

Specifically, we have the following theorem.

THEOREM 3.6 (Minimum Description Length) *Let* $\sum e^{-l(k)} \leq 1$. *An MDL principle to choose* k *is given as follows.*

$$\hat{k} = \arg\min_k \{\frac{1}{n} \log \frac{1}{\hat{f}_{k,\epsilon}(x^n)} + \frac{2 \log Card(\Theta_{k,\epsilon})}{n} + 2\frac{l(k)}{n}\} \tag{3.23}$$

Then with $\gamma_1 = 2 + 2a$, *we have risk bound*

$$\frac{1}{\gamma_1}[ED(f\|\hat{f}_{\hat{k}}) - D(f\|\mathcal{C})] \leq \min_k \{\gamma \frac{c_{f,*}^2}{k} + 2\frac{\log card(\Theta_{k,\epsilon})}{n} + 2\frac{l(k)}{n} + kdB\epsilon\} \tag{3.24}$$

Note that we can also choose an optimal $\epsilon^* = \frac{2}{nB}$. The risk bound becomes

$$\min_k \{\gamma \frac{c_{f,*}^2}{k} + 2\frac{kd \log(nABe/2)}{n} + 2\frac{l(k)}{n}\}$$

REMARK Here $l(k)$ has the interpretation of codelength of a uniquely decodable code to describe k. We are using logarithm base e. For the binary coding interpretation one would have $\sum 2^{-l(k)} \leq 1$ and use base 2 logarithms in 3.23 and 3.24.

The MDL principle can also be viewed as a penalized maximum likelihood principle. The penalties are the twice code-length for description of a model with k components. We introduce a 2 factor here because it arises in the proof. A recent discussion with Dr. Wu Chou, who is with Speech Processing Group at Lucent Bell Labs, indicates that he has empirical evidences to support usage of two times code-length as the penalty. □

Note that $\hat{f}_{\hat{k}}$ has a risk nearly as good as if we know the best k^* in advance. An extra term in the risk, $\frac{2l(k)}{n}$, is the price we pay for estimating k.

In the next chapter, we will digress from the main theme. We will discuss an information geometry which leads to the theorems about f^*. The existence and properties of f^* are essential to the proofs of the risk bounds.

Chapter 4

Projection Theories in A Space of Probability Measures: Information Geometry

Information geometry resembles projection theory in Hilbert spaces in many ways. I will present the information geometry and its connection with Hilbert space theory in this chapter. For a description of projection theory in Hilbert space, see appendix A. The Hilbert space case is based on the treatment in the course notes of Professor David Pollard at the Yale Stat 600 class in 1999.

In density estimation, we are interested in a space of probability measures or their densities when a common dominating measure is specified. Information geometry deals precisely with the properties of projection in a space of probability measures.

In such a space, convex subsets are of special importance. Let's look at a few examples of convex subsets in a space of probability measures.

EXAMPLE 2 (Moment Constraints) *Let S be a measurable function on a given*

measurable space. Then $C = \{Q : E_Q S = a\}$ *is a convex set of probability measures. Let* P_0 *be a fixed probability measure. In Kullback's method of minimum discrimination information inference, empirical moment constraints are given and one seeks a measure in* C *that achieves* $\inf_{Q \in C} D(Q||P_0)$ *or a sequence of measures that approach the infimum.*

EXAMPLE 3 (Mixture Density Estimation) *In the context of this thesis,* $C = \{\int \Phi_b P(db)\}$ *is the convex family of mixtures within which we seek maximum likelihood or minimum Kullback-Leibler divergence from the true distribution. Here each* x_i *is modeled as independent from a density possibly in* C*.*

EXAMPLE 4 (Bayes Mixtures) *Let* $P_{\underline{X}|\theta}$ *be a family of probability measures for random sequence* $\underline{X}$*, indexed by a parameter* θ*. Assignment of a prior* $H(d\theta)$ *leads to distribution for* $\underline{X}$ *of* $\int P_{\underline{X}|\theta} H(d\theta)$*, the collection of which forms a convex hull of statistical distributions.*

Under this setting, $D(P_{\underline{X}|\theta}||P_{\underline{X}})$ *represents the cumulative risk of predictive density estimation or the redundancy of a code based on* $P_{\underline{X}}$*. One may use information projection identities to show that if* $D(C||P_{\underline{X}})$ *is not zero, then* $P_{\underline{X}}$ *is inadmissible and may be replaced by an information projection* $P^*_{\underline{X}}$ *with* $D(p_{\underline{X}|\theta}||p^*_{\underline{X}})$ *strictly smaller (by the constant amount* $D(C||P_{\underline{X}})$ *) for all* θ*.*

4.1 Csiszar and Topsoe's Information Projection

As we know from convex projection in a Hilbert space, two essential elements are

1. Existence and uniqueness of the projection,

2. Characterization of the projection by a Pythagorean Inequality.

In the space of probability measures, Csiszar[1975] and Topsoe[1979] have shown the following key theorems about information projection.

DEFINITION 4.1 (I-projection) *Give a convex set $\mathcal{C}$ of probability measures and a measure P not necessarily in $\mathcal{C}$, a probability measure $P^* \in \mathcal{C}$ is called the information projection (or relative center of attraction) of P on $\mathcal{C}$ if for every sequence Q_n with $D(Q_n\|P) \to \inf D(\mathcal{C}\|P)$ we have $Q_n \to P^*$ (in total variation).*

THEOREM 4.1 (Existence and Uniqueness) *For any convex set $\mathcal{C}$ of probability measures and P for which $D(\mathcal{C}\|P) < \infty$, there exists a unique information projection P^*.*

PROOF See Topsoe[1979]. □

THEOREM 4.2 (Pythagorean Inequality) *Given P and a convex $\mathcal{C}$, the information projection P^* satisfies*

$$D(Q\|P) \geq D(Q\|P^*) + D(\mathcal{C}\|P) \text{ for every } Q \text{ in } \mathcal{C} \tag{4.1}$$

and consequently for any sequence Q_n with $D(Q_n\|P) \to D(\mathcal{C}\|P)$ we have

$$D(Q_n\|P^*) \to 0. \tag{4.2}$$

Moreover, either 4.1 or 4.2 characterizes the information projection.

PROOF See Topsoe[1979] which builds on earlier work of Csiszar [1975]. □

Csiszar's earlier work [1975] established the information projection inequality in the case that $\mathcal{C}$ is variation closed. Topsoe's extension provides the generalized information projection for arbitrary convex $\mathcal{C}$.

The T-C projection theorem has important applications in large deviation theory (see Csiszar [1984], also Cover and Thomas [1991]).

4.2 A New Information Projection Theory

In our theory, we reverse the order of the arguments in the K-L divergence. An analogous information projection theory is obtained. Applications to maximum likelihood estimation require this reversal of the order in the K-L divergence. We build upon a theory of Bell and Cover [1980], who in a portfolio selection contex developed the story under an assumption that a minimizer of $D(P\|Q), Q \in \mathcal{C}$ exists.

Again we consider a convex set $\mathcal{C}$ of probability measures. Let P be a probability measure of our interest. Define

$$D(P\|\mathcal{C}) = \inf_{Q\in\mathcal{C}} D(P\|Q).$$

Similar to T-C theory, we also want to establish existence, uniqueness and characterizing Pythagorean Identity of a projection P^* of P onto $\mathcal{C}$.

DEFINITION 4.2 (Reversed Information Projection) *Given a probability measure P with a density p and a convex set $\mathcal{C}$ of densities q, a function q^* is called the (reversed) information projection if for every q_n with $D(p\|q_n) \to D(p\|\mathcal{C})$, we have $\log q_n \to \log q^*$ in $L^1(P)$.*

THEOREM 4.3 (Properties of the Reversed I-Projection) *Let $\mathcal{C}$ be a convex set of probability measures Q with densities q and let P be a target measure with density p. Then the reversed I-projection q^* of P exists and is unique. Moreover it satisfies the following properties:*

1. $D(p\|q^*) = \inf_{q\in\mathcal{C}} D(p\|q)$,
2. $c_q = \int p\frac{q}{q^*} \leq 1, \forall q \in \mathcal{C}$,
3. $D(p\|q) \geq D(p\|q^*) + D(p\|\rho)$ *where* $\rho = \frac{pq/q^*}{c_q}$ *is a density depending on* q.

We will call q^* the "projection" in the sequel.

REMARK $D(p||\rho)$ is a disguised version of distance between q^* and q. As we will see it, it provides an upper bound for $\int p|\log(q^*) - \log(q)|$. So convergence of $D(p||\rho)$ to zero implies L^1-convergence of $\log(q)$. □

PROOF

1. This assertion is an immediate result of Lemma 4.3, "existence of a projection q^*".

2. This is proven in Lemma 4.4.

3. The Inequality immediately follows by re-arranging the difference in K-L divergence in Lemma 4.4 :

$$\begin{aligned} D(p||q) - D(p||q^*) &= \int p \log \frac{q^*}{q} \\ &= \int p \log \frac{p}{pq/q^*} \\ &= \int p \log \frac{p}{\rho} + \log \frac{1}{c_q} \end{aligned}$$

And we know $c_q \leq 1$, so $\log \frac{1}{c_q} \geq 0$.

□

4.2.1 Key Lemma: Characterizing Property of A Projection

First we need to establish a lemma allowing the interchange of integral and derivative to get a key property. Here we handle the case that q^* is in $\mathcal{C}$. Characterization of more general information projection is in subsequent lemma.

LEMMA 4.1 (characterize.projection) *Let $\mathcal{C}$ be a convex set of probability measures*

with densities and p be a density outside of C. Let q^* be a density in C. Then

$$D(p||q^*) = D(p||\mathcal{C}) = \min_{q \in \mathcal{C}} D(p||q)$$

if and only if

$$\int p\frac{q}{q^*} \leq 1, \forall q \in \mathcal{C}.$$

PROOF The "if" part is trivial: $\forall q \in \mathcal{C}$,

$$\begin{aligned} D(p||q^*) - D(p||q) &= \int p \log \frac{q}{q^*} \\ &\leq \int p(\frac{q}{q^*} - 1) \\ &\leq 0. \end{aligned}$$

REMARK The "if" part is true even if C is not convex. □

The "only if" part is the hard part:

For a given q in $\mathcal{C}$, construct $q_t = (1-t)q^* + tq, 0 \leq t \leq 1$. Let $D_t = D(p||q_t)$. Observe that $q_t \in \mathcal{C}$. So

$$D_0 \leq D_t, 0 \leq t \leq 1.$$

From the convexity of $-\log$, we observe that D_t is a convex function of t for $0 \leq t \leq 1$.

First we show that D_t is differentiable for $1 - \epsilon > t > \epsilon$, for a small $\epsilon > 0$.

$$\begin{aligned} D_t &= \int p \log \frac{p}{q_t} \\ &= \int p \log \frac{p}{(1-t)q^* + tq} \end{aligned}$$

For $\epsilon \leq t \leq 1-\epsilon < 1$, we have

$$\begin{aligned} |\frac{d}{dt}\log((1-t)q^* + tq)| &= |\frac{q-q^*}{(1-t)q^* + tq}| \\ &\leq max\{\frac{1}{1-t}, \frac{1}{t}\} \\ &\leq \frac{1}{\epsilon} \end{aligned}$$

So by the Dominated Convergence Theorem, we have that D_t is differentiable and

$$\frac{dD_t}{dt} = \int p\frac{q^* - q}{q_t}, 0 < t < 1.$$

Second we show that derivatives $\frac{dD_t}{dt} \geq 0$ for $t > 0$. Convexity of D_t gives the following:

$$D_t \leq \frac{h}{t+h}D_0 + \frac{t}{t+h}D_{t+h},$$

where $h > 0$ is small. Then by some algebra,

$$\frac{D_{t+h} - D_t}{h} \geq \frac{D_{t+h} - D_0}{t+h} \geq 0.$$

So $\frac{dD_t}{dt} \geq 0$ for $t > 0$.

At last, we show that $\frac{q^*-q}{q_t}, t > 0$ is a monotone decreasing sequence of functions as $t \to 0$. Take the derivative w.r.t. t,

$$\frac{d}{dt}\frac{q^* - q}{q_t} = \frac{(q^* - q)^2}{q_t^2} \geq 0.$$

In addition, for $0 \leq t \leq 1/2$, $\frac{q^*-q}{q_t}$ is bounded above by

$$\frac{q^* - q}{q_t}|_{t=1/2} = \frac{q^* - q}{q^*/2 + q/2}$$

$$= 2\frac{q^* - q}{q^* + q}$$
$$\leq 2.$$

Consequently, by the Monotone Convergence Theorem,

$$\int p\frac{q^* - q}{q_t} \to \int p\frac{q^* - q}{q^*}.$$

Then since $\int p\frac{q^* - q}{q_t} = D'(t) \geq 0$, it follows that $\int p\frac{q^* - q}{q^*} \geq 0$. Re-arrange it and we get

$$\int p\frac{q}{q^*} \leq 1.$$

□

REMARK R. Bell and T. Cover [1980,1988] proved this lemma using a truncation strategy. □

4.2.2 Existence of A Projection

First we establish a preliminary lemma.

LEMMA 4.2 (Continuity) *If $q_1, q_2, ..., q_n$ satisfy that $D(p||q_i) < \infty$, the function*

$$D(\underline{t}) = D(p||q_{\underline{t}})$$

is a continuous function over the range of $\underline{t}$, where

$$q_{\underline{t}} = \sum_{i=1}^{n} t_i q_i,$$

and $\underline{t} = (t_1, t_2, ..., t_n), 0 \leq t_i \leq 1, with \sum t_i = 1.$

PROOF Observe following domination:

$$\begin{aligned}
\log \frac{p(x)}{\sum_{i=1}^{n} t_i q_i(x)} &\leq \log \frac{p(x)}{\min_i q_i(x)} \\
&= \max_i \log \frac{p(x)}{q_i(x)} \\
&\leq \max_i \log^+ \frac{p(x)}{q_i(x)} \\
&\leq \sum_i \log^+ \frac{p(x)}{q_i(x)}.
\end{aligned}$$

Also observe that

$$\sum_i \int p(x) \log^+ \frac{p(x)}{q_i(x)} < \infty. \tag{4.3}$$

Also observe the following lower bound.

$$\begin{aligned}
\log \frac{p(x)}{\sum_{i=1}^{n} t_i q_i(x)} &\geq \log \frac{p(x)}{\max_i q_i(x)} \\
&= \min_i \log \frac{p(x)}{q_i(x)} \\
&\geq -\max_i \log^- \frac{p(x)}{q_i(x)} \\
&\geq -\sum_i \log^- \frac{p(x)}{q_i(x)}.
\end{aligned}$$

And

$$\sum_i \int p(x) \log^- \frac{p(x)}{q_i(x)} < \infty. \tag{4.4}$$

Consequently by the Dominated Convergence Theorem, we have

$$\begin{aligned}
\lim_{t_i \to a_i} D_t &= \lim_{t_i \to a_i} \int p \log \frac{p}{\sum_{i=1}^{n} t_i q_i} \\
&= \int p \lim_{t_i \to a_i} \log \frac{p}{\sum_{i=1}^{n} t_i q_i} \\
&= \int p \log \frac{p}{\sum_{i=1}^{n} a_i q_i}
\end{aligned}$$

$$= D_a.$$

□

Now we can establish the existence lemma.

LEMMA 4.3 *Assume $D(p||\mathcal{C})$ is finite. Then there exists a q^* such that $D(p||q^*) = D(p||\mathcal{C})$. And there exists a sequence $\tilde{q}_n \in \mathcal{C}$, such that $\tilde{q}_n$ converges to the q^* in the sense that:*

$$\int p|\log(\tilde{q}_n) - \log(q^*)| \to 0.$$

We also define the "information closure" of a set is the set that includes all $L^1(p)$ limits.

PROOF Consider a sequence of $q_n \in \mathcal{C}$ such that $D(p||q_n) \to D(p||\mathcal{C})$. Assume $D(p||q_n)$ finite for all n. Consider $q_1, q_2, ..., q_n$. Let $\mathcal{C}_n$ be the convex hull of $q_1, q_2, ..., q_n$. Elements in $\mathcal{C}_n$ can be written as $q_{\underline{t}}$ as in last lemma.

Define $D(\underline{t})$ as in the last lemma. Because it's a continuous function on a compact simplex, there exists a global minimum and this minimum is achieved by an element in $\mathcal{C}_n$. We call it $\tilde{q}_n$. By Lemma 4.1, we have for all $q \in \mathcal{C}_n$,

$$\int p\frac{q}{\tilde{q}_n} \leq 1$$

since $\tilde{q}_n$ is a projection of p on $\mathcal{C}_n$.

By such a construction, we find a sequence of $\tilde{Q}_n$ in $\mathcal{C}$ with properties

1. $D(p||\tilde{q}_n)$ is a non-increasing sequence converging to $D(p||\mathcal{C})$,

2. $\int p\frac{\tilde{q}_n}{\tilde{q}_m} \leq 1$ for all $n \leq m$.

Now we want to show $\log(\tilde{q}_m)$ is a Cauchy sequence in $L^1(P)$.

$$
\begin{aligned}
D(p\|\tilde{q}_n) - D(p\|\tilde{q}_m) &= \int p \log \frac{p}{p\tilde{q}_n/\tilde{q}_m} \\
&= \int p \log \frac{p}{\frac{p\tilde{q}_n}{\tilde{q}_m}/c_{m,n}} + \log \frac{1}{c_{m,n}} \qquad (4.5)
\end{aligned}
$$

where $c_{m,n} = \int \frac{p\tilde{q}_n}{\tilde{q}_m}$. Here $c_{m.n} \leq 1$ by Property 2. So both terms in 4.5 are non-negative.

Now let $n \leq m$ and $n, m \to \infty$, we have $D(P\|\tilde{Q}_n) - D(P\|\tilde{Q}_m) \to 0$ by Property 1. So both terms in 4.5 have to converge to zero as well. In particular, $D(P\|\frac{p\tilde{q}_n/\tilde{q}_m}{c_{m,n}}) \to 0$. Invoke the Pinsker-type inequality: (See Barron[1986] and Pinsker[1964] for proofs)

$$
\int p|\log(p) - \log(q)| \leq D(p\|q) + \sqrt{2D(p\|q)}.
$$

and we have

$$
\int p|\log \frac{p}{\frac{p\tilde{q}_n}{\tilde{q}_m}/c_{m,n}}| \to 0 \qquad (4.6)
$$

The other term in 4.5 goes to zero too, which yields

$$
\lim_{m,n\to\infty} c_{m,n} \to 1. \qquad (4.7)
$$

Therefore, combine equations (4.6) and (4.7),

$$
\lim_{m,n\to\infty} \int p|\log(\tilde{q}_n) - \log(\tilde{q}_m)| \to 0.
$$

So $\log(\tilde{q}_n)$ is a Cauchy sequence in $L^1(P)$. The fact that $L^1(P)$ is complete (see Lang[1993],pp 133) implies that $\log(\tilde{q}_n)$ converges to a $\log(q^*)$ in $L^1(p)$.

$$
\lim_{n\to\infty} \int p|\log(\tilde{q}_n) - \log(q^*)| \to 0. \qquad (4.8)
$$

Furthermore, a direct application of (4.8) yields

$$D(p||q^*) = \lim D(p||\tilde{q}_n) = D(p||\mathcal{C}).$$

□

REMARK Another consequence of 4.5 is

$$D(P||\tilde{Q}_n) - D(P||\tilde{Q}_m) \;\geq\; D(P||\frac{p\tilde{q}_n/\tilde{q}_m}{c_{m,n}}).$$

□

4.2.3 Properties of the Reversed I-Projection

LEMMA 4.4 *Let q^* be the limit identified for a sequence of q_n that achieves the* $\lim D(p||q_n) \to D(p||\mathcal{C})$. *Then*

$$\int p\frac{q}{q^*} \leq 1$$

for all $q \in \mathcal{C}$.

PROOF Expand $\mathcal{C}$ to $\mathcal{C}^* = \{\alpha q + (1-\alpha)q^* : q \in \mathcal{C}, 0 \leq \alpha \leq 1\}$.

First we show that $\log[(1-\alpha)e^z + \alpha]$ is a Lipschitz function.

$$\frac{d}{dz}\log[(1-\alpha)e^z + \alpha] = \frac{(1-\alpha)e^z}{(1-\alpha)e^z + \alpha} \leq 1.$$

It follows immediately that if $\int p|\log(q_n) - \log(q^*)| \to 0$, then $\int p|\log[(1-\alpha)q_n + \alpha q] - \log[(1-\alpha)q^* + \alpha q]| \to 0$.

This property implies that every elements of $\mathcal{C}^*$ can be approached by a sequence in $\mathcal{C}$ in K-L divergence. Then there are no elements in $\mathcal{C}^*$ that can achieve smaller

K-L distance from p than q^* because otherwise it will contradict the assumption that $\delta = \min_{q \in \mathcal{C}} D(p||q)$.

Therefore, $D(p||q^*)$ minimizes $D(p||\mathcal{C}^*)$. By Lemma 4.1, we have the desired property.

□

The reversed I-projection is in general a sub-probability density function. With a condition on f such that $f(x) > 0$ for a.e. λ on $\mathcal{X}$, q^* is a probability density function. In addition, it preserves the boundedness property of $q_k(x)$, where $q_k(x) \in \mathcal{C}$.

LEMMA 4.5 (Other Properties) *Let q^* be the limit identified for a sequence q_n that achieves $\lim D(p||q_n) \to D(p||\mathcal{C})$. Then q^* is a subprobability density function, that is, $\int q^*(x)\lambda(dx) \leq 1$.*

Furthermore, suppose for some ϕ in $\mathcal{C}$ with $\phi(x) > 0$ for a.e.[λ] in $\mathcal{X}$, we have $\log q(x)/\phi(x) \leq a$. Also suppose that $p(x) > 0$ a.e.[λ] in $\mathcal{X}$. Then q^* is a density function. In addition, boundedness of the density ratio q_n/ϕ implies boundedness of q^*/ϕ.

PROOF

We know that

$$\int p|\log(q_k) - \log(q^*)| \to 0.$$

Then $\log(q_k)$ goes to $\log(q^*)$ in P probability because of Chebyshev's inequality:

$$P\{|\log(q_k) - \log(q^*)| > \epsilon\} \leq \frac{E_P|\log(q_k) - \log(q^*)|}{\epsilon}.$$

Applying a continous function on a random variable that converges in probability, we still get convergence in probability (see, for instance, Durrett[1996] Chapter 1). So

$$exp(\log(q_k)) \to exp(\log(q^*))$$

in P probability since $exp(.)$ is a continuous function. Hence $q_k \to q^*$ in P probability.

Similarly

$$\int p|\log(q_k/p) - \log(q^*/p)| \to 0.$$

Note that we only deal with the set in which $f > 0$ in the P-integral. Thus $q_k/p \to q^*/p$ in P probability.

Applying Fatou's lemma on the sequence of random variables q_k/p, we have

$$\liminf \int p\frac{q_k}{p} \geq \int p\frac{q^*}{p} = \int q^*$$

since $q_k/p \to q^*/p$ in P probability (see also Durrett[1996] page 48 for a proof of this generalized Fatou's Lemma). Since $\int p\frac{q_k}{p} \leq 1$, we have $\int q^* \leq 1$.

Suppose $p(x) > 0$ a.e.$[\lambda]$ in $\mathcal{X}$. Take another density $\phi \in G$ with the properties $\phi > 0$ a.e. $[\lambda]$. Then we have $q_k/\phi \to q^*/\phi$ in P_ϕ probability through similar reasoning.

In addition, if we assume $e^{-a} \leq \frac{q_k}{\phi} \leq e^a$, we have

$$1 = P_\phi(q_k/\phi) \to P_\phi(q^*/\phi) = \int q^*$$

by the Bounded Convergence Theorem. So q^* is a probability density function.

Furthermore, consider the function $\frac{q_k}{\phi}I\{e^{-a} \leq \frac{q_k}{\phi} \leq e^a\}$. It also converges to $\frac{q^*}{\phi}I\{e^{-a} \leq \frac{q^*}{\phi} \leq e^a\}$ in P_ϕ probability. And it's bounded. So we have

$$1 = P_\phi(\frac{q_k}{\phi}I\{e^{-a} \leq \frac{q_k}{\phi} \leq e^a\}) \to P_\phi(\frac{q^*}{\phi}I\{e^{-a} \leq \frac{q^*}{\phi} \leq e^a\}) = \int q^* I\{e^{-a} \leq \frac{q^*}{\phi} \leq e^a\}.$$

So $e^{-a} \leq \frac{q^*}{\phi} \leq e^a$ a.e. $[q^*]$. □

REMARK The boundedness of q^* will be helpful for getting Kullback-Leibler risk bound. □

Chapter 5

Proofs of Main Results

In this chapter, we will prove the main results we presented in chapter 3. In section 1, we provide proofs of two inequalities and one iteration lemma. Both inequalities study upper bounds for log densities. The iteration lemma establishes that if a sequence of positive numbers satisfies the following iterative relationship:

$$D_k \leq (1-\alpha)D_{k-1} + \alpha^2 B, \tag{5.1}$$

for all $0 \leq \alpha \leq 1$, where B is a bounded constant, and $D_1 \leq 4B$, then we have a converging $D_k \to 0$ at a particular rate we will identify.

In section 2, we prove that a greedy algorithm generates a sequence of Kullback-Leibler numbers satisfying the iterative formula 5.1 and the condition on D_1.

In sections 3, 4 and 5, we show that the approximation and nearly maximum likelihood theorems are the results of the lemma in section 2.

In sections 6 and 7, we prove the risk bounds for fixed k and estimated k using MDL.

5.1 Preliminary Lemmas

We need a few preliminary lemmas for the proof of the main theorems.

5.1.1 An upper bound for $-\log(r)$

The first lemma establishes a tight upper bound for $-\log(r)$ at a range $r \geq r_0$.

LEMMA 5.1 *For $r \geq r_0 > 0$,*

$$-\log(r) \leq -(r-1) + \zeta(r_0)(r-1)^2. \tag{5.2}$$

where $\zeta(r) = \frac{-\log(r)-(r-1)}{(r-1)^2}$

REMARK We could just expand $-\log r$ at $r = 1$ to the second order term using a Taylor expansion,

$$-\log r \simeq -(r-1) + \frac{1}{2}(r-1)^2.$$

Is this an upper bound? Clearly it's not. Indeed, $-\log r \leq -(r-1) + \frac{1}{2}(r-1)^2$ only when $r \geq 1$. □

The above remark provides a motivation for the bound. Instead of Taylor expansion, let's consider the following quadratic bound for $-\log(r)$:

$$-\log r \leq -(r-1) + A(r_0)(r-1)^2.$$

We want to find $A(r_0)$ such that the inequality is true for $r \geq r_0$. And we want the constant $A(r_0)$ to be as small as possible.

The following lemma claims a general result on such quadratic bounds.

LEMMA 5.2 *Consider the function* $-\log(r)$ *and a quadratic functionh*$(r) = ar^2+br+c$. *Choose* a, b, c *such that following criteria are true.*

- $h(r)$ *match* $-\log(r)$ *at two arbitrary positive points* r_1, r_2. *Without loss of generality, assume* $r_1 < r_2$;
- *The two functions have the same first derivative at the bigger point* r_2.

Then we have the following inequality at the range $r > r_1$.

$$-\log(r) \leq h(r).$$

PROOF First, $h(r) \geq -\log(r)$ for $r \geq r_2$ because the first derivative of $h(r)$ is always larger than $-\log(r)$ for $r \geq r_2$. Then, the second derivative of $h(r)$ at r_2 is larger than $-\log(r)$. So $h(r)$ is larger than $-\log(r)$ at the neighborhood of r_2. But $-\log(r)$ has an increasing second derivative when r goes to zero. So eventually the value of $h(r)$ will go below $-\log(r)$ when r goes to zero. The two functions will intersect at some point, which is designed to be r_1. Therefore, $h(r)$ stays above $-\log(r)$ between r_1 and r_2 also. □

After doing calculus and algebra, a function defined as follows prove to be the coefficient A. And it's a strictly decreasing function of r.

LEMMA 5.3 *Define* $\zeta(r) = \frac{r-1-\ln r}{(r-1)^2}$. *Then* $\zeta(r)$ *is a strictly-decreasing function of* r *for* $r > 0$.

PROOF For $r \neq 0$ and $r \neq 1$, we can take the derivative of $\zeta(r)$. Denote it $\zeta'(r)$. We have

$$\zeta'(r) = -\frac{1}{(r-1)^2}(1 + \frac{1}{r} - \frac{2\ln r}{r-1})$$

$$
\begin{aligned}
&= -\frac{1}{r(r-1)^3}[r(r-1)+r-1-2r\ln r] \\
&= -\frac{1}{r(r-1)^3}(r^2-1-2r\ln r) \\
&= -\frac{1}{r(r-1)^3}G(r).
\end{aligned}
$$

We denote $G(r) = r^2 - 1 - 2r\ln r$. Let's take the derivative of $G(r)$. We obtain

$$
\begin{aligned}
\frac{dG(r)}{dr} &= 2r - 2\ln r - 2 \\
&= 2(r - 1 - \ln r) \geq 0.
\end{aligned}
$$

The equality holds at and only at $r = 1$.

Therefore, we know that $G(r)$ is a strictly increasing function of r (except at $r = 1$). Furthermore, we know $G(1) = 1 - 1 - 0 = 0$. We can then conclude that $G(r) < 0$ for $0 < r < 1$ and $G(r) > 0$ for $r > 1$. We can further conclude that $\zeta'(r) < 0$ for all $r > 0$ except at $r = 1$.

We can find the limit values of $\zeta(r)$ and $\zeta'(r)$ at $r = 1$ by a Taylor expansion of the $\ln r$ at $r = 1$ inside $\zeta(r)$. We obtain

$$
\begin{aligned}
\zeta(r) &= \frac{r-1-\ln r}{(r-1)^2} \\
&= \frac{r-1-[(r-1)-\frac{(r-1)^2}{2}+\frac{(r-1)^3}{3}+o(r-1)^3]}{(r-1)^2} \\
&= \frac{1}{2}-\frac{1}{3}(r-1)+o(r-1),
\end{aligned}
$$

where $\zeta(1) = \frac{1}{2}$ and $\zeta'(1) = -\frac{1}{3}$.

Now we have shown that $\zeta'(r) < 0$ for all $r > 0$. It's easy to show that

$$\lim_{r \to 0} \zeta(r) \longrightarrow +\infty \tag{5.3}$$

and

$$\lim_{r \to +\infty} \zeta(r) \longrightarrow 0. \tag{5.4}$$

So $\zeta(r)$ strictly decreases from $+\infty$ to 0 for $r > 0$.

□

REMARK We can apply above techniques for other non-polynomial monotone functions, e.g. $\frac{1}{r}$ and get quadratic upper bounds. For instance, we have that,

$$\frac{\frac{1}{r} + r - 2}{(r-1)^2}$$

is also monotone decreasing from $+\infty$ to 0. So we can find a quadratic upper bound for $\frac{1}{r}$. □

We also have a simple upper bound for $\zeta(r)$.

LEMMA 5.4

$$\zeta(r) \leq 1/2 + \log^{-}(r) \tag{5.5}$$

PROOF

Note that $\lim_{r \to 1} \zeta(r) = 1/2$. So $\zeta(r) \leq 1/2$ for $r \geq 1$ since $\zeta(r)$ is a monotonely decreasing function of r. For $r < 1$, $\log^{-}(r) = \log 1/r$. Thus ζ is like $-\log(r)$ when r small, and it is like $1/r$ when r is large. Some calculus and algebra will show that $1/2 + \log^{-}(r)$ provides a bound for small $r < 1$. This is also conveyed in the graph 5.1, which compares $\zeta(r)$ and $1/2 + \log^{-}(r)$ in the range (0,1].

□

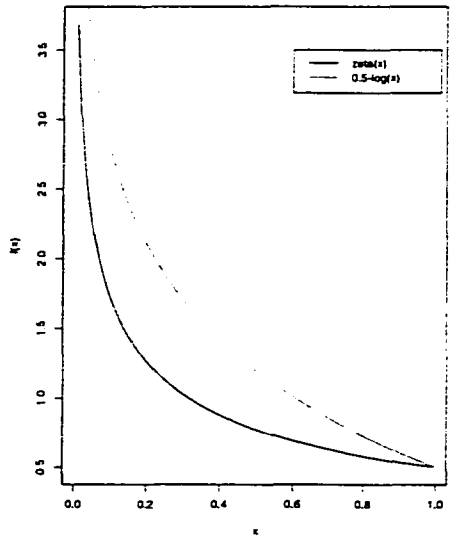


Figure 5.1: Compare zeta(r) with $-\log^{+}(r) + 1/2$

5.1.2 A Key Inequality

This inequality plays a key role in simplifying upper bounds we will use.

LEMMA 5.5 *For all* $r \geq 0$,

$$2[\frac{-\log r + r - 1}{(r-1)}] \leq \log r. \tag{5.6}$$

PROOF It's equivalent to prove for all $r \geq 0$, that

$$2[\frac{-\log r + r - 1}{(1-r)}] + \log r \geq 0.$$

Denote the left-hand side as $\zeta(r)$. It's easy to show by L'Hopital's Law that $\lim_{r \to 1} \zeta(r) = 0$.

For $0 \leq r < 1$, it's equivalent to show that $G(r) \geq 0$ where

$$G(r) = (1-r)\zeta(r) = 2(-\log r + r - 1) + (1-r)\log r.$$

We can show that $G'(r) \leq 0$. Indeed,

$$\begin{aligned} G'(r) &= \log\frac{1}{r} + 1 - \frac{1}{r} \\ &\leq (\frac{1}{r} - 1) + 1 - \frac{1}{r} = 0. \end{aligned}$$

Therefore, combining the fact that $G(1) = 0$ and $G'(r) \leq 0$ we have $G(r) \geq 0$ for $0 \leq r < 1$.

For $r \geq 1$, it's equivalent to show $G(r) \leq 0$. Again, combining the fact $G'(r) \leq 0$ and $G(1) = 0$, we have $G(r) \leq 0$ for $r > 1$.

Therefore, we have shown that $\zeta(r) \geq 0$ for all $r > 0$. □

5.1.3 Iteration Lemma

The following lemma establishes an iterative condition for a sequence to converge. It also establishes the rate of convergence.

LEMMA 5.6 (Iteration) *Suppose we have a sequence of positive numbers $D_k, k \geq 1$. D_k satisfy a recursive relation*

$$D_k \leq (1-\alpha_k)D_{k-1} + \alpha_k^2 B_k$$

for all $0 < \alpha_k < 1$, $k \geq 2$. And $B_k, k \geq 1$ is a non-decreasing sequence of bounded positive

numbers. Also suppose $D_1 \leq 4B_1$. *Then we have*

$$D_k \leq \frac{4B_k}{k}$$

for all $k \geq 1$.

PROOF First show the inequality for D_2. We have

$$\begin{aligned} D_2 &\leq (1-\alpha_2)D_1 + \alpha_2^2 B_2 \\ &\overset{D_1 \leq 4B_1}{\leq} (1-\alpha_2)4B_1 + \alpha_2^2 B_2 \\ &\overset{B_1 \leq B_2}{\leq} (1-\alpha_2)4B_2 + \alpha_2^2 B_2 \\ &= (\alpha_2 - 2)^2 B_2. \end{aligned}$$

Choose $\alpha_2 = 2/3$. Then

$$D_2 \leq 2B_2 = \frac{4B_2}{2}.$$

Suppose $D_{k-1} \leq \frac{4B_{k-1}}{k-1}$ for some $k \geq 3$. Then

$$D_k \leq (1-\alpha_k)\frac{4B_{k-1}}{k-1} + \alpha_k^2 B_k.$$

Choose $\alpha_k = \frac{2}{k}$. Then,

$$\begin{aligned} D_k &\leq (1-\frac{2}{k})\frac{4B_{k-1}}{k-1} + \frac{4}{k^2}B_k \\ &\leq (1-\frac{2}{k}+\frac{1}{k^2}-\frac{1}{k^2})\frac{4B_k}{k-1} + \frac{4B_k}{k^2} \\ &= \left[\left(\frac{k-1}{k}\right)^2 - \frac{1}{k^2}\right]\frac{4B_k}{k-1} + \frac{4B_k}{k^2} \\ &\leq \frac{k-1}{k^2}4B_k + \frac{1}{k^2}4B_k = \frac{4B_k}{k}. \end{aligned}$$

□

REMARK Under such a scheme, $1 - \alpha \geq 1/3$ because the largest α is $2/3$ when $k = 2, 3$. When $k > 3$, α will get smaller. This fact becomes useful in simplifying the upper bounds we will prove. □

Another related lemma is useful also.

LEMMA 5.7 (Iteration with Additional Error) *Suppose we have a sequence of positive numbers D_k, $k \geq 1$, and a fixed positive number ϵ, which satisfy*

$$D_k \leq (1 - \alpha_k)D_{k-1} + \alpha_k^2 B_k + \epsilon$$

for all $0 < \alpha_k < 1$, $k \geq 2$. And $B_k, k \geq 1$ is a non-decreasing sequence of bounded positive numbers. Also suppose $D_1 \leq 4B_1$. Then we have

$$D_k \leq \frac{4B_k}{k} + k\epsilon$$

for all $k \geq 1$.

PROOF

The proof is structurally similar to the proof of previous lemma. The first step is easy to check since we are only adding larger positive number to the bound. The induction step is checked as follows.

Suppose $D_{k-1} \leq \frac{4B_{k-1}}{k-1} + (k-1)\epsilon$ for some $k \geq 3$. Then

$$D_k \leq (1 - \alpha_k)\frac{4B_{k-1}}{k-1} + \alpha_k^2 B_k + (1 - \alpha_k)k\epsilon + \epsilon.$$

We still choose $\alpha_k = \frac{2}{k}$. Then consider the parts with ϵ. We have

$$\begin{aligned}(1-\alpha_k)k\epsilon + \epsilon &= (1-\frac{2}{k})(k-1)\epsilon + \epsilon \\ &= (k-2+2/k)\epsilon \\ &< k\epsilon.\end{aligned}$$

So

$$D_k \leq \frac{4B_k}{k} + k\epsilon.$$

□

5.2 A Generalized Greedy Algorithm

Proofs of approximation bounds and nearly maximum likelihood are applications of the following two lemmas. We also write $(1-\alpha)$ as $\bar{\alpha}$ from now on.

LEMMA 5.8 (Pointwise Iterative Inequality) *Let $f_k = \bar{\alpha} f_{k-1} + \alpha\phi$, where f_{k-1} and ϕ are probability densities in $\mathcal{C}$ and α is between 0 and 1. Let g be another probability density in $\mathcal{C}$. We have the following iterative relationship on the log ratio of densities,*

$$-\log\frac{f_k}{g} \leq -\log\bar{\alpha}\frac{f_{k-1}}{g} + \alpha\frac{\phi}{g}\log\frac{\bar{\alpha}f_{k-1}}{g} - \alpha\frac{\phi}{g} + \alpha^2\frac{\phi^2}{g^2}\gamma. \tag{5.7}$$

PROOF Note that right side is a quadratic of $\alpha\phi$. Since $\alpha\phi \geq 0$, we have

$$\frac{f_k}{g} \geq \bar{\alpha}\frac{f_{k-1}}{g}.$$

Apply LEMMA 5.3 to the pair, we have an inequality $\zeta(\frac{f_k}{g}) \leq \zeta(\bar{\alpha}\frac{f_{k-1}}{g})$ because ζ is a strictly decreasing function. Expand $\zeta(.)$. We have an upper bound for log ratio of

densities f_k and g,

$$-\log\frac{f_k}{g} \le -(\frac{f_k}{g}-1) + \frac{-\log\bar{\alpha}\frac{f_{k-1}}{g} + \bar{\alpha}\frac{f_{k-1}}{g} - 1}{(\bar{\alpha}\frac{f_{k-1}}{g}-1)^2}(\frac{f_k}{g}-1)^2. \tag{5.8}$$

Now we replace f_k by $\bar{\alpha}f_{k-1} + \alpha\phi$ in the right side.

First expand the quadratic $(\frac{f_k}{g}-1)^2 = (\bar{\alpha}\frac{f_{k-1}}{g} - 1 + \alpha\frac{\phi}{g})^2$ into a sum of the following three parts: $(\bar{\alpha}\frac{f_{k-1}}{g}-1)^2$, $\alpha^2(\frac{\phi}{g})^2$, and $2(\bar{\alpha}\frac{f_{k-1}}{g}-1)\alpha\frac{\phi}{g}$.

Then plug them back to the right side of 5.8. After some re-arrangement of terms and algebra, the bound turns into a sum of the following four parts:

1. $-(\bar{\alpha}\frac{f_{k-1}}{g} - 1 + \alpha\frac{\phi}{g})$,
2. $-\log\bar{\alpha}\frac{f_{k-1}}{g} + \bar{\alpha}\frac{f_{k-1}}{g} - 1$,
3. $\alpha^2(\frac{\phi}{g})^2\zeta(\bar{\alpha}\frac{f_{k-1}}{g})$,
4. and $2\alpha\frac{\phi}{g}\frac{-\log\bar{\alpha}\frac{f_{k-1}}{g}+\bar{\alpha}\frac{f_{k-1}}{g}-1}{(\bar{\alpha}\frac{f_{k-1}}{g}-1)}$.

Notice that the term $\bar{\alpha}\frac{f_{k-1}}{g} - 1$ appears in both expression 1 and 2 with opposite signs. So they cancel out. In the expression 3, we have a term $\zeta(\bar{\alpha}\frac{f_{k-1}}{g})$. Use 5.4 to get an upper bound for ζ,

$$\zeta(\bar{\alpha}\frac{f_{k-1}}{g}) \le 1/2 + \log^-\bar{\alpha}\frac{f_{k-1}}{g}$$

Assume $\bar{\alpha} \ge 1/3$. We have

$$\log^-\bar{\alpha}\frac{f_{k-1}}{g} \le \log^-\frac{1}{3}\frac{f_{k-1}}{g}$$

We lower bound $\min_x \frac{f_{k-1}}{g}$ as follows,

$$\begin{aligned}\min_x\frac{f_{k-1}}{g} &= \min_x\frac{\sum p_j\phi_j(x)}{\int\phi_b(x)P(db)} \\ &\ge \min_x\frac{\min_{\phi_b\in G}\phi_b}{\max_{\phi_b\in G}\phi_b}\end{aligned}$$

$$\geq e^{-a}.$$

Recall the definition of a and γ in section 3.2. Now we have

$$\zeta(\bar{\alpha}\frac{f_{k-1}}{g}) \leq 1/2 + \log 3 + a = \log 3\sqrt{e} + a = \gamma/4$$

Applying LEMMA 5.5 in the expression 4, we get a simple upper bound of $\log \bar{\alpha}\frac{f_{k-1}}{g}$ for the term $(2\frac{-\log \bar{\alpha}\frac{f_{k-1}}{g} + \bar{\alpha}\frac{f_{k-1}}{g} - 1}{(\bar{\alpha}\frac{f_{k-1}}{g} - 1)})$.

Now, the upper bound simplifies to a sum of the following four expressions:

1. $-\alpha\frac{\phi}{g}$,
2. $-\log \bar{\alpha}\frac{f_{k-1}}{g}$,
3. $\alpha^2(\frac{\phi}{g})^2\gamma/4$,
4. $\alpha\frac{\phi}{g}\log \bar{\alpha}\frac{f_{k-1}}{g}$.

We get the desired iterative relationship.

□

By taking expectation on both sides of above inequality, we can get an iterative relationship of the expected log density ratio, from which approximation and likelihood bounds fall out right away.

LEMMA 5.9 (Iterative Inequality) *Let $f_1 = \phi_{b_1}$. Suppose $f_k = \bar{\alpha}f_{k-1} + \alpha\phi_b, k = 2, 3...$, where ϕ_b is in G and α is between 0 and 1. Let g be any probability density in C, i.e. $g = \int \phi_b P(db)$ for some P. Let F be a probability measure on X space. We choose b_1 to minimize $E_F(-\log \frac{\phi_{b_1}(X)}{g(X)})$. We then choose b_k to minimize $E_F(-\log \frac{f_k(X)}{g(X)})$. Then we have for $k \geq 2$,*

$$E_F \log \frac{g}{f_k} \leq (1-\alpha)E_F \log \frac{g}{f_{k-1}} + \alpha^2 c_{F,P}^2 \gamma/4. \tag{5.9}$$

For $k = 1$, take $\alpha = 1$, then 5.9 yields $E_F(-\log \frac{\phi_{b_1}(X)}{g(X)}) \leq c_{F,P}^2 \gamma/4$. Consequently, from LEMMA 5.6, we have

$$E_F \log \frac{g}{f_k} \leq \frac{c_{F,P}^2 \gamma}{k} \tag{5.10}$$

PROOF Take expectation with respect to F at both sides of inequality 5.7.

$$E_F \log \frac{g}{f_k} \leq \underbrace{E_F(-\log \bar{\alpha} \frac{f_{k-1}}{g}) + E_F(\alpha \frac{\phi_b}{g} \log \frac{\bar{\alpha} f_{k-1}}{g})}_{A(b)} \tag{5.11}$$

$$\underbrace{+ E_F(-\alpha \frac{\phi_b}{g}) + E_F(\alpha^2 \frac{\phi_b^2}{g^2} \gamma/4)}_{B(b)} \tag{5.12}$$

Treat the right side of the inequality as a function of b. And we write a shorthand for it as $\pi(b)$. The greedy algorithm chooses b to minimize $E_F \log \frac{g}{f_k}$. Let b^* be the minimizer.

Let $L(b) = E_F \log \frac{g(X)}{(1-\alpha) f_{k-1}(X) + \alpha \phi_b(X)}$ denote the left side. Let $\pi(b)$ denote the right side. We have

$$\begin{aligned} L(b^*) &= \min_b L(b) \\ &\leq \int L(b) P(db) \\ &\leq \int \pi(b) P(db) \\ &= \int A(b) P(db) + \int B(b) P(db) \end{aligned}$$

where $A(b)$ and $B(b)$ are the two parts identified in 5.11 and 5.12. We look at those two parts separately.

$$\int A(b) P(db) = E_F(-\log \bar{\alpha} \frac{f_{k-1}}{g}) + \int E_F(\alpha \frac{\phi_b(X)}{g(X)} \log \frac{\bar{\alpha} f_{k-1}(X)}{g(X)}) P(db)$$

$$= E_F \log \frac{g}{f_{k-1}} - \log(\bar{\alpha}) - \alpha E_F \frac{\int \phi_b P(db)}{g} \log \frac{g}{\bar{\alpha} f_{k-1}}. \quad (5.13)$$

Observing that $\frac{\phi_b(x)}{g(x)} |\log \frac{g}{\bar{\alpha} f_{k-1}}|$ is integrable with respect to $P(db)F(dx)$ (since the log factor is bounded and $\frac{\phi_b(x)}{g(x)}$ integrates to 1 with respect to P), we see that the exchange of integration with respect to F and P is justified by Fubini's theorem.

Since $g = \int \phi_b P(db)$, the formula 5.13 is simplified to

$$(1-\alpha) E_F \log \frac{g}{f_{k-1}} - \log(\bar{\alpha}) + \alpha \log(1-\alpha). \quad (5.14)$$

Applying Fubini again, the part $\int B(b) P(db)$ yields

$$\alpha^2 E_F (\frac{\int \phi_b^2(x) P(db)}{g(x)^2}) \gamma/4 - \alpha \quad (5.15)$$

which is $\alpha^2 c_{F,P}^2 \gamma/4 - \alpha$.

Combining the 5.14 and 5.15, we have

$$E_F \log \frac{g}{f_k} \leq (1-\alpha) E_F \log \frac{g}{f_{k-1}} + \alpha^2 c_{F,P}^2 \gamma/4 + \alpha \log(1-\alpha) - \alpha - \log(1-\alpha). \quad (5.16)$$

Use the following fact:

$$\begin{aligned} & -\log(1-\alpha) - \alpha + \alpha \log(1-\alpha) \\ = \; & (1-\alpha) \log \frac{1}{1-\alpha} - \alpha \\ = \; & (1-\alpha) \log(1 + \frac{\alpha}{1-\alpha}) - \alpha \\ \leq \; & (1-\alpha) \frac{\alpha}{(1-\alpha)} - \alpha = 0. \end{aligned}$$

Now we have the desired inequality 5.9.

□

5.3 Approximation and Nearly Maximum Likelihood Bounds

Applying different F, g_P in the LEMMA 5.9, we get following three important results.

5.3.1 Approximation Error Bound When f is in $\mathcal{C}$

When f is an element in $\mathcal{C}$, f can be represented as $\int \phi_b P(db)$. Then replace g by f in LEMMA 5.9 and use f as the density of F. We have the desired approximation error bound for the greedy approximation procedure.

We choose ϕ and α to minimize $E_F(-\log \frac{f_k}{f})$. Then we have

$$E_F \log \frac{f}{f_k} \leq (1-\alpha) E_F \log \frac{f}{f_{k-1}} + \alpha^2 c_{F,P}^2 \gamma/4. \tag{5.17}$$

And consequently by LEMMA 5.6

$$D(f\|f_k) = E_F \log \frac{f}{f_k} \leq \frac{c_f^2 \gamma}{k}. \tag{5.18}$$

5.3.2 Approximation Error Bound When f is not in $\mathcal{C}$

When f is not an element in $\mathcal{C}$, we show that $D(f\|f_k)$ becomes not worse than $D(f\|g)$ for any $g \in \mathcal{C}$. For any $g_P = \int \phi_b P(db)$, apply LEMMA 5.9, we have

$$E_F \log \frac{g_P}{f_k} \leq \frac{c_{F,P}^2 \gamma}{k}. \tag{5.19}$$

Note that if F has a density f,

$$E_F \log \frac{g_P}{f_k} = D(f\|f_k) - D(f\|g_P). \tag{5.20}$$

So we have that for any $g_P \in \mathcal{C}$

$$D(f\|f_k) \leq D(f\|g_P) + \frac{c_{F,P}^2 \gamma}{k}. \tag{5.21}$$

5.3.3 Nearly Maximum Likelihood Bound

When we use empirical measure F_n in the expectation $E_F \log \frac{g_P}{f_k}$, we have a nearly Maximum Likelihood Bound.

If we choose ϕ and α to maximize $\frac{1}{n}\sum_{i=1}^n \log f_k(x_i)$, we have

$$\frac{1}{n}\sum_{i=1}^n \log \frac{\hat{f}_k}{g_P} \geq (1-\alpha)\frac{1}{n}\sum_{i=1}^n \log \frac{\hat{f}_{k-1}}{g_P} - \alpha^2 c_{F_n,P}^2 \gamma/4. \tag{5.22}$$

So we have for any $g_P \in \mathcal{C}$ and $g_P = \int \phi_b P(db)$,

$$\frac{1}{n}\sum_{i=1}^n \log \frac{\hat{f}_k}{g_P} \geq -\frac{c_{F_n,P}^2 \gamma}{k}. \tag{5.23}$$

Rearrange the terms to obtain

$$\frac{1}{n}\sum_{i=1}^n \log \hat{f}_k \geq \frac{1}{n}\sum_{i=1}^n \log g - \frac{c_{F_n,P}^2 \gamma}{k}. \tag{5.24}$$

Again, if we obtain the maximum likelihood estimate on quantized Θ_ϵ instead of Θ at each k, we introduce an extra quantization error. Now let $\hat{f}_{k,\epsilon}$ be the estimate on the quantized parameter space. Assume the smoothness condition as in Theorem 3.2. We

obtain the following result:

$$\frac{1}{n}\sum_{i=1}^{n}\log \hat{f}_{k,\epsilon} \geq \frac{1}{n}\sum_{i=1}^{n}\log g_P - \frac{4c^2_{F_n,P}\gamma}{k} - kdB\epsilon. \tag{5.25}$$

PROOF We just need to show that the drop in likelihood is bounded by $kdB\epsilon$ after we use a quantized parameter space.

We will show that at each step the likelihood drop at most $dB\epsilon$ as the result of using quantized parameter.

At kth step, let the MLE on Θ be $\hat{b}$ and the MLE on quantized parameter space Θ_ϵ be $\hat{b}_\epsilon$. Then $\sum_{j=1}^{d}|\hat{b}_j - \hat{b}_{\epsilon,j}| \leq d\epsilon$. By the smoothness condition,

$$\sup_x |\log(\phi_{\hat{b}}) - \log(\phi_{\hat{b}_\epsilon})| \leq Bd\epsilon.$$

Let $\hat{f}_k = (1-\alpha)\hat{f}_{k-1}+\alpha\phi_{\hat{b}}$. Let $\hat{f}^*_{k,\epsilon} = (1-\alpha)\hat{f}_{k-1}+\alpha\phi_{\hat{b}_\epsilon}$. Observe that $\log[(1-\alpha)e^z+\alpha]$ is a Lipschitz function because

$$\frac{d}{dz}\log[(1-\alpha)e^z+\alpha] = \frac{(1-\alpha)e^z}{(1-\alpha)e^z+\alpha} \leq 1.$$

Then

$$\begin{aligned}
& |\log \hat{f}_k(x) - \log \hat{f}^*_{k,\epsilon}(x)| \\
= \; & |\log[(1-\alpha)\hat{f}_{k-1}(x)+\alpha\phi_{\hat{b}}] - \log[(1-\alpha)\hat{f}_{k-1}(x)+\alpha\phi_{\hat{b}_\epsilon}]| \\
= \; & |\log[(1-\alpha)+\alpha e^{\log\frac{\phi_{\hat{b}}}{\hat{f}_{k-1}}}] - \log[(1-\alpha)+\alpha e^{\log\frac{\phi_{\hat{b}_\epsilon}}{\hat{f}_{k-1}}}] \\
\leq \; & |\log\phi_{\hat{b}_\epsilon} - \log\phi_{\hat{b}}|.
\end{aligned}$$

So

$$\sup_x |\log \hat{f}_k(x) - \log \hat{f}^*_{k,\epsilon}(x)| \leq Bd\epsilon.$$

So the likelihood difference is

$$\frac{1}{n}\sum(\log \hat{f}^*_{k,\epsilon} - \log \hat{f}_k) \geq -Bd\epsilon.$$

The difference is accumulated through k steps. Moreover from Lemma 5.7 "Iteration with Errors", we conclude that

$$\frac{1}{n}\sum_{i=1}^{n} \log \hat{f}_{k,\epsilon} \geq \frac{1}{n}\sum_{i=1}^{n} \log g_P - \frac{4c^2_{F_n,P}\gamma}{k} - kdB\epsilon \tag{5.26}$$

as desired. □

5.4 Risk Bounds For Fixed k

The proofs for risk bounds are organized as follows. We will first establish a Hellinger risk bound, which has a non-diminishing term of approximation error. Then we introduce a so-called approximation-adjusted Hellinger loss between f and g, defined as $\int f(\sqrt{\frac{g}{f^*}}-1)^2$, where f^* is the reversed information projection of f on $\mathcal{C}$. It turns out that by using this loss, we can get rid of the non-diminishing approximation error. In addition, after exploring the relationship between Hellinger distance and Kullback-Leibler divergence, we establish that the K-L risk bounds are just corollaries of Hellinger risk bounds.

5.4.1 Hellinger Risk Bound

We will consider estimates $\hat{f}_k$. We will also write it as $f(x|\hat{\theta}_k)$, where $\hat{\theta}_k \in \Theta_{k,\epsilon}$ is the parameters of k-component mixture density generated by the iterative maximum likelihood algorithm operating on the quantized parameter space. We will show that the assumption of quantized parameter doesn't affect the risk bound much because of smoothness condition we put on $\log(\phi_b)$. Meanwhile the assumption simplifies the proof.

Define the Affinity between two densities as

$$A(f,g) = \int \sqrt{f(x)g(x)}\lambda(dx).$$

The Affinity between product densities $f^n(x^n) = \prod f(x_i)$ and $g^n(x^n) = \prod g(x_i)$ is the product of affinities, that is,

$$\begin{aligned} A_n(f,g) &= \int \sqrt{f^n g^n} \\ &= \int \sqrt{f(x^n)g(x^n)}\lambda^n(dx^n) \\ &= (\int \sqrt{f(x)g(x)}\lambda(dx))^n \end{aligned}$$

$$= (\int \sqrt{fg})^n = (A(f,g))^n.$$

Here, $x^n = (x_1, ..., x_n)$ and the x_i's are independent.

First we establish the following inequality between the Hellinger distance and the Affinity.

LEMMA 5.10 (Hellinger Affinity Bound) *For two densities f and g,*

$$H^2(f,g) \leq 2 \log \frac{1}{A(f,g)}. \tag{5.27}$$

PROOF

$$\begin{aligned} H^2(f,g) &= \int (\sqrt{f(x)} - \sqrt{g(x)})^2 \lambda(dx) \\ &= \int f + g - 2\sqrt{fg} \\ &= 2 - 2\int \sqrt{fg} \\ &\leq -2 \log(\int \sqrt{fg}). \end{aligned}$$

The last inequality uses the fact that $\log(x) \leq x - 1$. □

For n-fold Hellinger distance, we have a n-fold Affinity bound,

$$\begin{aligned} nH^2(f,g) &\leq 2n \log \frac{1}{\int \sqrt{fg}} \\ &= 2 \log \frac{1}{(\int \sqrt{fg})^n} \\ &= 2 \log \frac{1}{A_n(f,g)}. \end{aligned}$$

We also establish a Hellinger distance bound for K-L distance in the following lemma. This lemma will prove handy when we try to get a K-L risk bound.

LEMMA 5.11 (Hellinger bound for K-L) *Let p, q be two probability densities. If $p(x)/q(x) \leq V$ for all x, then*

$$D(p\|q) \leq \Phi(V)H^2(p,q) \tag{5.28}$$

where $\Phi(V) = \frac{V\log V+1-V}{(\sqrt{V}-1)^2} \leq 2+\log(V)$.

PROOF See Yang and Barron [1998]. □

The following theorem bounds Hellinger risk in terms of Kullback approximation error in the context of general maximum likelihood. The essense of the technique can be applied to our iterative MLE and get bounds.

THEOREM 5.1 (Hellinger Risk Bound For MLE) *Let $X^n = (X_1, X_2, ..., X_n)$ be i.i.d. data with density f. Let $\mathcal{G}$ be a finite set of densities and let $\hat{f}$ be the choice in this set that maximize the likelihood $g(X^n) = \prod_{i=1}^n g(X_i)$. Then*

$$EH^2(f,\hat{f}) \leq \min_{g\in\mathcal{G}} D(f\|g) + \frac{2\log card(\mathcal{G})}{n}. \tag{5.29}$$

If also $f(x)/g(x) \leq V$ for all $g \in \mathcal{G}$, then

$$\frac{1}{\Phi(V)}D(f\|\hat{f}) \leq \min_{g\in\mathcal{G}} D(f\|g) + \frac{2\log card(\mathcal{G})}{n}. \tag{5.30}$$

PROOF We get an upper bound for the n-fold Hellinger distance between f and $\hat{f}$.

$$nH^2(f,\hat{f}) \leq 2\log\frac{1}{A_n(f,\hat{f})}. \tag{5.31}$$

Rewrite the right side of 5.31 as

$$\underbrace{2\log\left[\frac{1}{A_n(f,\hat{f})}\left(\frac{\hat{f}(X^n)}{f(X^n)}\right)^{1/2}\frac{1}{card(\mathcal{G})}\right]}_{*}+\underbrace{2\log\left(\frac{f(X^n)}{\hat{f}(X^n)}\right)^{1/2}}_{**}+2\log card(\mathcal{G}). \tag{5.32}$$

Taking the expected value in (5.38), we have an upper bound on the expected n-fold Hellinger distance:

$$nEH^2(f,\hat{f}) \le E(*)+E(**)+2\log card(\mathcal{G}). \tag{5.33}$$

Now we want to simplify the expression by getting rid of the randomness in $\hat{f}$. Let

$$\dagger = \frac{1}{A_n(f,\hat{f})}\left(\frac{\hat{f}(X^n)}{f(X^n)}\right)^{1/2}$$

and

$$\ddagger = \sum_{g\in\mathcal{G}}\frac{1}{A_n(f,g)}\left(\frac{g(X^n)}{f(X^n)}\right)^{1/2}. \tag{5.34}$$

Here the affinity $A_n(f,g)$ will serve as normalizing constants for the square root of density ratios, so that the expected values of the terms in the sum 5.34 are equal to 1.

Because $\hat{f}$ is an element in $\mathcal{G}$, we claim that

$$\dagger \le \ddagger.$$

Note that every term in the sum is positive. So the sum is an upper bound for any single term. Also observe that $\dagger$ is a randomly-chosen term in $\ddagger$. Therefore $\dagger \le \ddagger$ for all data sequences.

Replace $\dagger$ by $\ddagger$ in $E(*)$. We have the following upper bound for $E(*)$,

$$
\begin{aligned}
E(*) &= 2E\log\left[\frac{1}{A_n(f,\hat{f})}\left(\frac{\hat{f}(X^n)}{f(X^n)}\right)^{1/2}\frac{1}{card(\mathcal{G})}\right] \\
&\leq 2E\log[\sum_{g\in\mathcal{G}}\frac{1}{A_n(f,g)}\left(\frac{g(X^n)}{f(X^n)}\right)^{1/2}\frac{1}{card(\mathcal{G})}].
\end{aligned}
$$

Now using the concavity of the logarithm, we can bring the expectation inside of the logarithm and get another upper bound,

$$
\begin{aligned}
E(*) &\leq 2\log[\sum_{g\in\mathcal{G}}\underbrace{E\frac{1}{A_n(f,g)}\left(\frac{g(X^n)}{f(X^n)}\right)^{1/2}}_{=1}\frac{1}{card(\mathcal{G})}] \\
&= 2\log[1] \\
&= 0.
\end{aligned}
$$

For the rest of terms in the Hellinger bound 5.39, we use the fact that $\hat{f}$ was chosen to maximize the likelihood. We obtain that for all $g \in \mathcal{G}$,

$$
\begin{aligned}
E(**) &= E_f\log\frac{f(X^n)}{\hat{f}(X^n)} \\
&= E_f\log\frac{f(X^n)}{g(X^n)} + E_f\log\frac{g(X^n)}{\hat{f}(X^n)} \\
&\leq nD(f\|g)
\end{aligned}
$$

Since this is true for all g, we have

$$
E(**) \leq n\min_{g\in\mathcal{G}} D(f\|g) \tag{5.35}
$$

Combining $E(*)$ and $E(**)$, we have the desired result

$$EH^2(f,\hat{f}) \leq \min_{g\in\mathcal{G}} D(f\|g) + \frac{2\log card(\mathcal{G})}{n}. \tag{5.36}$$

The K-L risk bound follows immediately by using 5.28.

□

Now we are ready to prove the risk bound theorem for the iterative MLE.

THEOREM 5.2 (Hellinger Risk For Iterative MLE) *Let $\hat{f}_k$ be the estimator after k iterations. Let f be the density of the true distribution from which we sample data $X_1, ..., X_n$. Then*

$$EH^2(f,\hat{f}_k) \leq \inf_{g_P\in\mathcal{C}}\{D(f\|g_P) + \gamma\frac{c_{f,P}^2}{k}\} + 2\frac{\log card(\Theta_{k,\epsilon})}{n} + kdB\epsilon. \tag{5.37}$$

PROOF Replacing the $\hat{f}$ in the right side of 5.31 by $\hat{f}_k = \hat{f}_{k,\epsilon}$ and doing a similar arrangement, we have

$$\underbrace{2\log\left[\frac{1}{A_n(f,\hat{f}_k)}\left(\frac{\hat{f}_k(X^n)}{f(X^n)}\right)^{1/2}\frac{1}{card(\Theta_{k,\epsilon})}\right]}_{*} + \underbrace{2\log\left(\frac{f(X^n)}{\hat{f}_k(X^n)}\right)^{1/2}}_{**} + 2\log card(\Theta_{k,\epsilon}). \tag{5.38}$$

Taking the expected value in (5.38), we have an upper bound on the expected n-fold Hellinger distance:

$$nEH^2(f,\hat{f}_k) \leq E(*) + E(**) + 2\log card(\Theta_{k,\epsilon}). \tag{5.39}$$

By the same reasoning as in the previous theorem,

$$E(*) \leq 0.$$

For the rest of terms in the Hellinger bound 5.39, we use the fact that $\hat{f}_k$ was chosen to produce high likelihood $\hat{f}_k(X^n)$. We obtain

$$\begin{aligned} E(**) &= E_f \log \frac{f(X^n)}{\hat{f}_k(X^n)} \\ &= E_f \log \frac{f(X^n)}{g_P(X^n)} + E_f \log \frac{g_P(X^n)}{\hat{f}_k(X^n)} \\ &= nD(f\|g_P) + E_f \log \frac{g_P(X^n)}{\hat{f}_k(X^n)} \\ &\leq nD(f\|g_P) + \gamma \frac{c_{f,P}^2}{k} n + nkdB\epsilon \end{aligned}$$

for all $g_P \in \mathcal{C}$. The last inequality uses the Nearly Maximum Likelihood property of $\hat{f}_{k,\epsilon}$ in 3.20. Indeed, for all $g_P \in \mathcal{C}$,

$$E \frac{1}{n} \log \frac{g_P(x^n)}{\hat{f}_{k,\epsilon}(x^n)} \leq E\gamma \frac{c_{F_n,P}^2}{k} + kdB\epsilon = \gamma \frac{c_{f,P}^2}{k} + kdB\epsilon.$$

By combining $E(*)$ and $E(**)$, we have,

$$\begin{aligned} nEH^2(f, \hat{f}_k) &\leq E(*) + E(**) + 2 \log card(\Theta_{k,\epsilon}) \\ &\leq 0 + nD(f\|g_P) + \gamma \frac{c_{f,P}^2}{k} n + 2 \log card(\Theta_{k,\epsilon}) + nkdB\epsilon \\ &\leq n \min_{g_P \in \mathcal{C}} [D(f\|g_P) + \gamma \frac{c_{f,P}^2}{k}] + 2 \log card(\Theta_{k,\epsilon}) + nkdB\epsilon. \end{aligned}$$

Dividing both sides through by n, we get the desired Hellinger risk bound.

□

5.4.2 Approximation-adjusted Risk Bounds

In our application, a non-negligible Kullback-Leibler approximation error $D(f\|\mathcal{C}) = D(f\|f^*)$ remains when f is not in the information closure (see lemma 4.3) of the convex hull $\mathcal{C}$. In this section, we bound the Kullback risk difference $ED(f\|\hat{f}) - D(f\|f^*)$ and an associated approximation adjusted Hellinger risk. Thereby establishing that $\hat{f}$ converges to f^*.

For a given density f and each density $g \in \mathcal{C}$, we form a new density $\frac{fg}{f^*}/c$, where $c = \int \frac{fg}{f^*}$ is the normalizing constant. Here f^* is the information projection of f on $\mathcal{C}$ and satisfies that $D(f\|f^*) = \inf_{g\in\mathcal{C}} D(f\|g)$. By the information geometry we established in chapter 4, such an f^* always exists.

Then we define an Approximation-Adjusted Hellinger distance between f and g as

$$H^2(f, \frac{fg}{f^*}/c).$$

Note that the density ratio between f and $\frac{fg}{f^*}/c$, which is equal to $\frac{f^*c}{g}$, does not involve f anymore. We can control it as shown in Lemma 4.5. This property helps in getting a K-L risk bound from a Hellinger risk bound.

Again we seek a log-affinity upper bound for H^2. We first define f^*-adjusted affinity as follows. Given f and $\mathcal{C}$ with projection f^*, let

$$A(g) = \int f\sqrt{g/f^*} = E_f\sqrt{g(X)/f^*(X)},$$

and let

$$A_n(g) = E_f\sqrt{g(X^n)f^*(X^n)} = (A(g))^n$$

for $X_1, X_2, ..., X_n$ i.i.d. $\sim f$. Following lemma establishes an adjusted Affinity bound for

adjusted Hellinger distance.

LEMMA 5.12 *For the approximation adjusted Hellinger distance, we have an upper bound: for* $g \in \mathcal{C}$,

$$H^2(f, \frac{fg}{f^*}/c) + \log(1/c) \leq 2 \log \frac{1}{A(g)} \tag{5.40}$$

where $c = \int \frac{fg}{f^*}$.

PROOF The proof is similar to Lemma 5.10. Note that $\frac{fg}{f^*}/c$ is a density function. So

$$H^2(f, \frac{fg}{f^*}/c) \leq 2 \log \frac{1}{A(f, \frac{fg}{f^*}/c)} = 2 \log \frac{1}{A(g)} + \log c$$

□

Now we improve on the basic result in theorem 5.1.

THEOREM 5.3 (Adjusted Risk Bounds For MLE On a Finite Set of Densities) *Let* $\mathcal{G}$ *be a finite set of densities contained in a convex set* $\mathcal{C}$. *Let* $\hat{f}$ *maximize the likelihood over* $g \in \mathcal{G}$ *and let data* $X_1, ..., X_n$ *be i.i.d.* $\sim f$. *Then*

$$E[2 \log \frac{1}{A(\hat{f})}] \leq \min_{g \in \mathcal{G}} D(f\|g) - D(f\|f^*) + \frac{2 \log card(\mathcal{G})}{n}. \tag{5.41}$$

And consequently, if $\log f^*(x)/g(x) \leq 2a$, *then*

$$\frac{1}{\gamma_1}[ED(f\|\hat{f}) - D(f\|f^*)] \leq \min_{g \in \mathcal{G}} D(f\|g) - D(f\|f^*) + \frac{2 \log card(\mathcal{G})}{n} \tag{5.42}$$

where $\gamma_1 = 2 + 2a$.

PROOF Do a similar manipulation on $2 \log \frac{1}{A_n(\hat{f})}$ as we did in theorem 5.1.

$$2\log\frac{1}{A_n(\hat{f})} = \underbrace{2\log[\frac{1}{A_n(\hat{f})}\left(\frac{\hat{f}(X^n)}{f^*(X^n)}\right)^{1/2}\frac{1}{card(\mathcal{G})}]}_{*} + \underbrace{2\log\left(\frac{f^*(X^n)}{\hat{f}(X^n)}\right)^{1/2}}_{**} + 2\log card(\mathcal{G}).$$

Now take expected value on both sides,

$$2E_f\log\frac{1}{A_n(\hat{f}_k)} \leq E_f(*) + E_f(**) + 2\log card(\Theta_{k,\epsilon}).$$

Again $E_f(*) \leq 0$ through the same reasoning.

And

$$\begin{aligned} E_f(**) &= E_f\log\left(\frac{f^*(X^n)}{\hat{f}(X^n)}\right) \\ &= n[D(f\|\hat{f}) - D(f\|f^*)] \\ &\leq n\min_{g\in\mathcal{G}}[D(f\|g) - D(f\|f^*)] \end{aligned}$$

where the last inequality is the result of $\hat{f}$ being the MLE in $\mathcal{G}$.

Now we get the desired result. The K-L risk result is a straightforward application of lemma 5.28. □

For our iterative MLE estimator, we can apply the above theorem and obtain the risk bounds.

THEOREM 5.4 (Adjusted Risk Bounds For Iterative MLE) *Use the same set-*

ting as in Theorem 5.2. Under adjusted Hellinger risk, we have

$$E[\int f(\sqrt{\frac{\hat{f}_k}{f^*c}}-1)^2-\log c]\leq \gamma\frac{c_{f,*}^2}{k}+2\frac{kd\log(A/\epsilon)}{n}+kdB\epsilon. \tag{5.43}$$

And consequently,

$$\frac{1}{\gamma_1}[ED(f\|\hat{f}_k)-D(f\|f^*)]\leq \gamma\frac{c_{f,*}^2}{k}+\frac{2kd\log(A/\epsilon)}{n}+kdB\epsilon. \tag{5.44}$$

PROOF First we have Affinity bound,

$$nH^2(f,f\hat{f}_k/(f^*c))+n\log\frac{1}{c}\leq 2\log\frac{1}{A_n(\hat{f}_k)}. \tag{5.45}$$

Do a similar manipulation on $2\log\frac{1}{A_n(\hat{f}_k)}$ as we did in the previous theorem,

$$\begin{aligned}2\log\frac{1}{A_n(\hat{f}_k)} &= \underbrace{2\log[\frac{1}{A_n(\hat{f}_k)}\left(\frac{\hat{f}_k(X^n)}{f^*(X^n)}\right)^{1/2}\frac{1}{card(\Theta_{k,\epsilon})}]}_{*}\\ &\quad+\underbrace{2\log\left(\frac{f^*(X^n)}{\hat{f}_k(X^n)}\right)^{1/2}}_{**}+2\log card(\Theta_{k,\epsilon}).\end{aligned}$$

Now take expected value on both sides,

$$2E_f\log\frac{1}{A_n(\hat{f}_k)}\leq E_f(*)+E_f(**)+2\log card(\Theta_{k,\epsilon}).$$

Again $E_f(*)\leq 0$ through the same reasoning.

By the Nearly Maximum Likelihood property, for any $q_P = \int \phi_b(x)P(db)$,

$$\frac{1}{n}\log \hat{f}_k(X^n) \geq \frac{1}{n}\log q_P(X^n) - \gamma\frac{c^2_{F_n,P}}{k} - kdB\epsilon.$$

Take the expected value with respect to f,

$$E\log \hat{f}_k \geq E\log q_P - \gamma\frac{c^2_{f,P}}{k} - kdB\epsilon.$$

Hence,

$$E\log \hat{f}_k \geq \sup_P\{E\log q_P - \gamma\frac{c^2_{f,P}}{k}\} - kdB\epsilon.$$

The right side is at least

$$E\log f^*(x) - \gamma\frac{c^2_{f,*}}{k} - kdB\epsilon$$

because $E\log\frac{f}{f^*} = \inf_P E\log\frac{f}{q_P}$ and $c^2_{f,*}$ is the limit infimum of $c^2_{f,P}$ for q_P with $E\log q_P \to E\log f^*$. So

$$E(**) \leq n(\gamma\frac{c^2_{f,*}}{k} + kdB\epsilon).$$

So

$$2\log\frac{1}{A_n(\hat{f}_k)} \leq n(\gamma\frac{c^2_{f,*}}{k} + kdB\epsilon)) + 2\log card(\Theta_{k,\epsilon}).$$

Now dividing 5.45 through by n we have the desired upper bound.

$$EH^2(f, f\hat{f}_k/(f^*c)) - \log(c) \leq \frac{2}{n}\log\frac{1}{A_n(\hat{f}_k)} \leq \gamma\frac{c^2_{f,*}}{k} + kdB\epsilon + 2\frac{\log card(\Theta_{k,\epsilon})}{n}. \quad (5.46)$$

Consequently, we have a bound for Kullback-Leibler risk,

$$\frac{1}{\gamma}[ED(f\|\hat{f}_k) - D(f\|f^*)] \leq \gamma\frac{c^2_{f,*}}{k} + \frac{2\log(card(\Theta_{k,\epsilon}))}{n} + kdB\epsilon. \quad (5.47)$$

in view of the following arguments.

In equation 5.28, let $p = f$ and $q = \frac{f\hat{f}_k}{cf^*}$. Then

$$V = \max_x \frac{f^* c}{\hat{f}_k} \leq \max_x \frac{f^*}{\hat{f}_k} \leq e^{2a}.$$

The first inequality is true because $c \leq 1$. The second inequality comes from Lemma 4.5. Then

$$2 + \log V \leq 2 + 2a = \gamma_1 \leq \gamma.$$

The second inequality is true because $\gamma = 4[\log(3\sqrt{e}) + a]$ (see chapter 3). Thus

$$D(f \| \frac{f\hat{f}_k}{f^*}/c) \leq \gamma_1 H^2(f, \frac{f\hat{f}_k}{f^*}/c). \tag{5.48}$$

Denoting the right-side of 5.46 to be $\eta(k, n, \epsilon)$, we have,

$$H^2(f, \frac{f\hat{f}_k}{f^*}/c) \leq \eta(k, n, \epsilon) + \log c.$$

So the 5.48 becomes

$$[D(f\|\hat{f}_k) - D(f\|f^*)] + \log c \leq \gamma_1[\eta(k, n, \epsilon) + \log c].$$

Thus,

$$D(f\|\hat{f}_k) - D(f\|f^*) \leq \gamma_1\eta(k, n, \epsilon) + (\gamma_1 - 1)\log c \leq \gamma_1\eta(k, n, \epsilon).$$

The inequality comes from the facts that $\gamma_1 > 1$ and $\log c \leq 0$.

□

REMARK We get ride of the non-diminishing approximation error by adjusting our loss function accordingly for the f^*. □

5.5 Risk Bounds For Estimated k Using MDL

We will first prove a theorem that's an improvement of Barron and Cover [1991]. Then we treat the special case of selecting number of components in the mixture models.

THEOREM 5.5 (Risk Bounds for MDL) *Let $\mathcal{G}$ be a countable collection of densities included in a convex set $\mathcal{C}$. Let $L(g)$ satisfy $\sum_{g\in\mathcal{G}} e^{-L(g)} \leq 1$, and let $\hat{f}_{MDL}$ achieve $\min_{g\in\mathcal{G}}[\log(1/g(X^n)) + 2L(g)]$, where data $X_1, ..., X_n$ are i.i.d. $\sim f$. Then*

$$2E\log\frac{1}{A(\hat{f}_{MDL})} \leq \min_{g\in\mathcal{G}}[D(f\|g) - D(f\|f^*) + \frac{2L(g)}{n}] \tag{5.49}$$

where f^* is the reversed information projection of f on $\mathcal{C}$. And consequently if $\frac{f^*(x)}{g(x)} \leq e^{2a}$, then

$$\frac{1}{\gamma_1}[D(f\|\hat{f}_{MDL}) - D(f\|f^*)] \leq \min_{g\in\mathcal{G}}[D(f\|g) - D(f\|f^*) + \frac{2L(g)}{n}] \tag{5.50}$$

where $\gamma_1 = 2 + 2a$.

PROOF The proof of this theorem is almost identical to the proof for the case of mixture models. For brevity, we only give the proof of mixture models in detail. □

The proof of risk bounds for models selected by the MDL principle is similar to the proofs in the preceding sections. In our case of mixture models, as is common with MDL applications, $\mathcal{G}$ consists of a union of families $\mathcal{G}_k$ of models indexed by a parameter k. For us k is the number of components in the mixture and $\mathcal{G}_k = \{f_k(x|\theta) : \theta \in \Theta_{k,\epsilon}\}$, and for $g = f_{k,\theta}$, $L(g)$ is

$$L(g) = L(k) = \log card(\Theta_{k,\epsilon}) + l(k).$$

Here $l(k)$ satisfies $\sum_k e^{-l(k)} \leq 1$ and is the code length for describing an integer k. And $card(.)$ is the cardinality of the model class. Hence $\log Card(\Theta_{k,\epsilon})$ is the code length for

describing a particular parameter with k components. (See Cover and Thomas [1991] for a detailed explanation of codelength.)

We use the following MDL principle in choosing k for every ϵ,

$$\hat{k} = \arg\min_k \frac{1}{n}[\log \frac{1}{\hat{f}_k(X^n)} + 2L(k)]$$

where $\hat{f}_k(X^n)$ is the likelihood achieved by the iterative maximum likelihood estimate. There is a factor 2 in front of the codelength L. We will show that it comes up naturally in the proof of the following theorem.

THEOREM 5.6 (Risk Bound For Mixture Models with MDL-selected k) *Choose $\hat{k}$ following above prescription. We have a risk bound*

$$\frac{1}{\gamma_1}[ED(f\|\hat{f}_{\hat{k}}) - D(f\|f^*)] \leq \min_k\{\frac{\gamma c_{f,*}^2}{k} + 2L(k)/n + kdB\epsilon\}.$$

PROOF We first establish an Adjusted-Hellinger risk bound for $\hat{f}_{\hat{k}}$. Let f^* be the projection. Again, we define

$$A(g) = \int f\sqrt{\frac{g}{f^*}}$$

and

$$A_n(g) = \int f(x^n)\sqrt{\frac{g(x^n)}{f^*(x^n)}}dx^n.$$

Then for $c = \int f\hat{f}_{\hat{k}}/f^*$, we have

$$\begin{aligned} n\frac{1}{\gamma_1}[ED(f\|\hat{f}_{\hat{k}}) - D(f\|f^*)] &\leq nE[H^2(f, \frac{f\hat{f}_{\hat{k}}}{f^*}/c) - \log c] \\ &\leq 2nE\log\frac{1}{A(\hat{f}_{\hat{k}})} \end{aligned}$$

$$
\begin{aligned}
&= 2E\log\frac{1}{A_n(\hat{f}_{\hat{k}})} \\
&= \underbrace{2E\log[\frac{1}{A_n(\hat{f}_{\hat{k}})}\left(\frac{\hat{f}_{\hat{k}}(X^n)}{f^*(X^n)}\right)^{1/2}\frac{1}{card(\Theta_{\hat{k},\epsilon})}e^{-l(\hat{k})}]}_{*} \\
&\quad \underbrace{+2E\log\sqrt{\frac{f^*(X^n)}{\hat{f}_{\hat{k}}(X^n)}}+2\log card(\Theta_{\hat{k},\epsilon})+2l(\hat{k})}_{**}.
\end{aligned}
$$

We will write $\hat{f}_{\hat{k}}(X^n)$ as $f(X^n|\hat{\theta},\hat{k})$, where $\hat{\theta}\in\Theta_{k,\epsilon}$ and $\hat{k}\in\{1,2,...\}$. We upper bound part * by

$$
2E\log\dagger = 2E\log\sum_{k\in\{1,2,...\}}e^{-l(k)}\frac{1}{card(\Theta_{k,\epsilon})}\sum_{\theta\in\Theta_{k,\epsilon}}\frac{1}{A_n(f_{\theta,k})}\left(\frac{f(X^n|\theta,k)}{f^*(X^n)}\right)^{1/2}. \tag{5.51}
$$

By the concavity of logarithm, we get:

$$
2E\log(\dagger)\le 2\log(E(\dagger)). \tag{5.52}
$$

Observe that

$$
\begin{aligned}
E(\dagger) &= \sum_{k\in\{1,2,...\}}e^{-l(k)}\frac{1}{card(\Theta_{k,\epsilon})}\sum_{\theta\in\Theta_{k,\epsilon}}\frac{E(\frac{f(X^n|\theta,k)}{f^*(X^n)})^{1/2}}{A_n(f_{\theta,k})} \\
&= \sum_k e^{-l(k)}\frac{1}{card(\Theta_{k,\epsilon})}\sum_{\theta\in\Theta_{k,\epsilon}}1 \\
&= \sum_k e^{-l(k)}\le 1
\end{aligned}
$$

since $E(\frac{f(X^n|\theta,k)}{f^*(X^n)})^{1/2}=A_n(f_{\theta,k})$ by the definition of A_n. So the part $*\le 2\log 1=0$. Now we have

$$
\frac{1}{\gamma_1}[ED(f\|\hat{f}_{\hat{k}})-D(f\|f^*)]\le\frac{1}{n}E\log\frac{f^*(X^n)}{\hat{f}_{\hat{k}}(X^n)}+2\frac{L(k)}{n}.
$$

By the definition, $\hat{k}$ is the minimizer of the right side among all k's. So we have

$$\begin{aligned}
\frac{1}{n}E\log\frac{f^*(X^n)}{\hat{f}_{\hat{k}}(X^n)}+2\frac{L(\hat{k})}{n} &\leq \min_k[\frac{1}{n}E\log\frac{f^*(X^n)}{\hat{f}_k(X^n)}+2L(k)/n] \\
&\leq \min_k[\frac{1}{n}E\log\frac{f^*(X^n)}{f^*(X^n)}+\gamma\frac{c_{f,*}^2}{k}+kdB\epsilon+2L(k)/n] \\
&= \min_k[\gamma\frac{c_{f,*}^2}{k}+kdB\epsilon+2L(k)/n].
\end{aligned}$$

The second inequality is the result of the nearly maximum likelihood we have established for $\hat{f}_k$. We now have the desired upper bound for the risk. □

Chapter 6

Discussions

6.1 Number of Components

In our method, we do not assume there is a true number of components. We determine the number of components to minimize the statistical risk. The more components we have, the smaller the approximation error and the larger the estimation error will be.

On the other hand, in the situation that the truth is a k-component Gaussian mixture, it's interesting to know how the method behaves.

6.2 Curse of Dimensionality

The estimation procedure still involves an optimization over d dimensional parameters. So even though we have used a greedy algorithm to reduce the search from $d \times k$ dimension to d dimension, the search in d dimension can still be difficult. We can still encounter multi-modality.

The EM algorithm for mixture density estimation searches for the global MLE in kd dimensions. It will be interesting to compare the Greedy algorithm to the performance

of EM. The Nearly Maximum Likelihood Property implies that our algorithm should do nearly as well as a global search for the MLE.

6.3 L^2 distance

We have a similar result for using L^2 distance in my prospectus [1997]. The problem is that the constant is much larger. Instead of the log ratio of densities, the ratio of densities appears in the bound.

The next theorem shows that a greedy algorithm based on L^2 distance gives us $O(1/k)$ error with a much worse constant.

Let $G = \{\phi_b, b \in \Theta\}$ be a set of parametric densities. Denote the $\mathcal{L}^2$ norm by $\|\cdot\|^2$ Suppose density $f(x)$ is in the closure of convex hull of G.

THEOREM 6.1 (Iterative density approximation in L-2 distance) *Suppose f_1 is chosen to satisfy $\|f_1 - f\|^2 = \inf_{\phi_b \in G} \|\phi_b - f\|^2$. Define f_k iteratively by $f_k = (1-\alpha)f_{k-1} + \alpha\phi_b$. And choose α and b to minimize $\|f - f_k\|^2$. Then for every $k \geq 1$,*

$$\|f - f_k\|^2 \leq \frac{b_f^2}{k} \tag{6.1}$$

where

$$b_f^2 = sup_{\phi_b \in G} \|\phi_b\|^2 \tag{6.2}$$

PROOF The proof is a direct application of Barron [1993] Theorem 5.

□

The following example shows that the constant b_f grows exponentially with dimension d.

Let G be d dimensional Gaussian family with a fixed variance σ^2. We compute the $\mathcal{L}^2$ norm of $\phi_b = N(b, \sigma^2 I_{d,d})$ to be

$$\int g^2 dx = (\frac{1}{4\pi})^{d/2} (\frac{1}{\sigma})^d. \tag{6.3}$$

When $\sigma < \frac{1}{\sqrt{4\pi}}$, the constant grows exponentially with dimension d. Remember that the constant in the bound using K-L distance grows linearly with d.

Appendix A

Background: Projection in Hilbert Space

A.1 Definition of Hilbert Space

A Hilbert space $\mathcal{H}$ is a vector space equipped with an inner product $< .,. >$, which satisfies following properties:

- linearity

$$< \alpha f + \beta g, h >= \alpha < f, h > + \beta < g, h >,$$

- symmetry

$$< f, g >=< g, f >,$$

- positivity

$$< f, f > \geq 0 \text{ with equality if and only if } f = 0,$$

- completeness if f_n is a Cauchy sequence, f_n converges.

Define norm $\|f\| = \sqrt{< f, f >}$. Completeness means that if $\|f_m - f_n\| \to 0$ as $m, n \to \infty$, then $\exists f \in \mathcal{H}$, such that $f_n \to f$.

A trivial example of Hilbert space is the Euclidean space. It's a finite-dimensional Hilbert space. A prime example of non-trivial Hilbert space is $\mathcal{L}^2(\mu)$ defined as following. Let μ be a measure in a measurable space $(\Omega, \mathcal{B})$. Let $\mathcal{L}^2(\mu)$ be the set of measurable functions f with $\mu f^2 < \infty$. Then define the inner product as

$$< f, g >= \mu(fg).$$

$< f, g >$ is well-defined because of inequality $|fg| \leq f^2 + g^2$.

REMARK $\mathcal{L}^2(\mu)$ is not a genuine Hilbert space yet, with above definition of inner product. The "positivity" condition is not satisfied. But if we consider $L^2(\mu)$ be a set of equivalent classes. Define $[f] = \{g \in \mathcal{L}^2(\mu), g = f a.s.[\mu]\}$, and $< [f], [g] >=< f, g >$, then $L^2(\mu)$ is a genuine Hilbert space. □

A.2 Existence and Uniqueness of Projection in Hilbert Space

THEOREM A.1 *Let $\mathcal{H}$ be a Hilbert space, $\mathcal{H}_l \subset \mathcal{H}$ and $\mathcal{H}_l$ is a closed subspace. For any $f \in \mathcal{H}$, there exists a $f_0 \in \mathcal{H}_l$ such that $\|f - f_0\| = \inf_{h \in \mathcal{H}_l} \|f - h\|$. Also $< h, f - f_0 >= 0$ for all $h \in \mathcal{H}_l$.*

PROOF Define

$$\delta = \inf_{h \in \mathcal{H}_l} \|f - h\|.$$

Because δ is the greatest lower bound, we can choose a sequence $h_n \in \mathcal{H}_l$, so that $\|f - h_n\| \to \delta$. We can show that h_n is a Cauchy sequence.

Use Parallelogram identity in a normed vector space, we have

$$2\|f - h_m\|^2 + 2\|f - h_n\|^2 = \|2f - h_m - h_n\|^2 + \|h_m - h_n\|^2.$$

By the definition of h_n, re-arranging terms in above identity, we have that $\forall \epsilon > 0$,

$$\|h_m - h_n\|^2 \leq \epsilon.$$

Therefore, we have a Cauchy sequence. By completeness, there exists a $f_0 \in \mathcal{H}$ such that $h_n \to f_0$. And $\mathcal{H}_l$ is closed. It follows that $f_0 \in \mathcal{H}_l$.

Furthermore,

$$\delta \leq \|f - f_0\|$$

and

$$\|f - f_0\| \leq \|f - h_n\| + \|h_n - f_0\| (\text{triangular inequality}) \leq \delta$$

So $\|f - f_0\| \to \delta$.

Orthogonality. Let $h \in \mathcal{H}_l$, then $f_0 + th \in \mathcal{H}_l$ for all $t \in R$. So

$$\|f - (f_0 + th)\|^2 \geq \delta^2.$$

And

$$\|f - (f_0 + th)\|^2 = \|f - f_0\|^2 + 2t < f - f_0, h > + t^2\|h\|^2.$$

i.e.

$$2t < f - f_0, h > + t^2\|h\|^2 \geq 0.$$

This is true for all t. It follows $< f - f_0, h >= 0$.

□

A.3 Projection onto Convex Subsets in a Hilbert Space

Sometimes we are also interested in projection onto convex subsets.

Example 1: Let $(\Omega, \mathcal{F}, \mu)$ be a measurable space. Consider a Hilbert space $L^2(\mu)$. Define $\mathcal{C} = \{X \in L^2(\mu), \mu X = a\}$. $\mathcal{C}$ is a convex subset of $L^2(\mu)$.

Results for closed subspace can be carried into closed convex subset as the following theorem shows.

THEOREM A.2 *Let $\mathcal{H}$ be a Hilbert space, $\mathcal{H}_t \subset \mathcal{H}$ and $\mathcal{H}_t$ is a closed convex subset. For any $f \in \mathcal{H}$, there exists a $f_0 \in \mathcal{H}_t$ such that $\|f - f_0\| = \inf_{h \in \mathcal{H}_t} \|f - h\|$. Also $< f_0 - h, f - f_0 > \geq 0$ for all $h \in \mathcal{H}_t$, consequently we have a Pythagorean identity:*

$$\|f - h\|^2 \geq \|f - f_0\|^2 + \|f_0 - h\|^2.$$

PROOF

The proof of the existence of a projection onto a subspace used only the property that $\frac{h_m}{2} + \frac{h_n}{2}$ remains in the subspace. This property holds true for convex subsets also. Therefore, the existence of a projection carries immediately into the case of convex subsets.

Orthogonality doesn't hold any more. But we have a Pythagorean identity. The result is proven by replacing $f_0 + th$ with $(1 - \alpha)f_0 + \alpha h$ $(0 \leq \alpha \leq 1)$ in the above proof of Orthogonality.

□

REMARK We need to assume a *closed* convex subset to get existence of a projection in the subset.

The characterizing property of projection onto a convex subset is the Pythagorean identity. Even if we don't define a inner product, hence not working with a Hilbert space, we can still prove the Identity for some measure of distance. We will do that in the next

section. □

Appendix B

Background: Some useful inequalities in information theory

There are some very useful inequalities between different measures of distances. They are crucial in getting many results.

LEMMA B.1 *For two probability measures P, Q with densities p and q with respect to a common dominating measure $m(dx)$,*

$$\begin{aligned} D(P\|Q) &\geq -2\log\int\sqrt{p(x)q(x)}m(dx) && \text{(B.1)} \\ &\geq \int(\sqrt{p(x)}-\sqrt{q(x)})^2 m(dx) && \text{(B.2)} \\ &\geq \frac{1}{4}(\int|p(x)-q(x)|m(dx))^2 && \text{(B.3)} \end{aligned}$$

PROOF For inequality B.1,

$$\begin{aligned} D(P\|Q) &= \int p(x)\log\frac{p(x)}{q(x)}m(dx) \\ &= -2\int p(x)\log(\frac{q(x)}{p(x)})^{1/2}m(dx) \end{aligned}$$

$$\begin{aligned} &\geq -2\log\int p(x)(\frac{q(x)}{p(x)})^{1/2}m(dx) \\ &= -2\log\int\sqrt{p(x)q(x)}m(dx) \end{aligned}$$

For B.2, we apply inequality $\log(x) \leq x - 1$.

$$\begin{aligned} -2\log\int\sqrt{p(x)q(x)}m(dx) &\geq -2(\int\sqrt{pq} - 1) \\ &= 2 - 2\int\sqrt{pq} \\ &= \int(\sqrt{p} - \sqrt{q})^2 \end{aligned}$$

For B.3, consider using Cauchy-Schwartz,

$$\begin{aligned} \int|p-q| &= \int(\sqrt{p} - \sqrt{q})(\sqrt{p} + \sqrt{q}) \\ &\leq \sqrt{\int(\sqrt{p} - \sqrt{q})^2\int(\sqrt{p} + \sqrt{q})^2} \\ &\leq 2\sqrt{\int(\sqrt{p} - \sqrt{q})^2} \end{aligned}$$

□

REMARK Hellinger distance between probability measures is upper bounded by Kullback-Leibler distance. The square of L1 distance is in turn bounded by Hellinger distance. So a convergence in Kullback-Leibler divergence implies Hellinger and L1 convergence. □

Kullback-Leibler divergence behaves like a squared distance. If we take the absolute value, positive part, and negative part of $\log(\frac{p}{q})$ respectively, the new measures of distance behave like the square root of D.

LEMMA B.2 *For two densities p and q,*

1.

$$\int p|\log\frac{p}{q}| \leq D + \sqrt{2D}.$$

2.

$$\int p\log^+\frac{p}{q} \leq D + \frac{1}{2}\sqrt{2D}.$$

3.

$$\int p\log^-\frac{p}{q} \leq \frac{1}{2}\sqrt{2D}.$$

PROOF Observe that $|\log(p) - \log(q)| = \log^+\frac{p}{q} + \log^-\frac{p}{q}$. Define $A = \{x : q > p\}$.

$$\begin{aligned}\int p\log^-\frac{p}{q} &= \int_A p\log\frac{q}{p} &\text{(B.4)}\\ &\leq p(A)(\log\frac{q(A)}{p(A)}) &\text{(B.5)}\end{aligned}$$

The inequality comes by normalizing $p(x)x \in A$ and using positivity of K-L divergence. Now use $\log(x) \leq x - 1$.

$$\begin{aligned}p(A)(\log\frac{q(A)}{p(A)}) &\leq q(A) - p(A) &\text{(B.6)}\\ &= \frac{1}{2}L^1(p,q) &\text{(B.7)}\\ &\leq \frac{1}{2}\sqrt{2D} &\text{(B.8)}\end{aligned}$$

□

REMARK Note that $q(A) - p(A)$ is the total variance distance between p and q. Also note that L^1 distance is twice total variance distance. And the L^1 distance is upper bounded by $\sqrt{2D}$ (see Csiszar[1967], Kullback[1967]). □

Bibliography

[1] Akaike, H[1974]. A new look at the statistical identification model. IEEE Trans. Auto. Control 19. 716-723.

[2] Banerjee, Saibal and Rosenfeld, Azriel[1993]. Model-based cluster analysis. Pattern Recognition. Vol. 26, No. 6, 963-974, 1993.

[3] Banfield, Jeffrey and Raftery, Adrian[1993]. Model-based Gaussian and Non-Gaussian Clustering. Biometrics 49, 803-821.

[4] Barron, Andrew[1993]. Universal Aroximation Bounds for Superpositions of a Sigmoidal Function. IEEE Transactions on Information Theory. Vol. 39, No. 3, 930-945

[5] Barron, Andrew[1986]. Entropy and the Central Limit Theorem. Annals of Probability 14: 336-342. April 1986.

[6] Barron, Andrew[1985]. The Strong Ergodic Theorem for Densities: Generalized Shannon-McMillan-Breiman Theorem. Annals of Probability 13: No.4. 1292-1303.

[7] Barron, Andrew and Cover, Thomas [1991]. Minimum complexity density estimation. IEEE Infomation Theory 37, 1034-1054 J.

[8] Barron, Andrew and Sheu, Chyong-Hwa [1991]. Aroximation of Density Functions by Sequences of Exponential Families. Annals of Statistics. 1991. Vol. 19. No.3, 1347-1369.

[9] Barron, A., Rissanen, J. and Yu, B. [1997] The Minimum Description Length Principle in Coding and Modeling. Technical Report. Department of Statistics. Yale University.

[10] Beran, R.J. [1977]. Minimum Hellinger distance estimates for parametric models. Ann. Statist. 5, 445-463.

[11] Bell, Robert and Cover, Thomas[1980]. Competitive Optimality of Logarithmic Investment. Mathematics of Operations Research, 5: 161-166, 1980.

[12] Bell, Robert and Cover, Thomas[1988]. Game-theoretic optimal portfolios. Management Science, 34: 724-733, 1988.

[13] Cover, Thomas and Thomas, Joy [1991]. The elements in information theory.

[14] Csiszar, I.[1967]. Information-type measures of difference of probability distributions and indirect observations. Studia Sci. Math. Hungar. 2 299-318.

[15] Csiszar, I.[1975]. I-divergence Geometry of Probability Distributions And Minimization Problems. Annals of Probability 3: 146-158. 1975.

[16] Csiszar, I. [1984]. Sanov Property, Generalized I-Projection and A Conditional Limit Theorem. Annals of Probability 12: 768-793. 1984.

[17] Devroye Luc[1987]. A Course in Density Estimation. Birkhauser.

[18] Diebolt, Jean and Robert, Christian[1994]. Estimation of finite mixture distributions through Bayesian Sampling. J.R.Statist.Soc. B(1994), 56, No. 2. 353-375.

[19] Everitt, B.S. and Hand, D.J.[1981]. Finite Mixture Distributions. Chapman and Hall.

[20] Feng, Z.D. and McCulloch, C.E. [1996] Using bootstrap likelihood ratios in finite mixture models. J.R. Statist. Soc. B, 58, 609-617.

[21] Hartigan, John[1975]. Clustering Algorithms. John Wiley.

[22] Hartigan, John[1985]. A failure of likelihood ratio asymptotics for normal mixtures. Proc. Berkeley Conference in Honor of Jerzy Neyman and Jack Keifer. Wadsworth Advanced Books, Monterey, CA and Institute of Mathematical Statistics. Hayward, CA. 789-806.

[23] Jones, Lee[1992]. A simple lemma on greedy aroximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. Annals of Statist. 1992, Vol. 20, No. 1, 608-613.

[24] Kent, John and Tyler, David[1991]. Redescending M-estimates of multivariate location and scatter. Annals of Statist. 1991, Vol. 19, No. 4, 2102-2119.

[25] Kullback, S.[1967]. A lower bound for discrimination in terms of variation. IEEE Trans. Information Theory IT-13 126-127.

[26] Lang, Serge[1993] Real and Functional Analysis. Springer-Verlag.

[27] Lee, W.S, Bartlett, P. , Williamson, R. [1995]. Efficient Agnostic Learning of Neural Networks with Bounded Fan-in. 6th Australian Conference on Neural Networks, Sydney, 6th-8th February, 1995.

[28] Leroux, Brian G.[1992]. Consistent estimation of a mixing distribution. Annals of Statist. 1992, vol 20. No. 3, 1350-1360.

[29] Liang, Z., Jaszczak, R.J., and Coleman, R.E. [1992]. Parameter Estimation of Finite Mixtures Using the EM algorithm and Information Criteria with Alication to Medical Image Processing. IEEE Transactions on Nuclear Science, VOL. 39, No. 4, 1992.

[30] McLachlan, G.J. and Basford, K.E. [1988]. Mixture Models: Inference and Alications to Clustering. New York: Dekker.

[31] O'Sullivan, Finbarr [1994]. Images From Dynamic Positron Emission Tomography Studies. Statistical Methods in Medical Research. Vol. 3. 87-101.

[32] O'Sullivan, Finbarr [1993]. Estimation From Multichannel Image Data. JASA. Vol.88. 209-220.

[33] Richardson, Silvia, Green, Peter [1997]. On Bayesian Analysis of Mixtures with an Unknown number of Components. J.R. Statist. Soc. B (1997) 59, No. 4. 731-792.

[34] Ripley, B.D.[1996]. Pattern Recognition and Neural Networks. Cambridge University Press 1996.

[35] Schwartz, Gideon[1978]. Estimating the dimension of a model. Annals of Statist. Vol.6, No. 2, 461-464.

[36] Sclove, Stanley[1987]. Alication of model-selection criteria to some problems in multivariate analysis. Psychometrika. Vol.52, No.3, 333-343.

[37] Sclove, Stanley[1983]. Alication of the Conditional Population-mixture Model to Image Segmentation. IEEE Pattern Analysis and Machine Intelligence. Vol. 5, 428-433.

[38] Scott, David W.[1992]. Multivariate Density Estimation: Theory, Practice, and Visualization. John Wiley & Sons, INC. QA276.8.S28.

[39] Silverman, B.W.[1986]. Density Estimation for Statistics and Data Analysis. Chapman and Hall. QA276.8.

[40] Simar, L.[1976]. Maximum likelihood estimation of a compound Poisson process. Annals of Statist. 4. 1200-1209.

[41] Titterington, D.M., Smith, A.F.M. and Makov, U.E. [1985]. Statistical Analysis of Finite Mixture Distributions. Chichester: Wiley.

[42] Thompson, James, and Tapia, Richard.[1990]. Nonparametric Function Estimation, Modeling and Simulation. SIAM. QA276.8.T37.

[43] Weigend, Andreas, Chin, Elion, and Zimmermann, Heinz [1999]. Computing Portfolio Risk Using Gaussian Mixtures and Independent Component Analysis. Proceedings of 1999 Conference on Computational Intelligence for Financial Engineering (CIFEr99) New York City, March 28-30, 1999.

[44] Wolfowitz, J.[1957]. The minimum distance method. Ann. Math. Statist. 28, 75-88.

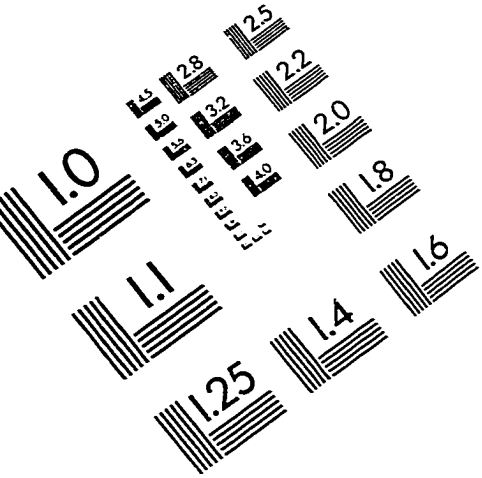
[45] Wong, Wing Hung and Shen, Xiaotong [1995]. Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. Ann. Statist. 1995. Vol. 23, No. 2, 339-362.

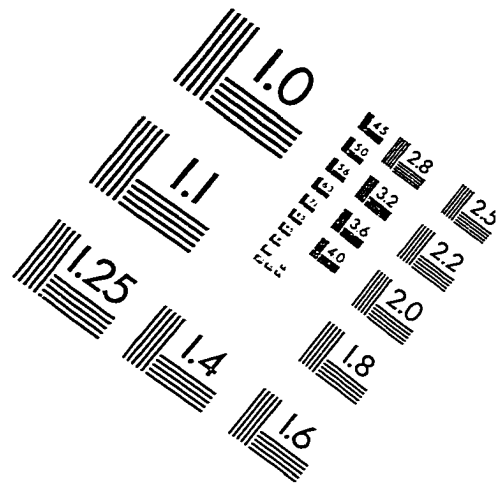
[46] Xu, YueWu[1993]. Asymptotic Problems when there is a lack of identifiability. Ph.D prospectus, Yale University, Department of Statistics.

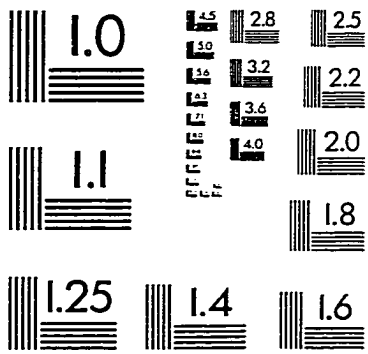
[47] Yang, Yuhong and Barron, Andrew[1998]. An Asymptotic Property of Model Selection Criteria. IEEE Transaction on Information Theory. Vol. 44. No. 1. January 1998.

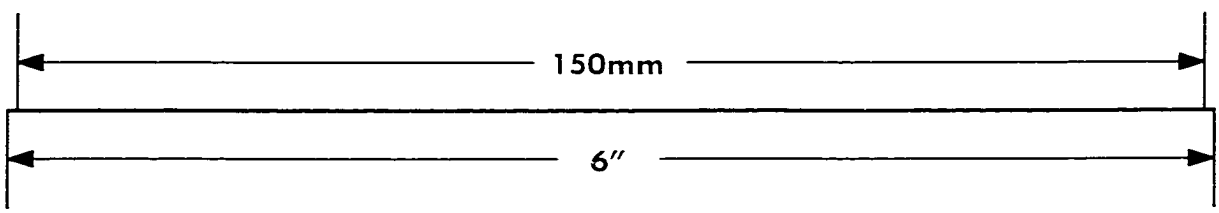
[48] Yatracos, Yannis G.[1985]. Rates of convergence of minimum distance estimators and kolmogorov's entropy. Annals of Statistics. 1985, Vol. 13, No. 2, 768-774.

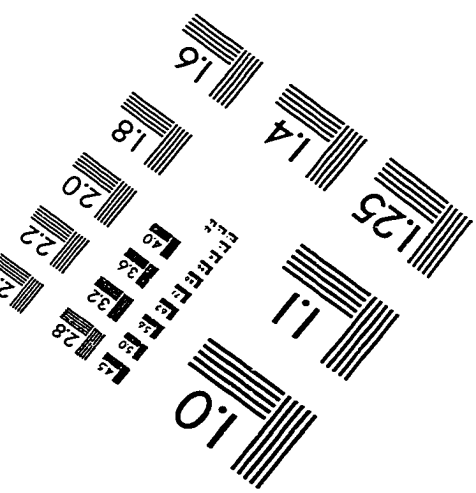
IMAGE EVALUATION
TEST TARGET (QA-3)

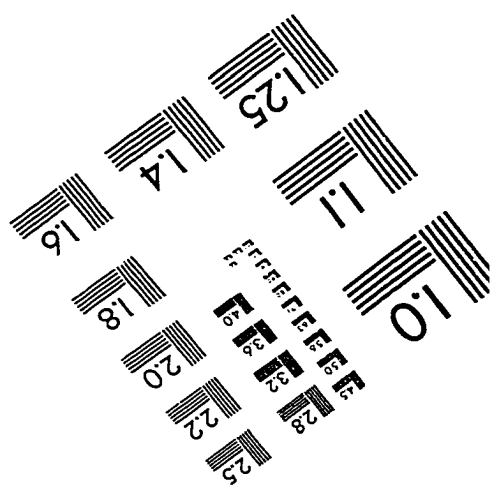












APPLIED IMAGE, Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989