Exact Minimax Predictive Density Estimation and MDL

Feng Liang* and Andrew Barron†

December 5, 2003

Abstract

The problems of predictive density estimation with Kullback-Leibler loss, optimal universal data compression for MDL model selection, and the choice of priors for Bayes factors in model selection are interrelated. Research in recent years has identified procedures which are minimax for risk in predictive density estimation and for redundancy in universal data compression. Here, after reviewing some of the general story, we focus on the case of location families. The exact minimax procedures use an improper uniform prior on the location parameter. We illustrate use of the minimax optimal procedures with data previously used in a study of robustness of location estimates. Plus we discuss applications of minimax MDL criteria to variable selection problems in regression.

1 Introduction

Suppose we are about to transmit a data string $y = (y_1, \dots, y_n)$ and we assume that the underlying data generating process is some distribution from a parametric family with probability density function $p(y \mid \theta)$ depending on a d dimensional parameter vector θ taking values in $\Theta \subset \mathbb{R}^d$. If the parameter θ were known to us, by Shannon coding theory, the ideal code length would be equal to $\log 1/p(y \mid \theta)$ where for now we ignore the requirement of integer code length and finite precision representation of the numbers. Such a code length is optimal in the following two senses: First, it is the shortest code on average, giving entropy as the shortest expected code length; Second, it is competitively optimal [5][8]. Without the knowledge of θ, we in fact code the data with some other distribution, say $q(y)$ with code length $\log 1/q(y)$. The corresponding excess average code length (*expected redundancy*) is given by the Kullback-Leibler divergence

$$\mathbb{E}_{y|\theta} \log \frac{p(y \mid \theta)}{q(y)}. \tag{1}$$

A minimax optimal coding strategy is one that achieves the minimax expected redundancy equal to

$$\min_q \max_\theta \mathbb{E}_{y|\theta} \log \frac{p(y \mid \theta)}{q(y)}. \tag{2}$$

*F. Liang is with the Institute of Statistics and Decision Sciences, Duke University, Durham, NC 27708-0251 (e-mail: feng@stat.duke.edu)

†A. Barron is with the Department of Statistics, Yale University, New Haven, CT 06520 (e-mail: andrew.barron@yale.edu)

By a result from decision theory [12][9][13][16], the minimax code length (2) agrees with the maximin value

$$\begin{aligned}
&\max_w \min_q \int w(\theta)\ \mathbb{E}_{y|\theta} \log \frac{p(y \mid \theta)}{q(y)} d\theta \\
= \quad &\max_w \int w(\theta)\ \mathbb{E}_{y|\theta} \log \frac{p(y \mid \theta)}{p_w(y)} d\theta,
\end{aligned}$$

where w is a prior distribution on Θ and p_w is the corresponding Bayes mixture (marginal) density $p_w(y) = \int w(\theta) p(y \mid \theta) d\theta$ which minimizes the Bayes average redundancy $R_w(q) = \int w(\theta) \mathbb{E}_{y|\theta} \log p(y \mid \theta)/q(y) d\theta$. Thus the mixture density $p_w(y)$ provides the optimal code length $\log 1/p_w(y)$ for model selection by description length criteria. Likewise, the mixture density $p_w(y)$ is also the key ingredient in Bayes factors for model selection.

Previous work has shown that the mixture code p_{w_J} with Jeffreys' prior w_J (proportional to the root of the determinant of the Fisher information matrix) is asymptotically minimax when the square root of the determinant of the information matrix is integrable [3][4][23]. However, for some cases including location families, the Jeffreys' prior is improper (the root determinant of the information matrix is not integrable) and the minimax redundancy is infinite.

We may express both $q(y)$ and $p(y \mid \theta)$ in the predictive form. For example, the joint density $q(y) = q(y_1, \ldots, y_n)$ is given by

$$q(y) = \prod_{m=0}^{n-1} q(y_{m+1} \mid y^m), \quad y^m = (y_1, \ldots, y_m).$$

Then we have the following identity

$$\begin{aligned}
\mathbb{E}_{y|\theta} \log \frac{p(y \mid \theta)}{q(y)} &= \mathbb{E}_{y|\theta} \log \frac{\prod_m p(y_{m+1} \mid y^m, \theta)}{\prod_m q(y_{m+1} \mid y^m)} \\
&= \sum_m \mathbb{E}_{y|\theta} \log \frac{p(y_{m+1} \mid y^m, \theta)}{q(y_{m+1} \mid y^m)}. \qquad (3)
\end{aligned}$$

Each term in the right side of equation (3) is the Kullback-Leibler risk of the predictive density estimator for the $(m+1)$th observation based on the previous m observations, $q(\cdot \mid y^m)$. That is, the expected redundancy (1) is precisely the accumulated Kullback-Leibler risk of the sequence of predictive density estimators. The connection between optimal coding and statistical estimation is not a surprise because we know that codes correspond to probability distributions by the fundamental Kraft inequality [8].

For each m, a minimax strategy can be constructed by specifying the predictive distribution $\{q_m^*(\cdot \mid y^m)\}$ which is the solution of

$$\min_q \max_\theta \ \mathbb{E}_{y^{m+1}|\theta} \log \frac{p(y_{m+1} \mid y^m, \theta)}{q(y_{m+1} \mid y^m)}. \qquad (4)$$

Here we summarize some of the recent results of the authors reported in [20] in which we studied certain transformation families, including location families, scale families and combined location and scale families. There we showed that when conditioning on sufficiently many initial observations $(m \geq d)$, the minimax redundancy is finite and is achieved by a particular generalized Bayes rule.

For example, for location families, the minimax procedure is generalized Bayes using the uniform (Lebesgue) prior. Though the priors are improper, the posterior based on enough initial observations are proper (i.e. $\int p(y^m \mid \theta)w(\theta)d\theta$ is finite for each y^m). The product of those sequential minimax estimators, $q^*_m(y_{m+1} \mid y^m)q^*_{m+1}(y_{m+2} \mid y^{m+1}) \cdots q^*_{n-1}(y_n \mid y^{n-1})$, specifies a valid predictive density for $(y_{m+1}, \ldots, y_n)$ conditioning on the previous m observations. In general, this product is not the minimax solution of the total expected redundancy for the future $(n-m)$ observations

$$\min_q \max_\theta \mathbb{E}_{y^n|\theta} \log \frac{p(y_{m+1}, \ldots, y_n \mid y^m, \theta)}{q(y_{m+1}, \ldots, y_n \mid y^m)}, \tag{5}$$

because the values of θ which maximize the risk in (4) may differ at various m. Thus the sum of individual minimax risks is an upper bound on the minimax total risk in (5). Nevertheless, for location and scale problems, we exhibit a constant risk minimax procedure, so it simultaneously provides a minimax solution for both the individual risk and the total risk (5), and in such cases the sum of the individual minimax risks is equal to the minimax total.

When we report predictive density estimates, as we shall do in section 3, it is convenient to do so through the value of the log reciprocal, $\log_2 1/q^*(y_{m+1} \mid y^m)$, not only because they add up nicely to give the total code length, but also because informally it will show, for unusual values of y_i, a degree to which that value is surprising and thereby forces a longer description.

The Minimum Description Length (MDL), as a criterion in model selection, was introduced by Rissanen [22] (see two review papers by Barron et al. [5], and by Hansen and Yu [14]). The idea of MDL is to first represent each model by a universal distribution and then choose the one with the shortest description length for the observed data. In this framework, a good model is the one that captures most features of the data and hence can describe the data in a short code. Our results on exact minimaxity of predictive density estimation provide a means to construct the underlying universal coding scheme for MDL, with the code length achieving the minimax redundancy.

As in other papers in the present collection, and in the above-mentioned review papers, there has been a trend in recent years to compare universal procedures not to the codelength $\log 1/p(y|\theta)$ that would have been the best for a hypothetical distribution governing the data (as in the traditional definition of redundancy as given in equation (1)), but rather to study the regret $\log 1/q(y) - \log 1/p(y|\hat{\theta}(y))$ in which the universal codelength $\log 1/q(y)$ is compared to the to the shortest codelength with hindsight $\min_\theta log 1/p(y|\theta)$, corresponding to the maximum likelihood estimate $\hat{\theta}(y)$. For any strategy q, if one takes the expected value of this regret it differs from the expected redundancy by an amount $E_{y|\theta} \log p(y|\hat{\theta}(y))/p(y|\theta)$. Now in general this difference could depend on θ. However, for location and scale families we find that this difference between expected regret and expected redundancy is a constant (independent of θ, as well as independent of q), and the same conclusion of constancy of the expected difference holds in our setting in which one conditions on an initial set of observations. Thus our procedures for location and scale families, which are exactly minimax for expected redundancy (Kullback-Leibler risk) are also exactly minimax for expected regret.

The editors have asked that we also comment further about the nature of asymptotic expressions for regret and Kullback-Leibler risk, and how, if at all, our exact minimax procedures relate to that asymptotics. For i.i.d. sampling from smooth parametric families, as we have said, the Bayes procedures with Jeffreys prior provide asymptotically minimax expected regret and expected redundancy, provided the square root of the determinant of the Fisher information $I(\theta)$ is integrable [3][4][23][5]. In that case, with no need to condition on initial data, the total code length has expected regret that is asymptotically of the form $(d/2)\log(n/2\pi) + \log\int |I(\theta)|^{1/2} + o(1)$, where d is the parameter dimension and $o(1)$ tends to zero as the sample size n tends to infinity as shown in [3][4]. The same asymptotics hold also for the minimax individual sequence regret, as reviewed in [5], though as mentioned there it requires substantial modification of Jeffreys prior when outside exponential families.

Continuing with our focus on expected regret or expected redundancy (Kullback-Leibler risk) incorporating conditioning on a initial sample of size m, we see that if m as well as n are large, then taking the difference in the risk expressions at the final size n and the initial size m, many of the terms cancel away, leaving a conditional redundancy of $(d/2)\log n/m + o(1)$ where $o(1)$ tends to zero as m and $n > m$ tends to infinity. However, unlike the total redundancy, such asymptotic differences do not reveal much role for the choice of procedure, as the results of the [3][4] show that $(d/2)\log n/m + o(1)$ is the asymptotic conditional redundancy of Bayes procedures for all choices of smooth prior. A considerably more refined asymptotic analysis is in Hartigan [15] where he shows that the Kullback-Leibler risk for one-step ahead predictive density estimation with a sample of size k has asymptotic expression $(d/2)(1/k) + c(\theta, w)/k^2 + o(1/k)^2$, where $c(\theta, w)$ depends in a somewhat complicated way on the parameter value and the derivate of the log of the prior density $w(\theta)$ as well as the form of the parametric family. Summing Hartigan's risk expression for k from m to n permits a refined conditional redundancy expression of the form $(d/2)\log n/m + 2c(\theta, w)/m + o(1/m)$ that is sensitive to the choice of procedure (through the choice of the prior w). Thus the asymptotics of expected conditional redundancy, as well as the Kullback-Leibler risk, motivate Hartigan's study of the minimax properties of $c(\theta, w)$ initiated in [15] (one may see also [2] Emerson []). For each family one has a differential inequality to solve to determine if a suggested level C is indeed a minimax bound (that is, one addresses whether there is a prior w such that $c(\theta, w) \leq C$ for all θ). It is reassuring that the priors shown in our work to be exact minimax for finite sample sizes in the special cases of location and scale families do fit in Hartigan's theory as asymptotically minimax.

The remaining of the paper is arranged as follows: in section 2 we summarize ideas from [20] showing the minimaxity for the case of location families. In section 3, we show how to use our result to calculate the MDL criterion value to do model selection on some real data sets, which were used before in a study of robustness by Stigler [25]. The application of variable selection in linear regression model is discussed in section 4 and some additional discussion is given in section 5.

2 Exact Minimax Coding Strategy

In this section, we summarize the derivation of the minimax procedure q^* for location families. It is the simplest case among the transformation families covered by the authors in [20].

Suppose the observations $y^{m+1} = (y_1, \ldots, y_{m+1})$ are from a location family, that is,

$$y_i = z_i + \theta,$$

for $i = 1, \ldots, m+1$ where $\theta \in \mathbb{R}^d$ is an unknown location parameter and $z^{m+1} = (z_1, \ldots, z_{m+1})$ has a known distribution with a joint density denoted by p_0. Then the density for y^{m+1} is given by

$$p(y^{m+1} \mid \theta) = p_0(y^{m+1} - \theta),$$

where $y^{m+1} - \theta$ is a shorthand notation for $(y_1 - \theta, \ldots, y_{m+1} - \theta)$. From now on, we will use p_0 and p as generic expressions for their corresponding marginal and conditional densities. For example, $p_0(z_{m+1})$ denotes the marginal density for z_{m+1} and $p_0(z_{m+1} \mid z^m)$ denotes the conditional density for z_{m+1} given z_1 through z_m.

Without any knowledge of θ, the predictive distribution we use for coding y_{m+1} is denoted by $q(\cdot \mid y^m)$. The expected redundancy (or the *risk* for predictive density estimation) is equal to

$$\mathbb{E}_{y^{m+1}|\theta} \log \frac{p_0(y_{m+1} - \theta)}{q(y_{m+1} \mid y^m)}. \tag{6}$$

Let us first focus on the class of *location invariant* predictive density estimators. For any number $a \in \mathbb{R}^d$, a location invariant estimator q satisfies the following equality,

$$q(y_{m+1} \mid y^m) = q(y_{m+1} - a \mid y^m - a). \tag{7}$$

Supposing our estimator q is location invariant, we can apply the invariance property (7) with $a = y_1$ to equation (6) and obtain

$$\begin{aligned} & \mathbb{E}_{y^{m+1}|\theta} \log \frac{p_0(y_{m+1} - \theta)}{q(y_{m+1} - y_1 \mid 0, y_2 - y_1, y_m - y_1)} \\ = \;& \mathbb{E} \; \log \frac{p_0(y_{m+1} - \theta)}{q(u_m \mid 0, u_1, \ldots, u_{m-1})} \end{aligned} \tag{8}$$

where $u_i = y_{i+1} - y_1$, for $i = 1, \ldots, m$. Notice that u_i is also equal to $z_{i+1} - z_1$ which has a density not depending on the unknown parameter θ. Let $p_u(u_m \mid u^{m-1})$ denote the density for u_m given u^{m-1} derived from p_0. We have the quantity (8) equal to

$$\mathbb{E} \; \log \frac{p_0(y_{m+1} - \theta)}{p(u_m \mid u^{m-1})} + \mathbb{E}_{u^{m-1}} \big[\mathbb{E}_{u_m} \log \frac{p_u(u_m \mid u^{m-1})}{q(u_m \mid 0, u^{m-1})} \big].$$

Notice that the second term in the above quantity is an expected Kullback-Leibler divergence which is always bigger than or equal to zero and is equal to zero if and only if

$$q(u_m \mid 0, u^{m-1}) = p_u(u_m \mid u^{m-1}). \tag{9}$$

Reexpressing in terms of the y_i's and applying the invariance property of q, we have that equation (9) is equivalent to

$$q(y_{m+1} \mid y^m) = p_u(y_{m+1} - y_1 \mid y_2 - y_1, \ldots, y_m - y_1).$$

So the best invariant estimator q^* is the one equal to the right side of the above equality. This analysis for the best invariant density estimator with Kullback-Leibler loss is analogous to that originally given by Pitman [21] (cf. [12], pp. 186–187) for finding the best invariant estimator of θ with squared error loss.

To get a final expression for q^*, we calculate $p_u(u_m \mid u^{m-1}) = p(u^m)/p(u^{m-1})$ and replace u_i by $y_{i+1} - y_1$. Since $u_i = z_{i+1} - z_1$, the joint density for $(z_1, u_1, \ldots, u_{m-1})$ is equal to $p_0(z_1, u_1 + z_1, \ldots, u_{m-1} + z_1)$. Integrating out z_1, we obtain the joint density $p_u(u^{m-1})$ which, when reexpressed in terms of y_i's, is

$$\int p_0(y_1 - \theta, y_2 - \theta, \ldots, y_m - \theta) d\theta = \int p(y^m \mid \theta) d\theta.$$

So the best invariant estimator q^* is equal to

$$q^*(y_{m+1} \mid y^m) = \frac{\int p(y^{m+1} \mid \theta)d\theta}{\int p(y^m \mid \theta)d\theta}, \tag{10}$$

which can be interpreted as the generalized Bayes procedure with the uniform (improper) prior $w(\theta)$ constant on $\mathbb{R}^d$ (Lebesgue measure) for location families.

To show that the best invariant estimator q^* is minimax among all the estimators, we use a result from decision theory (see Ferguson [12]) that constant risk plus extended Bayes implies minimax. The constant risk is a consequence of the location invariance property of q^*.

For a procedure q to be an extended Bayes means that there exists a sequence of Bayes procedures $\{p_{w_k}\}$ with proper priors w_k such that their Bayes risk differences $R_{w_k}(q) - R_{w_k}(p_{w_k})$ go to zero, as $k \to \infty$. Recall that the Bayes procedure p_{w_k} is

$$p_{w_k}(y_{m+1} \mid y^m) = \frac{\int_\Theta p(y^{m+1} \mid \theta)w_k(\theta)d\theta}{\int_\Theta p(y^m \mid \theta)w_k(\theta)d\theta},$$

and the Bayes risk $R_{w_k}(q)$ is

$$R_{w_k}(q) = \int w_k(\theta)\mathbb{E}_{y^{m+1}|\theta} \log \frac{p(y_{m+1} \mid y^m, \theta)}{q(y_{m+1} \mid y^m)} d\theta.$$

The Bayes risk difference for q^* is

$$R_{w_k}(q^*) - R_{w_k}(p_{w_k}) = \mathbb{E}^{w_k}_{y^{m+1}} \log \frac{p_{w_k}(y_{m+1} \mid y^m)}{q^*(y_{m+1} \mid y^m)},$$

where $\mathbb{E}^{w_k}_{y^{m+1}}$ means the expectation is taken with respect to the Bayes mixture $p_{w_k}(y^{m+1})$.

By the chain rule of information theory, the Bayes risk difference is bounded by the total risk difference conditioning on only one observation, without loss of generality, say y_1, and this total risk difference is

$$\begin{aligned} & \mathbb{E}^{w_k}_{y^{m+1}} \log \frac{p_{w_k}(y_2, \ldots, y_{m+1} \mid y_1)}{q^*(y_2, \ldots, y_{m+1} \mid y_1)} \\ = \; & \mathbb{E}^{w_k}_{y^{m+1}}[- \log \frac{\int p(y^{m+1} \mid \theta)w_k(\theta)\frac{1}{w_k(\theta)}d\theta}{\int p(y^{m+1})w_k(\theta)d\theta}] + \mathbb{E}^{w_k}_{y_1}[- \log \int p(y_1 \mid \theta)w_k(\theta)d\theta], \end{aligned} \tag{11}$$

where we use the fact that the density for y_1 given θ, $p(y_1 \mid \theta) = p_0(y_1 - \theta)$, is also a density for θ by the symmetry between y_1 and θ, hence $\int p(y_1 \mid \theta)d\theta = 1$.

Invoking Jensen's inequality ($g(\mathbb{E}X) \leq \mathbb{E}g(X)$ for a convex function g) for both terms in (11), we obtain the Bayes risk difference is less than or equal to

$$\int w_k(\theta) \log w_k(\theta)d\theta - \mathbb{E}^{w_k}_{y_1} \int p_0(y_1 - \theta) \log w_k(\theta)d\theta. \tag{12}$$

By choosing the prior w_k to be normal with mean zero and variance k and changing variables, we finally express the bound (12) as C/k where C is a constant, the second moment of the distribution of z_1. So the Bayes risk difference goes to zero when k goes to infinity, provided that the distribution p_0 has finite second moment. The paper [20] goes further to show the extended Bayes property under a weaker logarithmic moment condition and for other transformation families.

We close this section by noting that for location families, when conditioning on any one observation, say the ith, the minimax predictive density for the rest of the observations is

$$\frac{\int p(y^n \mid \theta)d\theta}{\int p(y_i \mid \theta)d\theta},$$

which reduces to $\int p(y^n \mid \theta)d\theta$ since the denominator is equal to 1 for location families. Thus we obtain the same value no matter which single observation one conditions on.

3 Model Selection in Robust Estimation

We often encounter the problem of estimating the location parameter for some data. The sample mean is a good estimator when the data satisfies the normality assumption, but it can be a very bad one when the data is actually from a heavy-tailed distribution like the Cauchy. The predictive densities are used in formulation of optimal criteria for selection among various shapes of the density to use in the location problem.

Various robust estimators for the location parameter, such as the sample median, have been proposed and compared [1][17][25]. The mean, as a location estimator, works well for data from normal-like distribution because the sample mean is the Maximal Likelihood Estimator (MLE) of the location. Some other robust estimates also correspond to the MLE of certain distributions. We use the mean of the minimax predictive density estimator, $\int \tilde{y}q^*(\tilde{y} \mid y^m)d\tilde{y}$, which arose importantly in the work of Pitman [21]. It is the mean of the posterior density of θ using the uniform prior (when z has mean 0), which Pitman showed to be the minimax estimator of location with squared error loss. It is a nice confluence of decision-theoretic properties that the minimax estimator of location is the mean of the minimax predictive density estimator.

Next we will pick some data sets which have been used before in comparing performances for different robust procedures and calculate the exact minimax MDL for various models to see which one is preferred and to see whether our model selection result is consistent with the results from the robustness study. Here we focus attention on four families of densities: normal, double exponential, Huber, uniform and Cauchy. The double exponential density is $p(y \mid \theta) = p_0(y - \theta)$ with

$$p_0(y) = \frac{1}{2}e^{-|y|}.$$

Its MLE is the sample median. The Huber's density ([17] page 71) is $p(y \mid \theta) = p_0(y - \theta)$ with

$$p_0(y) = \begin{cases} Ce^{-y^2/2}, & |y| \leq k, \\ Ce^{-k|y|+k^2/2}, & |y| > k, \end{cases}$$

where $k = 1.5$. Its MLE is known as the Huber P15 estimator, which is the solution of

$$\sum_{i=1}^{n} \min(k, \max(-k, y_i - \theta)) = 0.$$

The data sets we use in this paper are from Stigler's study for robust location estimators [25]. They are all taken from famous experiments such as 18th century attempts to determine the distance from the earth to the sun and the density of the earth, and 19th century attempts to determine the speed of light. For all 20 data sets, we calculate the description length under the five different models: normal, Huber's density, double exponential, uniform and Cauchy. Combined location and scale families are considered here. The corresponding minimax predictive density is a generalized Bayes using a uniform prior on the location and log-scale parameters, as shown in our work [20]. In terms of model selection using MDL, we find that Cauchy is not supported by the data, which is consistent with what Stigler found "... the data sets considered tend to have slightly heavier tails than the normal, but that a view of the world through Cauchy-colored glasses may be overly-pessimistic."

Table 1 shows more detailed results on model selection for one of Stigler's data sets (table 4, data set 5) with $n = 21$ observations, which are from Short's 1763 determinations of the parallax of the sun. We focus your attention first on the columns with header "minimax". Each entry denotes the log reciprocal of the minimax predictive density, $\log_2[1/q^*(y_i \mid y^{i-1})]$, for the ith observation conditioning on the previous $(i-1)$ observations, using the indicated family of density. Since combined location and scale families are considered here, we have to condition on at least two observations, that is, $i = 3, 4, \ldots, 21$. The totals used for model selection are $\log_2[1/q^*(y_3, \ldots, y_n \mid y_1, y_2)]$, which have interpretations both for minimax code length (MDL) and for Bayes factors. Plug-in type estimators, $p(y_i \mid \hat{\theta}_{i-1}, \hat{\sigma}_{i-1})$, have also been used, where $\hat{\theta}_{i-1}$ and $\hat{\sigma}_{i-1}$ are the MLE based on the previous $i-1$ observations. The product of the plug-in rules arose in the prequential approach to statistical inference studied by Dawid [10][11]. For comparison purposes, we include them in Table 1 too. For this data set, the description lengths based on minimax predictive densities are much shorter than those based on MLE plug-in densities. The two outliers, 10.04 and 10.48, apparently have larger contributions to the totals than the other observations. Surprisingly, the description length for the 5th observation 9.71 is pretty long, especially for the coding strategies using plug-in densities. This is because, without knowing the true parameters, 9.71 does look like an outlier among the first 5 observations, even though it is not among all the 21 observations. We can see that all the minimax predictive densities handel this situation much better than plug-in densities, because they have already taken the unknown location and scale into consideration by averaging. The extreme case is uniform: using MLE plug-in densities, we will have infinity description length once the new observation is outside the range of the previous ones. Note that, for the minimax procedure, the total description length $\log_2[1/q^*(y_3, \ldots, y_n \mid y_1, y_2)]$ does not depend on the order of which the $n-2$ observations $y_3, \ldots, y_n$ are presented, while for plug-in procedure, it does. We randomly permutate the 21 observation 1000 times and calculate the corresponding description length based on plug-in and minimax for normal and double exponential. We find the description lengths based on minimax procedures are much less variant than those based on plug-in procedures.

Further remarks on some practicalities of data compression and prediction may be helpful. The data were, of course, not given as infinite precisions real numbers, but rather they were given to the nearest hundredth. These correspond naturally to intervals of width 1/100 for each observation. The probabilities of these intervals would be the integrals of the densities. Since the densities here do not change perceptibly over these small intervals, the probability is the computed density value times the interval width. Correspondingly, one can report log reciprocal probabilities from Table 1 simply by adding $\log_2 100$ to the entries for each observation. These sum to give the total $\log_2[1/\mathrm{Prob}(y_2, \ldots, y_n|y_1)]$, which, when rounded up to an integer, is the length in bits of the

Table 1. PREDICTIVE DENSITY ESTIMATES

Contributions of each observation to total code lengths

y_i	Normal		Huber's Density		Double Exp		Cauchy		Uniform	
	plug-in	minimax	plug-in	minimax	plug-in	minimax	plug-in	minimax	plug-in	minimax
8.43	-	-	-	-	-	-	-	-	-	-
9.09	-	-	-	-	-	-	-	-	-	-
8.5	2.24	1.81	0.17	1.12	0.54	1.06	0.35	0.92	-0.6	0.99
8.44	2.33	1.52	0.02	0.56	-1.60	0.11	2.79	-1.02	-0.6	0.4
9.71	9.13	1.81	10.86	4.06	47.28	4.55	14.92	5.60	∞	3.96
8.07	3.49	1.65	2.00	1.89	6.03	1.73	7.45	3.78	∞	2.73
8.36	2.27	1.45	0.73	1.00	-0.46	0.51	1.90	1.00	0.71	1.2
8.6	1.99	1.41	0.38	0.64	0.24	0.25	3.38	0.79	0.71	1.13
9.11	3.39	1.47	0.93	1.06	5.64	1.94	6.27	8.79	0.71	1.08
8.66	1.99	1.40	0.26	0.47	-0.19	0.19	2.07	1.97	0.71	1.04
8.58	2.03	1.39	0.23	0.41	-1.41	-0.12	1.05	1.94	0.71	1
9.54	6.41	1.59	2.93	2.52	7.50	3.47	6.43	2.51	0.71	0.98
8.34	2.59	1.42	0.81	0.91	0.49	0.60	1.83	2.02	0.71	0.95
8.55	2.08	1.39	0.34	0.47	-0.99	-0.13	0.86	1.89	0.71	0.94
9.03	2.67	1.41	0.54	0.64	2.39	1.42	4.48	1.98	0.71	0.92
10.04	8.91	1.90	6.30	4.90	8.39	5.03	6.91	11.74	∞	4.87
9.04	2.22	1.39	0.55	0.66	1.65	1.17	3.24	2.57	0.98	1.16
8.71	2.02	1.38	0.43	0.53	-0.33	0.16	0.76	2.54	0.98	1.15
10.48	10.03	2.27	8.00	6.21	8.70	5.72	6.45	3.39	∞	6.37
8.31	2.89	1.45	1.31	1.36	0.93	1.11	1.41	2.64	1.27	1.42
8.67	2.06	1.38	0.71	0.80	-0.52	0.21	0.44	2.53	1.27	1.41
Total	70.76	29.49	37.50	30.21	84.25	28.97	72.99	57.58	∞	33.7

Table 1: Log reciprocal of predictive densities (description lengths) for Short's 1763 determinations of the parallax of the sun (in seconds of a degree). Short's data are listed in the 1st column (from reference [25] data set 5 in Table 4). The row labelled Total provides total code lengths or log Bayes factors for model selection. The $\hat{\theta}$ row gives location estimates based on all 21 observations. For plug-in density estimation, these are the sample mean, Huber's P15, the sample median, and the Cauchy MLE, and for minimax estimation, these are the mean of the predictive densities (Pitman estimators)

.

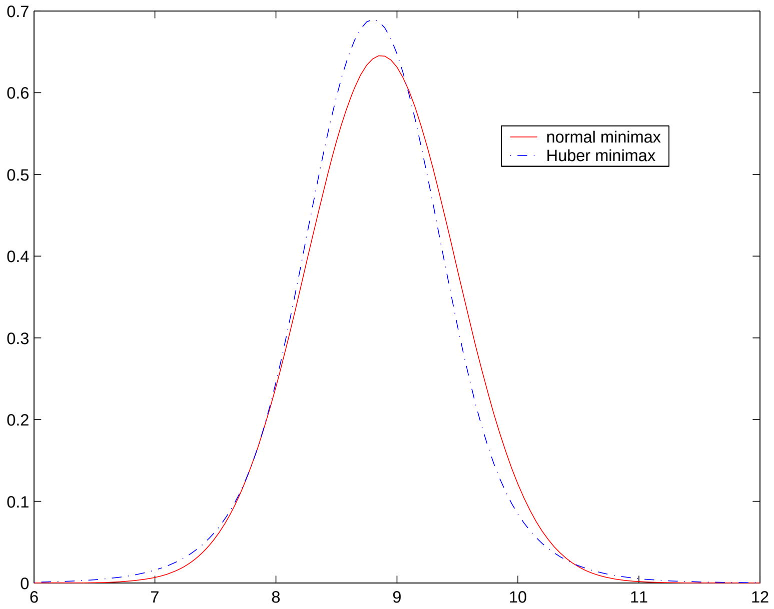

Figure 1: Plot of the minimax predictive densities of normal and Huber location families.

Shannon code for $y_2, \ldots, y_n$ given the value of y_1.

For model selection, one may inspect which of the five minimax predictive distributions provides the shortest $l(y^n) = \log_2 1/q^*(y_2, \ldots, y_n|y_1)$. Then to convert this to a code length, one add $(n-1)\log_2 100$ to convert it to log reciprocal probabilities as required for the Shannon code, and one adds two extra bits to communicate which of the four model is used in the final description.

We recommend before committing to a model selection that one should consider instead the use of model averaging for data compression and prediction. Indeed as we now briefly demonstrate, model averaging provides a shorter code length. To explain, let $\pi(\cdot)$ be a distribution on the model index M, and let $\log 1/\text{Prob}(y \mid M) + \log 1/\pi(M)$ be the total code length for the data using a selected model $M = \hat{M}$, where the term $\log 1/\pi(M)$ is to describe which model is used. On the other hand, if we encode the data with respect to the mixture distribution, it yields code length $\log 1/\sum_M \text{Prob}(y \mid M)\pi(M)$ which is always shorter, because the sum is always greater than any of its terms. The relative contribution of an individual term to the sum is given by its posterior weight $\pi(M \mid y) = \pi(M)\text{Prob}(y|M)/\sum_{M'} \pi(M')\text{Prob}(y|M')$. When this weight for a selected model $\hat{M}$ is nearly one, the mixture code and the model selection based code have essentially the same length. Otherwise it is advantageous to code with the mixture.

For the example problem, we have five models $M = 1, 2, \ldots, 5$, we use $\pi(M) = 1/5$, and the total code lengths are all computed conditional on two observations. For the given data, none of these five individual models stand out as giving much higher probability (shorter code length) than the others as seen in the row labelled Totals. Therefor, coding with respect to the mixture will be better than with model selection.

The corresponding approach for prediction in statistics is called Bayesian model averaging (BMA) [18]. The model averaging can be implemented in one path through the data via a Bayesian update. At observation $i+1$, the partial product $q^*(y_2, \ldots, y_i \mid y_1, M)$ is updated for each of the models. It is used to give posterior weights $\pi(M \mid y^i)$ for each model in the predictive density:

$$q_{ave}(y_{i+1} \mid y^i) = \sum_M \pi(M \mid y^i) q^*(y_{i+1} \mid y^i, M).$$

The final predictive density estimator is $q_{ave}(y \mid y^n)$ and the corresponding final location estimator is then $\hat{\theta}_{ave} = \sum_M \hat{\theta}_M \pi(M \mid y^n)$, where $\hat{\theta}_M$ is the minimax location estimator (Pitman estimator) associated with the predictive density $q^*(y \mid y^n, M)$ for the component M. In our case, the weights $\pi(M \mid y^n)$ are proportional to $2^{-l(y^n|M)}$ where the values for $l(y^n \mid M)$ are given in the row of totals in Table 1. For these data the final location estimate is $\hat{\theta}_{ave} = 8.78$.

4 Variable Selection in Linear Regression

Consider a linear regression model where we have observations y_i and the corresponding possible explanatory variables (also called covariates, predictors, or regressors) $x_{i1}, \dots, x_{id}$. We use γ to index the possible subsets of the d variables, and $x_{i\gamma}$ to denote the column vector of the covariates in γ. Given a subset of variables γ, the observations are modelled by

$$y_i = x_{i\gamma}^t \theta_\gamma + \epsilon_i,$$

where θ_γ is a vector of unknown parameters with dimension equal to d_γ, the size of the subset γ. We all know that the more variables one includes in the regression model, the better will be the fit to the data, at the possible expense of generalizability to new cases. Such an phenomenon is called "overfitting". To avoid it, statisticians look for an subset of variables to achieve a trade-off between fitting errors and model complexity.

If we assume the error ϵ_i has a density function p_0, then the density for y_i is given by

$$p(y \mid \theta, \gamma) = p_0(y - x_{i\gamma}^t \theta_\gamma).$$

Such a distribution family is a generalized location family. Similar analysis to what we did for location families can be applied and it reveals, as shown in [20], that the exact minimax predictive density estimator q^* is the Bayes estimator with uniform prior over the parameter space $\mathbb{R}^{d_\gamma}$, conditioning on $m \geq d_\gamma$ observations.

In ordinary regression models, we often assume that the random error ϵ_i's are normal$(0, \sigma^2)$. Consider first the case that σ^2 is known. The corresponding minimax MDL criterion for variable selection chooses the subset of variables, γ, such that one minimizes

$$\text{MDL}_\gamma = \frac{1}{2\sigma^2}[\text{RSS}_N(\gamma) - \text{RSS}_m(\gamma)] + \frac{1}{2} \log \frac{|(S_N(\gamma)|}{|S_m(\gamma)|},$$

where $S_m(\gamma) = \sum_{i=1}^m x_{i\gamma} x_{i\gamma}^t$ and $\text{RSS}_m(\gamma) = \|y - x_\gamma^t \hat{\theta}_{\gamma,m}\|^2$, respectively, are the information matrix and the residual sum of squares using m observations. Similarly for $S_N(\gamma)$ and $\text{RSS}_N(\gamma)$. Here $|\cdot|$ denotes the determinant. For model selection, we evaluate the criterion for various choices of explanatory variables x_γ (provided $d_\gamma \leq m$), and pick the one that minimizes this optimal description length criterion.

When σ^2 is unknown, we found in [20] that the minimax procedure q^* is generalized Bayes procedure with a uniform prior on the location and log-scale parameters and the corresponding MDL criterion is given by

$$\frac{N-d_\gamma}{2}\log \mathrm{RSS}_N(\gamma) - \frac{m-d_\gamma}{2}\log \mathrm{RSS}_m(\gamma) + \frac{1}{2}\log\frac{|S_N(\gamma)|}{|S_m(\gamma)|} - \log\frac{\Gamma(\frac{N-d_\gamma}{2})}{\Gamma(\frac{m-d_\gamma}{2})}.$$

5 Some Additional Discussion

The priors we showed to provide the exact minimax MDL criterion (uniform on location and log-scale parameters) were suggested by researchers from other perspective before. For example, it is related with the Intrinsic Bayes Factor (IBF) introduced by Berger and Pericchi [7] for the Bayesian model selection. Again, the prior is improper. So they condition on a training sample. The minimal size of the conditioning data for our minimax MDL result agrees with the minimal size of training sample in the IBF, which is the smallest number among those which provide proper predictive densities. Our work provides decision-theoretic optimality (for Kullback risk) of the given choice of priors for IBF and MDL.

The concept of conditioning arises very naturally in time series analysis and in the framework of *prediction without refitting* (see Speed and Yu [24]) where it is of interest to do prediction for some future data based on an initial data set. But when the data does not come with natural order, it is not clear how to implement the exact MDL because of its dependency on the initial data set. A similar problem is encountered in defining intrinsic Bayes factors by Berger and Pericchi [7]. To remove the dependence and increase the stability for the training sample, Berger and Pericchi proposed different averages (such as arithmetic, geometric and median) over all possible training samples. Such an approach can be carried over to the exact MDL, but the description length interpretation may be lost.

The initial observations are used to convert the improper prior to a proper posterior. Therefore one way to avoid conditioning is to find a minimax Bayes procedure which is based on a proper prior. A recent result by the authors [19] has shown that there exists a proper Bayes minimax predictive density estimator with smaller risk than q^* everywhere provided that the dimension is bigger than 4, for normal location families.

Under current investigation is the extent to which the proper Bayes minimax density estimation solution extends to the regression setting. One special case is when the initial design matrix S_m and the total design matrix S_N are proportional to each other. Then a proper prior can be used to assign a description length for the whole data with the property that after the description of the first m observations, the description of the rest is minimax optimal (as well as proper Bayes). Moreover, compared to the minimax code with uniform prior, it provides everywhere smaller (conditional) description length. It is under current investigation whether this result can be extended to more general design matrices.

References

[1] D. F. Andrews, P. J. Bickel, F. R. Hampel, P. J. Huber, W. H. Rogers and J. W. Tukey. (1972) *Robust Estimates of Location: Survey and Advances.* Princeton Univ. Press.

[2] M. Aslan. *Asymptotically Minimax Bayes Predictive Densities.* Ph.D. dissertation, Yale University, 2002.

[3] A. R. Barron and B. S. Clarke. (1990) Information-theoretic asymptotics of Bayes methods. *IEEE Transactions on Information Theory*, 36:453–471.

[4] A. R. Barron and B. S. Clarke. (1994) Jeffrey's prior is asymptotically least favorable under entropy risk. *Journal of Statistical Planning and Inference*, (41):37–60.

[5] A. R. Barron, J. Rissanen, and B. Yu. (1998) The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, 44:2743–2760.

[6] J.O. Berger. (1980) *Statistical Decision Theory: Foundations, Concepts, and Methods.* Springer-Verlag, New York.

[7] J.O. Berger and L.R. Pericchi. (1996) The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91:109–122.

[8] T. Cover and J. Thomas. (1991) *Elements of Information Theory.* New York: John Wiley & Sons.

[9] L. D. Davisson and A. Leon-Garcia. (1980) A source matching approach to finding minimax codes. *IEEE Transactions on Information Theory*, pages 166-174, March.

[10] A. P. Dawid. (1984) Present position and potential developments: Some personal views, statistical theory, the prequential approach. *J. Roy. Statist. Soc. A*, vol. 147, pages 278–292.

[11] A. P. Dawid. (1991) Prequential analysis, stochastic complexity and Bayesian inference. Bayesian statistics 4, *Proceedings of the Fourth Valencia International Meeting.* Clarendon Press, Oxford.

[12] T. S. Ferguson. (1967) *Mathematical Statistics, A Decision Theoretic Approach.* New York: Academic Press.

[13] R. Gallager. (1979) Source coding with side information and universal coding. Tch. Rep. LIDS-P-937, MIT Lab. Inform. Decision Syst., Cambridge, MA.

[14] M. Hansen and B. Yu. (2001) Model selection and minimum description length principle. *Journal of the American Statistical Association*, 96:746–774.

[15] J. Hartigan. The maximum likelihood prior. *The Annals of Statistics*, 26(6):2083–2103, 1998.

[16] D. Haussler. (1997) A general minimax result for relative entropy. *IEEE Transactions on Information Theory*, pages 1276–1280, July.

[17] P. J. Huber. (1981) *Robust Statistics.* John Wiley & Sons.

[18] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. (1999) Bayesian model averaging: A tutorial. *Statistical Science*, 14(4):382–417.

[19] F. Liang. (2002) *Exact minimax strategies for predictive density estimation and data compression.* Ph.D. Dissertation, Yale University, 2002.

[20] F. Liang and A. R. Barron. (2002) Exact minimax strategies for predictive density estimation, data compression and model selection. Accepted by *IEEE Transactions on Information Theory.* Summary appears in *Proceedings of the 2002 IEEE International Symposium on Information Theory.*

[21] E. J. G. Pitman, "The estimation of location and scale parameters of a continuous population of any given form," *Biometrika*, vol. 30, 1939.

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

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

[24] T. Speed and B. Yu. (1993) Model selection and prediction: normal regression. *J. Inst. Statist. Math.*, 45:35–54.

[25] Stephen M. Stigler (1977) Do robust estimators work with real data? *The Annals of Statistics*; Vol 5:6, 1055-1098.