Exact Minimax Procedures for Predictive Density Estimation and Data Compression

A Dissertation

Presented to the Faculty of the Graduate School

of

Yale University

in Candidacy for the Degree of

Doctor of Philosophy

by

Feng Liang

Dissertation Director: Professor Andrew R. Barron

May 2002

UMI Number: 3046185

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

Abstract

Exact Minimax Procedures for Predictive Density Estimation and Data Compression

Feng Liang

2002

For problems of model selection in regression, we determine an exact minimax universal data compression strategy for the minimum description length (MDL) criterion. The analysis also gives the best invariant and indeed minimax procedure for predictive density estimation in location families, scale families and location-scale families, using Kullback-Leibler loss. The exact minimax rule is a generalized Bayes using a uniform (Lebesgue measure) prior on the location parameter for location families and on the log-scale for the scale families, and the product measure on the combined location-scale families. Such improper priors are made proper by conditioning on an initial set of observations.

Our proof for the minimaxity already implies the admissibility for location families in one dimension. However, it is well known that there might exist a better estimator than the constant minimax estimator in high dimension. For example, for normal location families, the sample mean is not admissible when dimension is three or higher (Stein, 55). Moreover, there exists a proper Bayes estimator which is minimax and produces better risk everywhere than the sample mean (Strawderman, 71), when dimension is bigger than four. We present an analogous result for predictive density estimation, using Kullback-Leibler loss.

Contents

Chapter 1

Introduction

Suppose we observed some data from a normal distribution with standard variance and unknown mean. So what is a *good* density estimator for the next observation? Of course, I should first define what I mean about a *good* estimator.

1.1 Problem Statement

Let $\tilde{Y} = (\tilde{Y}_1, \ldots, \tilde{Y}_n)$ be a random vector to which we wish to assign a distribution given observed data $Y = (Y_1, \ldots, Y_m)$. For each model it is assumed that there is a parametric family of distributions $P_{Y|\theta}$ and $P_{\tilde{Y}|Y,\theta}$ with densities $p(y \mid \theta)$ and $p(\tilde{y} \mid y, \theta)$, depending on a d-dimensional parameter vector θ taking values in a parameter space Θ, possibly consisting of all of $\mathbb{R}^d$. To each choice of predictive distribution $Q_{\tilde{Y}|Y}$ with density $q(\tilde{y} \mid y)$ we incur a loss given by the Kullback-Leibler information divergence

$$D(P_{\tilde{Y}|Y,\theta} \| Q_{\tilde{Y}|Y}) = \int p(\tilde{y} \mid y, \theta) \log \frac{p(\tilde{y} \mid y, \theta)}{q(\tilde{y} \mid y)} d\tilde{y}. \tag{1.1}$$

Our interest is in the minimax risk

$$R = \min_{Q} \max_{\theta \in \Theta} \mathbb{E}_{Y|\theta} D(P_{\tilde{Y}|Y,\theta} \| Q_{\tilde{Y}|Y}) \tag{1.2}$$

and in the determination of a predictive distribution $Q_{\tilde{Y}|Y}$ that achieves it. In universal data compression [23][5], the value $\log 1/q(\tilde{y} \mid y)$ corresponds to the length of description of $\tilde{y}$ given y in the absence of knowledge of θ, and the expected Kullback-Leibler loss arises as the excess average code length (redundancy)

$$\mathbb{E}_{Y,\tilde{Y}|\theta}\left[\log 1/q(\tilde{Y}\,|\,Y) - \log 1/p(\tilde{Y}\,|\,Y,\theta)\right].$$

The optimal choice of $q(\tilde{y}\,|\,y)$ is the one providing the minimax redundancy.

In this thesis, I provide exact solution to this minimax problem for certain families of densities parameterized by location or scale. Implications are discussed for predictive density estimation and for the Minimum Description Length (MDL) criterion.

Density Estimation

In density estimation, our aim is to estimate the density function for $\tilde{Y}$ using the data Y in the absence of knowledge of θ. The risk function is the expected Kullback-Leibler loss $R(\theta, q) = \mathbb{E}_{Y|\theta} D(P_{\tilde{Y}|\theta} \| Q_{\tilde{Y}|Y})$. Estimators $q(\tilde{y}\,|\,y)$ are required to be non-negative and to integrate to one for each y, and as such can be interpreted as predictive densities for $\tilde{y}$ given y. Though it may be customary to use plug in type estimators $q(\tilde{y}\,|\,y) = p(\tilde{y}\,|\,\hat{\theta}(y))$, one finds that the optimal density estimators (from Bayes and minimax perspectives) take on the form of an average of members of the family with respect to a posterior distribution given y. We remind the readers of the Bayes optimality property: with prior w and Kullback-Leibler loss, the Bayes risk $R_w(q) = \int R(\theta, q) w(\theta) d\theta$ is minimized by choosing q to be the Bayes predictive density

$$p_w(\tilde{y}\,|\,y) = \int p(\tilde{y}\,|\,y,\theta) w(\theta\,|\,y) d\theta = \frac{\int_\Theta p(y,\tilde{y}\,|\,\theta) w(\theta) d\theta}{\int_\Theta p(y\,|\,\theta) w(\theta) d\theta}. \tag{1.3}$$

Indeed for all q the Bayes risk difference $R_w(q) - R_w(p_w)$ reduces to the expected KL divergence between p_w and q which is positive unless $q = p_w$.

A procedure is said to be generalized Bayes if it takes the same form as in (1.3), with a possibly improper prior (i.e. $\int w(\theta) d\theta$ might not be finite), but proper posterior. Such generalized Bayes procedures arise in our examination of minimax optimality.

I prove that for location families with Kullback-Leibler loss, a minimax procedure is the generalized Bayes using a uniform (Lebesgue) prior. A similar conclusion holds if for a univariate scale parameter $\theta \neq 0$ such that $Y_i = \theta^{-1} Z_i$ where now the minimax procedure uses a uniform prior on $\log |\theta|$. Likewise when one has both multivariate location ($\theta_1 \in \mathbb{R}^d$) and univariate scale ($\theta_2 \neq 0$) parameters such that $Y_i = \theta_2^{-1} Z_i + \theta_1$, the minimax procedure uses Lebesgue product measure on θ_1 and $\log |\theta_2|$.

Partial results (showing the procedure that minimizes risk among invariant estimators) are given for families defined by other groups of transformations including linear transformations $Y_i = \theta^{-1} Z_i$ for $d \times d$ non-singular matrices θ and affine transformations $Y_i = \theta_2^{-1} Z_i + \theta_1$ where $\theta_1 \in \mathbb{R}^d$, θ_2 is non-singular $d \times d$ matrix. The best invariant density estimator uses the prior $1/|\theta|^d$ (where $|\theta|$ denotes the absolute value of the determinant of matrix θ) for linear transformation families, and the prior $1/|\theta_2|^d$ (with respect to Lebesgue product measure on the coordinates of θ_1 and θ_2) for affine families.

For normal location families, I give a proper Bayes estimator which is minimax and produces smaller risk everywhere than the constant minimax estimator. This work is related with Strawderman's [22] proper Bayes estimator for multivariate normal mean vector.

Minimum Description Length

Of particular historical and practical importance is the problem of model selection in linear regression, first considered from the MDL perspective by Rissanen [16]. Suppose we have a total of N observations Y_i which may be predicted using given d-dimensional explanatory vectors x_i for $i = 1, 2, \ldots, N$. One may describe such outcomes using a

Gaussian distribution, in which for given θ and σ^2, each Y_i is modeled as independent Normal$(x_i^t\theta, \sigma^2)$ for $i = 1, 2, \ldots, N$. If σ^2 is fixed and θ is estimated, these models lead to description length criteria of the form

$$\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - x_i^t\hat{\theta})^2 + \frac{d}{2}\log N + c.$$

In Rissanen's original two-stage code formulation, the parameter θ is estimated by least squares and the term $(\frac{d}{2}\log N + c)$ corresponds to the length of description of the coordinates of $\hat{\theta}$ to precision of order $1/\sqrt{N}$. Various values for c have arisen in the literature corresponding to different schemes of quantification of θ, or to the use of mixture or predictive coding strategies rather than two-stage [19]. Asymptotics in N have also played a role in justifying the form of the criterion [5]. When several candidates are available for the explanatory variables x, the model selection criterion picks out the subset of the variables that leads the shortest total description length achieving the best trade off between sum of squared errors and the complexity of the model $(d/2)\log N + c$.

In this thesis I show that if one conditions on m initial observations with m at least as large as the parameter dimension d, then for any regression problem and for all prediction horizon lengths $n \geq 1$, an exact minimax strategy is to use a mixture-based code (or predictive distribution) where the prior is taken to be uniform over θ in $\mathbb{R}^d$ (made proper by conditioning on the initial observations). As a particular case of the general theory, the exact minimax strategy for linear regression models with Gaussian errors is studied. The exact minimax strategy leads to the description length criterion of the form

$$\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - x_i^t\hat{\theta}_N)^2 + \frac{1}{2}\log|\sum_{i=1}^{N} x_i x_i^t| - c_m,$$

where I specify the exact form of c_m (it is $c_m = \frac{1}{2\sigma^2}\sum_{i=1}^{m}(y_i - x_i^t\hat{\theta}_m)^2 + \frac{1}{2}\log|\sum_{i=1}^{m} x_i x_i^t|$). If we set $R_N = \frac{1}{N}\sum_{i=1}^{N} x_i x_i^t$, then the main terms in the penalty are $\frac{d}{2}\log N +$

$\frac{1}{2}\log|R_N|$. For some x_i's (e.g. those evolving according to some nonstationary time series models), the sum $\sum_{i=1}^{N} x_i x_i^t$ may grow at faster rates, e.g. order N^2 rather than N, leading to $\frac{1}{2}\log|NR_N|$ of order $d\log N$ rather than $\frac{d}{2}\log N$. In general it is better to retain the $\frac{1}{2}\log|NR_N|$ determinate form of the penalty rather than the $\frac{d}{2}\log N$. Thus the determinant of the information matrix $\sum_{i=1}^{N} x_i x_i^t$ plays a key role in the exact minimax strategy for regression. Previous work has identified the role of the information matrix in asymptotically optimal two-stage codes [2], in stochastic complexity (Bayes mixture codes) [2][18][3] and in asymptotically minimax code [4][20] when the parameter space is restricted so that the square root of the determinant of the information matrix is integrable.

Priors providing asymptotically minimax codes in [4] are modifications of Jeffreys' prior (proportional to the root of the determinant of the information matrix), historically important [13][11] because of a local invariance property – small diameter Kullback-Leibler balls have approximately the same prior probability in different parts of the parameter space. For the regression problem and other unconstrained location and scale families the Jeffreys' prior is improper (root determinant information is not integrable) commensorate with infinite minimax redundancy. Nevertheless, conditioning on sufficiently many initial observations produces proper posterior distributions and finite maximal risk (conditional redundancy). Conditioning on initial observations can change the asymptotically optimal prior from what it was in the unconditional case. In particular, with conditioning, the optimal prior need not be Jeffreys'. Nevertheless, the procedures we show to be exactly minimax (with conditioning) do coincide with the use of Jeffreys' prior for location or scale families.

1.2 Layout of This Thesis

This dissertation is arranged as following:

the problem has already formulated in section 1 in this Chapter. Implications like density estimation and model selection are also discussed. I do not have a "historical review" in this Chapter, instead I will discuss those related work in each Chapter.

In Chapter 2, I am going to introduce a class of estimators which are invariant under certain transformations such as location shift. One property of invariant estimators is that they have constant risk. The best invariant estimators are calculated for some transformation families such as location families. Examples for some familiar parametric families are given. The understanding of invariance through groups of transformations and the connection between best invariant estimators and right Haar measure are given in the discussion section.

In Chapter 3, I prove that the best invariant estimators are minimax for location families, scale families and the multivariate location and univariate scale families, if conditioning on enough initial data. The minimax risk is instead infinity if not conditioning on enough data set. The proof for minimaxity already implies the admissibility in one dimension. For normal location family, I find that the constant minimax estimator is not admissible when dimension is three or higher. The similar analysis reveals the minimax estimator for regression under Kullback-Leibler loss. Consequently, we can use such a minimax estimator to derive a criterion for model selection in regression.

The minimax estimator, which is also the best invariant estimator with constant risk, is a generalized Bayes estimator with the improper uniform prior on location parameter for location families. In Chapter 4, for normal location family, I will give a proper Bayes estimator which is also minimax and produces better risk everywhere than the constant minimax estimator, provided that the dimension is bigger than four. This piece of work is related with Strawderman's proper Bayes estimator in point estimation for normal location.

Chapter 2

Best Invariant Estimators

2.1 Location Families

Consider first location families. We are to observe $Y = (Y_1, \ldots, Y_m)$ and want to encode or provide predictive distribution for the future observations $\tilde{Y} = (\tilde{Y}_1, \ldots, \tilde{Y}_n)$, where $Y_i = Z_i + \theta$, $\tilde{Y}_i = \tilde{Z}_i + \theta$ with unknown $\theta \in \mathbb{R}^d$. We assume that $Z = (Z_1, \ldots, Z_m)$ and $\tilde{Z} = (\tilde{Z}_1, \ldots, \tilde{Z}_n)$ have a known joint density $p_{Z,\tilde{Z}}$. Then the joint density for Y and $\tilde{Y}$ is given by $p(y, \tilde{y} \,|\, \theta) = p_{Z,\tilde{Z}}(y - \theta, \tilde{y} - \theta)$. We use $y - \theta$ and $\tilde{y} - \theta$ as shorthand notations for $y_1 - \theta, \ldots, y_m - \theta$ and $\tilde{y}_1 - \theta, \ldots, \tilde{y}_n - \theta$, respectively. When the context is clear, we will write $p_{Z,\tilde{Z}}$ as p.

Our first goal is to find the best invariant estimator or coding strategy $q^*(\tilde{y} \,|\, y)$.

Definition 1 *A procedure q is invariant under location shift, if for each $a \in \mathbb{R}^d$ and all $y, \tilde{y}$, $q(\tilde{y} \,|\, y + a) = q(\tilde{y} - a \,|\, y)$.*

That is, adding a constant a to the observations $y = (y_1, \ldots, y_m)$ shifts the density estimator for $\tilde{y}$ by the same amount a. Consequently, if we shift both y and $\tilde{y}$ by the same amount, the value of $q(\tilde{y} \,|\, y)$ is unchanged,

$$q(\tilde{y} + a \,|\, y + a) = q(\tilde{y} \,|\, y). \tag{2.1}$$

Proposition 1 *Invariant procedures have constant risk.*

Proof: Applying the invariance of q, we obtain

$$R(\theta, q) = \mathbb{E}_{Y,\tilde{Y}|\theta} \log \frac{p(Y-\theta \mid \tilde{Y}-\theta)}{q(\tilde{Y}-\theta \mid Y-\theta)} = \mathbb{E}_{Z,\tilde{Z}} \log \frac{p(\tilde{Z} \mid Z)}{q(\tilde{Z} \mid Z)}, \tag{2.2}$$

which is equal to $R(0, q)$, a quantity not depending on θ. Thus Proposition 1 is proved. □

Now we derive the best invariant procedure. The idea is to express the risk in terms of transformed variables that are invariant to the location shift: here $\tilde{Z}_j - Z_1$, $Z_i - Z_1$ for $j = 1, \ldots, n$ and $i = 2, \ldots, m$. Applying the invariance property (2.1) with $a = -Z_1$ in equation(2.2), we obtain

$$R(\theta, q) = \mathbb{E}_{Z,\tilde{Z}} \log \frac{p(\tilde{Z} \mid Z)}{q(\tilde{Z} - Z_1 \mid 0, Z_2 - Z_1, \cdots, Z_m - Z_1)}.$$

Define $\tilde{U} = \tilde{Z} - Z_1$, $U_1 = Z_1$ and $U_i = Z_i - Z_1$ for $i = 2, \ldots, m$. Then $\tilde{U}$ given $U_2, \ldots, U_m$ will have a conditional density function $p(\tilde{u} \mid u_2, \ldots, u_m)$ which we show provides the optimal q. Indeed for any q, the risk satisfies

$$\begin{aligned} R(\theta, q) &= \mathbb{E}_{\tilde{Z},Z} \log \frac{p(\tilde{Z} \mid Z)}{q(\tilde{U} \mid 0, U_2, \ldots, U_m)} \\ &\geq \mathbb{E}_{\tilde{Z},Z} \log \frac{p(\tilde{Z} \mid Z)}{p(\tilde{U} \mid U_2, \ldots, U_m)}, \end{aligned} \tag{2.3}$$

because the difference

$$\mathbb{E} \log \frac{p(\tilde{U} \mid U_2, \ldots, U_m)}{q(\tilde{U} \mid 0, U_2, \ldots, U_m)} = \mathbb{E}_{U_2,\ldots,U_m} [\mathbb{E}_{\tilde{U} \mid U_2,\ldots,U_m} \log \frac{p(\tilde{U} \mid U_2, \ldots, U_m)}{q(\tilde{U} \mid 0, U_2, \ldots, U_m)}]$$

is an expected Kullback-Leibler divergence that is greater than or equal to zero, and it is equal to zero (i.e. achieves the smallest risk) if and only if $q(\tilde{u} \mid 0, u_2, \ldots, u_m) = p(\tilde{u} \mid u_2, \ldots, u_m)$.

This analysis for the best invariant density estimator with Kullback-Leibler loss is analogous to that originally given by Pitman [15] (cf. Ferguson [10], page 186-187) for finding the best invariant estimator of θ with squared error loss.

Now we solve for $p(\tilde{u}\,|\,u_2,\ldots,u_m) = p(u_2,\ldots,u_m,\tilde{u})/p(u_2,\ldots,u_m)$. Note that the mapping from $Z,\tilde{Z}$ to $U,\tilde{U}$ has unit Jacobian. So the joint density $p(u,\tilde{u})$ is given by $p_{\tilde{Z},Z}(u_1,u_2+u_1,\ldots,u_m+u_1,\tilde{u}+u_1)$. Integrating out u_1, we obtain

$$p(u_2,\ldots,u_m,\tilde{u}) = \int p_{Z,\tilde{Z}}(u_1,u_2+u_1,\ldots,u_m+u_1,\tilde{u}+u_1)du_1. \tag{2.4}$$

Observe that $u_i = z_i - z_1 = y_i - y_1$ for $i = 2,\ldots,m$ and $\tilde{u} = \tilde{z} - z_1 = \tilde{y} - y_1$, then (2.4) is equal to

$$\int p_{Z,\tilde{Z}}(u_1, y_2 - y_1 + u_1, \ldots, y_m - y_1 + u_1, \tilde{y} - y_1 + u_1)du_1.$$

Letting $\theta = y_1 - u_1$, we may express this integral as

$$\int p_{Z,\tilde{Z}}(y_1 - \theta, y_2 - \theta, \ldots, y_m - \theta, \tilde{y} - \theta)d\theta = \int p(y,\tilde{y}\,|\,\theta)d\theta.$$

Similarly, we obtain $p(u_2,\ldots,u_m) = \int p(y\,|\,\theta)d\theta$. Thus the conditional density for $\tilde{u}$ given $u_2,\ldots,u_m$ (expressed as a function of y and $\tilde{y}$) is the ratio,

$$p(\tilde{u}\,|\,u_2,\ldots,u_m) = \frac{\int p(y,\tilde{y}\,|\,\theta)d\theta}{\int p(y\,|\,\theta)d\theta}, \tag{2.5}$$

which we denote as $q^*(\tilde{y}\,|\,y)$. One can check that q^* is an invariant procedure under location shift. Our analysis at inequality (2.3) and following show that this predictive density q^* has the smallest risk among all invariant estimators. It is also the unique best invariant one due to the strict convexity of the KL loss. So we get the following proposition.

Proposition 2 *The unique best invariant predictive density for a location family is*

$$q^*(\tilde{y}\,|\,y) = \frac{\int p(y,\tilde{y}\,|\,\theta)d\theta}{\int p(y\,|\,\theta)d\theta}.$$

The procedure q^* we have showed to be the best invariant can be interpreted as a generalized Bayes procedure with uniform (improper) prior $w(\theta)$ constant on $\mathbb{R}^d$ (Lebesgue measure) for location families. Bayes prediction densities are not invariant in general, except for certain improper priors, identified in Hartigan [11] as *relatively invariant* priors, for which $w(\theta + t) = c(t)w(\theta)$, e.g., $w(\theta) = ce^{a\theta}$. A corollary then of Proposition 2 is that the relatively invariant prior with the smallest constant risk is the uniform prior on $\mathbb{R}^d$ $(w(\theta) = c)$.

2.2 Other Transformation Families

Similarly we can derive the best invariant predictive density estimator for other transformation families, such as

1. Linear Transformation family: $Y_i = \theta^{-1}Z_i$, $\tilde{Y}_i = \theta^{-1}\tilde{Z}_i$, where θ is a non-singular $d \times d$ matrix and

$$p(y, \tilde{y} \mid \theta) = |\theta|^{m+n} p_{Z,\tilde{Z}}(\theta y, \theta \tilde{y})$$

   Specially, when $d = 1$, it is called a univariate *scale* family.

2. Affine family: $Y_i = \theta_2^{-1}Z_i + \theta_1$, $\tilde{Y}_i = \theta_2^{-1}\tilde{Z}_i + \theta_1$, $\theta_1 \in \mathbb{R}^d$, θ_2 non-singular $d \times d$ matrix

$$p(y, \tilde{y} \mid \theta) = |\theta_2|^{m+n} p_{Z,\tilde{Z}}(\theta_2(y - \theta_1), \theta_2(\tilde{y} - \theta_1))$$

3. Multivariate location with univariate scale: same as in affine family with $\theta_1 \in \mathbb{R}^d$, but with scalar $\theta_2 \in \mathbb{R} - 0$.

$$p(y, \tilde{y} \mid \theta) = |\theta_2|^{(m+n)d} p_{Z,\tilde{Z}}(\theta_2(y - \theta_1), \theta_2(\tilde{y} - \theta_1))$$

Definition 2 *A procedure q is invariant under linear transformation if for any non-singular $d \times d$ matrix b and all y, $\tilde{y}$, $q(\tilde{y} \mid by) = \frac{1}{|b|^n} q(b^{-1}\tilde{y} \mid y)$. Thus*

$$|b|^n q(b\tilde{y} \mid by) = q(\tilde{y} \mid y). \tag{2.6}$$

It is invariant under affine transformation if for any $a \in \mathbb{R}^d$ and non-singular $d \times d$ matrix b,

$$|b|^n q(b(\tilde{y} - a) \mid b(y - a)) = q(\tilde{y} \mid y). \tag{2.7}$$

Likewise, it is invariant for multivariate location with univariate scale if for any $a \in \mathbb{R}^d$ and non-zero scalar b,

$$|b|^{nd} q(b(\tilde{y} - a) \mid b(y - a)) = q(\tilde{y} \mid y). \tag{2.8}$$

Suppose q is invariant under linear transformation, then the risk $R(\theta, q)$ is equal to

$$\mathbb{E}_{Y,\tilde{Y}|\theta} \log \frac{|\theta|^n p(\theta Y \mid \theta \tilde{Y})}{|\theta|^n q(\theta \tilde{Y} \mid \theta Y)} = \mathbb{E}_{Z,\tilde{Z}} \log \frac{p(\tilde{Z} \mid Z)}{q(\tilde{Z} \mid Z)}.$$

Thus q has constant risk. Similarly, the risk is constant for affine transformation families and affine invariant estimators. Likewise, for multivariate location with univariate scale families. A parallel result to Proposition 2 is given below for the three families.

Proposition 3 *The unique best invariant predictive density is*

$$q^*(\tilde{y} \mid y) = \frac{\int_\Theta \frac{1}{|\theta|^d} p(y, \tilde{y} \mid \theta) d\theta}{\int_\Theta \frac{1}{|\theta|^d} p(y \mid \theta) d\theta}$$

for a linear transformation family,

$$q^*(\tilde{y} \mid y) = \frac{\int_\Theta \frac{1}{|\theta_2|^d} p(y, \tilde{y} \mid \theta) d\theta}{\int_\Theta \frac{1}{|\theta_2|^d} p(y \mid \theta) d\theta}$$

for an affine family where $d\theta$ denotes integration with respect to both location parameter θ_1 in $\mathbb{R}^d$ and scale parameter θ_2 in $\mathbb{R}^{d \times d}$, and

$$q^*(\tilde{y} \mid y) = \frac{\int_\Theta \frac{1}{|\theta_2|} p(y, \tilde{y} \mid \theta) d\theta}{\int_\Theta \frac{1}{|\theta_2|} p(y \mid \theta) d\theta}$$

for a multivariate location with univariate scale family, where $d\theta$ denotes integration with respect to both the location parameter θ_1 in $\mathbb{R}^d$ and the scale parameter θ_2 in $\mathbb{R} - \{0\}$.

Proof: As we studied before, for all three transformation families, we have the risk $R(\theta, q)$ equal to

$$\mathbb{E}_{Z,\tilde{Z}} \log \frac{p(\tilde{Z} \mid Z)}{q(\tilde{Z} \mid Z)}. \tag{2.9}$$

For linear transformation, let Z_1^d denote $(Z_1, \ldots, Z_d)$, the $d \times d$ matrix with Z_i in the ith column for $i = 1, \ldots, d$. Define

$$\tilde{U} = (Z_1^d)^{-1}\tilde{Z}, \; U_i = (Z_1^d)^{-1}Z_i, \; i = d+1, \ldots, m. \tag{2.10}$$

Note that those variables are invariant to linear transformation of the Z_i and $\tilde{Z}$, so that

$$\tilde{U} = (Y_1^d)^{-1}\tilde{Y}, \; U_i = (Y_1^d)^{-1}Y_i, \; i = d+1, \ldots, m, \tag{2.11}$$

where Y_1^d is the $d \times d$ matrix formed from the initial portion of Y.

Applying the invariance property (2.6) in (2.9) with $b = (Z_1^d)^{-1}$, then in a manner similar to the proof for location families (Proposition 2), the best invariant estimator q^* satisfies

$$q^*(\tilde{u} \mid e_1, \ldots, e_d, u_{d+1,m}) = \frac{p(u_{d+1,m}, \tilde{u})}{p(u_{d+1,m})}, \tag{2.12}$$

where e_i is the i^{th} column of the $d \times d$ identity matrix and $u_{d+1,m} = (u_{d+1}, \ldots, u_m)$.

Next we derive the expression (in terms of y and $\tilde{y}$) for both sides of (2.12). By the mapping between $U, \tilde{U}$ and $Z, \tilde{Z}$ given in (2.10), the joint density for $Z_1, \ldots, Z_d, U_{d+1,m}$ and $\tilde{U}$ is given by

$$|z_1^d|^{m+n-d} p_{Z,\tilde{Z}}(z_1, \ldots, z_d, z_1^d u_{d+1}, \ldots, z_1^d u_m, z_1^d \tilde{u}),$$

where $|z_1^d|$ denotes the absolute value of the determinant of the matrix z_1^d and $|z_1^d|^{m+n-d}$ comes out as the Jacobian. Rewriting $u_{d+1,m}$ and $\tilde{u}$ using (2.11) and changing the variables of integration $z_1^d = (z_1, \ldots, z_d)$ to $\theta = z_1^d (y_1^d)^{-1}$, a $d \times d$ matrix, we obtain

$$p(u_{d+1,m}, \tilde{u}) = |y_1^d|^{n+m} \int p(y, \tilde{y} \mid \theta) d\theta.$$

Then the conditional distribution is equal to

$$|y_1^d|^n \frac{\int p(y, \tilde{y} \,|\, \theta) d\theta}{\int p(y \,|\, \theta) d\theta}.$$

On the other hand, using the equalities at (2.11) and the invariance property of q^*, we have the left side of (2.12) equal to $|y_1^d|^n q^*(\tilde{y} \,|\, y)$. So

$$q^*(\tilde{y} \,|\, y) = \frac{\int p(y, \tilde{y} \,|\, \theta) d\theta}{\int p(y \,|\, \theta) d\theta}.$$

For the affine families, define the variables

$$\tilde{U} = [Z_2^{d+1} - Z_1 \mathbf{1}]^{-1}(\tilde{Z} - Z_1), \quad U_i = [Z_2^{d+1} - Z_1 \mathbf{1}]^{-1}(Z_i - Z_1), \quad i = d+2, \ldots, m, \tag{2.13}$$

where $\mathbf{1} = (1, \ldots, 1)$ is the row vector of all one's and thus $Z_1 \mathbf{1}$ is the matrix with d identical columns Z_1. One can see that the variables $\tilde{U}$ and U_i are invariant to affine transformation of the Z_i's and $\tilde{Z}$. Applying the invariance property (2.8) with $a = -Z_1$ and $b = [Z_2^d - Z_1 \mathbf{1}]^{-1}$ in (2.9), we find the best invariant estimator q^* satisfies $q^*(\tilde{u} \,|\, 0, e_1, \ldots, e_d, u_{d+2,m}) = p(\tilde{u} \,|\, u_{d+2,m})$. The remainder of the proof is the same as the one above for the linear transformation families.

For multivariate location with univariate scale families, define a scalar random variable W which is the first coordinate of the vector $Z_2 - Z_1$. The last $d-1$ coordinates divided by W is defined to be V (thus $(1, V) = (Z_2 - Z_1)/W$) and we define

$$\tilde{U} = \frac{\tilde{Z} - Z_1}{W}, \; U_i = \frac{Z_i - Z_1}{W}, \; i = 3, \ldots, m. \tag{2.14}$$

After applying the invariance property with $a = -Z_1$ and $b = 1/W$, it turns out that the best invariant estimator q^* satisfies $q^*(\tilde{u} \,|\, 0, (1, v), u_{3,m}) = p(\tilde{u} \,|\, v, u_{3,m})$. The joint density for $V, U_{3,m}$ and $\tilde{U}$ is given by

$$\int |w|^{(m+n-1)d-1} p_{Z,\tilde{Z}}(z_1, (w, wv) + z_1, wu_3 + z_1, \ldots, wu_m + z_1, w\tilde{u} + z_1) dz_1 dw. \tag{2.15}$$

Let b be the first coordinate of vector $Y_2 - Y_1$. Note V, $\tilde{U}$ and U_i's are invariant to location shift and univariate scale, that is, V is equal to the last $d-1$ coordinates of $Y_2 - Y_1$ divided by b and

$$\tilde{U} = \frac{\tilde{Y} - Y_1}{b}, \ U_i = \frac{Y_i - Y_1}{b}, \ i = 3, \ldots, m.$$

Plug them back into equation (2.15) and change variable w to θ_2 with $w = \theta_2 b$, then (2.15) is equal to

$$\int |\theta_2|^{(m+n-1)d-1} |b|^{(m+n-1)d} p_{Z,\tilde{Z}}(z_1, \theta_2(y_2-y_1)+z_1, \ldots, \theta_2(y_m-y_1)+z_1, \theta_2(\tilde{y}-y_1)+z_1) dz_1 d\theta_2.$$

Change variable again with $\theta_1 = y_1 - z_1/\theta_2$ whose Jacobian is $|\theta_2|^d$ to obtain

$$\begin{aligned} & \int |\theta_2|^{(m+n)d-1} |b|^{(m+n-1)d} p_{Z,\tilde{Z}}(\theta_2(y - \theta_1), \theta_2(\tilde{y} - \theta_1)) d\theta_1 d\theta_2 \\ = \ & |b|^{(m+n-1)d} \int \frac{1}{|\theta_2|} p(y, \tilde{y} \,|\, \theta_2, \theta_1) d\theta_1 d\theta_2. \end{aligned}$$

The rest of the proof is then the same as we have given for other transformation families. □

2.3 Examples

The best invariant estimator q^* is calculated for some examples in which we have m observations $Y_1, \ldots, Y_m$ and want to estimate the density for the next observation $\tilde{Y}$. Let $Y_{(i)}$ be the i^{th} order statistic (the i^{th} smallest value) among $Y_1, \ldots, Y_m$.

Shifted exponential family: $p(\tilde{y} \,|\, \theta) = \exp(-(\tilde{y} - \theta)) 1_{\{y \geq \theta\}}$.

$$q^*(\tilde{y} \,|\, Y_1, \ldots, Y_m) = \begin{cases} \frac{m}{m+1} e^{-(\tilde{y} - Y_{(1)})} & \text{if } \tilde{y} \geq Y_{(1)} \\ \frac{m}{m+1} e^{-m(Y_{(1)} - \tilde{y})} & \text{if } \tilde{y} < Y_{(1)} \end{cases}$$

In Figure 2.1, the true density is plotted in the solid line and the three crosses indicate the three observations. We know that the Maximal Likelihood Estimator (MLE) for the location parameter is equal to the smallest observation

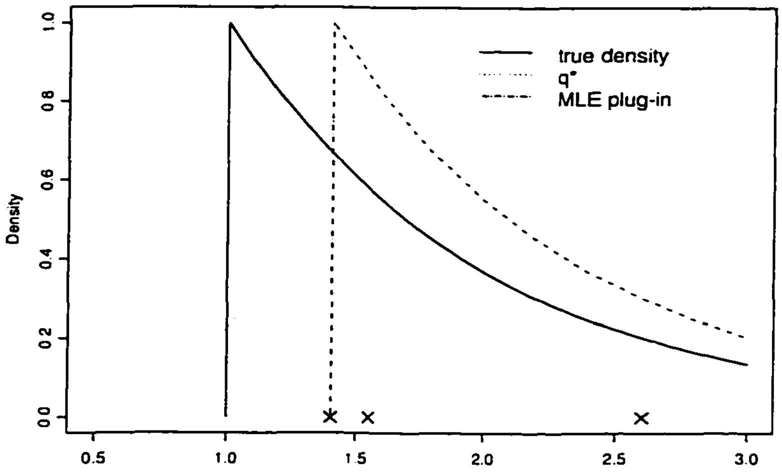

Figure 2.1: Plot of true density vs. q^* for shifted exponential family.

$Y_{(1)}$. The corresponding MLE plug-in estimator for the density is plotted in the dash-dot line. We can see there is a gap between the true density and the MLE plug-in estimator, which causes infinite loss. Some calculations reveal that the best invariant estimator q^* has finite risk equal to $\log(1 + \frac{1}{m})$. To avoid the infinite loss, the best invariant estimator q^*, plotted in the dashed line in Figure 2.1, distributes a small portion $(\frac{1}{m+1})$ of the total mass on the left of $Y_{(1)}$ and puts the remaining mass on the right. This is an example in which the optimal (best among invariant estimators) estimator is not in the same parametric family as the truth.

Uniform family (with scale parameter): $p(\tilde{y} \mid \theta) = |\theta| 1_{\{0 \leq \theta \tilde{y} \leq 1\}}$. Even though θ can take any value in $\mathbb{R}$ except 0, we will know θ is positive or negative once one observation is given. Here suppose Y_1 is positive, then θ is ranging from 0 to

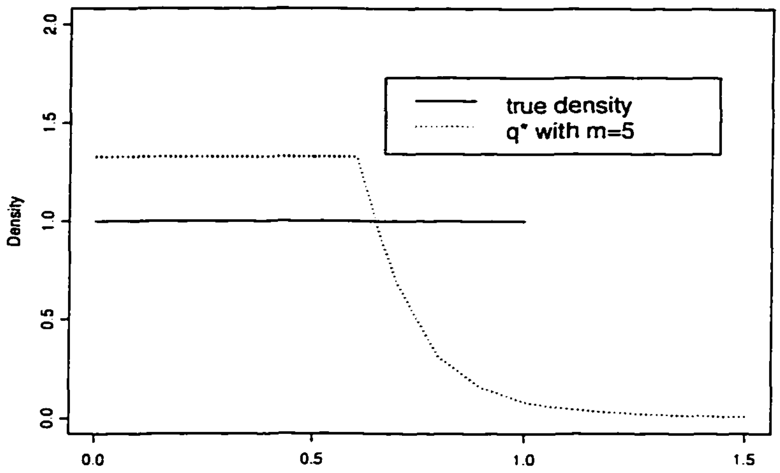

Figure 2.2: Plot of true density ($\theta = 1$) vs. q^* ($m = 5$) for uniform family with scale parameter.

∞.

$$
\begin{aligned}
q^*(\tilde{y} \mid Y_1, \ldots, Y_m) &= \frac{\int p(\tilde{y}, Y_1, \ldots, Y_m \mid \theta)\frac{1}{\theta}d\theta}{\int p(Y_1, \ldots, Y_m \mid \theta)\frac{1}{\theta}d\theta} \\
&= \begin{cases} \frac{m}{m+1}\frac{(Y_{(m)})^m}{\tilde{y}^{m+1}} & \text{if } \tilde{y} > Y_{(m)} \\ \frac{m}{m+1}\frac{1}{Y_{(m)}} & \text{if } \tilde{y} \le Y_{(m)} \end{cases}
\end{aligned}
$$

Figure 2.2 plots the true density (solid line) and the best invariant estimator q^* (dashed line) for $\theta = 1$ and $m = 5$.

Normal Location: Normal(θ, σ^2), θ unknown and σ^2 fixed with $p(y \mid \theta) = \phi_{\sigma^2}(y - \theta)$.

$$q^*(\tilde{y} \mid Y_1, \ldots, Y_m) = \phi_{\sigma^2(1+\frac{1}{m})}(\tilde{y} - \bar{Y}).$$

This is the normal density with mean $\bar{Y} = (1/m)\sum_{i=1}^m Y_i$ and a slightly larger variance $\sigma^2(1 + \frac{1}{m})$. The MLE plug-in estimator, a normal density with mean $\bar{Y}$, but variance σ^2, is also invariant. The risk for the MLE plug-in estimator is equal to $(\frac{d}{2})(\frac{1}{m})$, which is bigger than the risk of q^*, $(\frac{d}{2})\log(1 + \frac{1}{m})$.

Normal location and scale: Normal(θ, σ^2), $\theta \in \mathbb{R}^d$, $\sigma^2 \geq 0$, both unknown. The best invariant estimator is proportional to $(1 + \|\tilde{y} - \bar{y}\|^2/\hat{s}^2c)^{-md/2}$, that is

$$q^*(\tilde{y} \mid Y_1, \ldots, Y_m) = \frac{\Gamma(\frac{md}{2})}{\Gamma(\frac{(m-1)d}{2})} \frac{[(1+1/m)\hat{s}^2]^{-d/2}}{[\pi(m-1)d]^{d/2}} \left[\frac{1}{(m-1)d} \frac{\|\tilde{y} - \bar{Y}\|^2}{(1+1/m)\hat{s}^2} + 1\right]^{-md/2}$$

where $\hat{s}^2 = \sum_{i=1}^m \|Y_i - \bar{Y}\|^2/((m-1)d)$ is the sample variance. Thus $T = (\tilde{Y} - \bar{Y})/[(1+\frac{1}{m})\hat{s}^2]^{1/2}$ is assigned a predictive distribution which is the multivariate t distribution with $(m-1)d$ degrees of freedom.

Uniform on Parallelograms: $p(\tilde{y} \mid \theta_1, \theta_2) = |\theta_2| 1_{(0,1)\times(0,1)}(\theta_2(\tilde{y} + \theta_1))$, where $\theta_1 \in \mathbb{R}^2$ and θ_2 is a 2×2 matrix with determinant not equal to 0. Conditioning on at least three observations, one can show that the best invariant density estimation q^* is constant in the convex hull spanned by the observations, and tapers down toward zero as one moves away from the convex hull.

2.4 Discussion

In this section, we will briefly review some results about invariant decision problems through groups of transformations. For more detail, please refer to [9][25][6].

In a decision problem, we have a sample space $\mathcal{X}$, a family of densities with parameter space Θ and an action space $\mathcal{A}$. Suppose there is a group of transformations (one-to-one and onto) G on the sample space, that is, all the transformations from G form a group using the usual *composition* as the group operator.

The parameter space is said to be *invariant under the group* G if , for every $g \in G$ and $\theta \in \Theta$, there exists an unique $\theta^* \in \Theta$ such that the corresponding density for $g(Y)$ is $p(y \mid \theta^*)$. We can denote θ^* by $\bar{g}(\theta)$, then $\bar{G} = \{\bar{g} : g \in G\}$ is the induced group of transformations on Θ into itself. Consequently, the following two equalities hold true:

$$P_\theta(g(X) \in A) = P_{\bar{g}(\theta)}(X \in A)$$

and

$$\mathbb{E}_{\theta}[f(g(X))] = \mathbb{E}_{\bar{g}(\theta)}[f(X)].$$

A loss function $L(\theta, a)$ is said to be *invariant under the group G* if , for every $g \in G$ and $a \in \mathcal{A}$, there exists an $a^* \in \mathcal{A}$ such that $L(\theta, a) = L(\bar{g}(\theta), a^*)$ for all θ. The action a^* is denoted by $\tilde{g}(a)$, and then $\tilde{G} = \{\tilde{g} : g \in G\}$ is a group of transformation of the action space $\mathcal{A}$ into itself.

Now we restate some of the invariant decision problems we considered in this Chapter using the idea of transformation groups.

Example 1 For location family with $\mathcal{X} = \mathbb{R}^d$, consider the transformation group $G = \{g_c : c \in \mathbb{R}^d\}$, where $g_c(x) = x + c$. This group is called the *additive group* or *location group*. The corresponding transformation on the parameter space is $\bar{g}_c(\theta) = \theta + c$, and the one on the action space is $\tilde{g}_c(q) = q(\tilde{y} - c; y)$, since

$$\begin{aligned} L(\bar{g}_c(\theta), \tilde{g}_c(q_{\tilde{Y}|y})) &= \mathbb{E}_{\tilde{Y}|\theta+c} \log \frac{p(\tilde{Y} \mid y, \theta + c)}{q(\tilde{Y} - c; y)} \\ &= \mathbb{E}_{\tilde{Y}-c|\theta} \log \frac{p(\tilde{Y} - c \mid y, \theta)}{q(\tilde{Y} - c; y)} \\ &= \mathbb{E}_{X|\theta} \log \frac{p(X \mid y, \theta)}{q(X; y)} \\ &= L(\theta, q), \end{aligned}$$

where we change the integration variable $\tilde{Y} - c$ to X at the third equality. So this decision problem is invariant under the location group.

Example 2 For an affine family with $\mathcal{X} = \mathbb{R}$, consider the transformation group $G = \{g_{b,c}(x) = bx + c, b \in \mathbb{R}, c \in \mathbb{R}\}$. It can be checked that the decision problem for the affine family is invariant under that group.

Next we give the definition of right Haar density.

Definition 3 *A density v_r is a right Haar density on $\bar{G}$ if for any set $A \in \bar{G}$ and all*

$\bar{g}_0 \subset \bar{G}$, it satisfies

$$\int_{A\bar{g}_0} v_r(y)dy = \int_A v_r(y)dy.$$

Similarly, we can define left Haar density (measure).

For example, the uniform is the right Haar measure for the location group since if we shift a set A to $A + c$, its measure is unchanged.

The best invariant estimators, we identified in this Chapter, are generalized Bayes estimators with improper priors which are made proper by conditioning. Those improper priors are the same as the right Haar measures for the corresponding transformation groups on the parameter space Θ which leave the decision problem invariant. This is not a coincidence. It is proved that under some conditions the best invariant estimator is the generalized Bayes estimator using the right Haar measure (Berger [6]). The calculations in Section 1 and 2 in this Chapter, which follow Pitman's technique, provide a way to understand this general result without any knowledge of group theory.

Chapter 3

Minimax Estimators

Since the risk is constant for invariant predictive density estimators, the best invariant estimator q^* is the minimax procedure among all invariant procedures. Hunt-Stein theory provides a means by which to show that under some conditions the best invariant rule is in fact minimax over all rules, and this strategy has proved effectively in parameter estimation and hypothesis testing [24][14]. The same technique might be carried over to prove the same conjecture for predictive density estimation. In this paper, we provide a proof based on the fact from decision theory that constant risk plus extended Bayes implies minimax (see Appendix A). We use tools from Information Theory to confirm that our best invariant procedure, which is known to have constant risk, is extended Bayes and hence minimax.

3.1 Location Families

We first work with location families. Recall a location family has observations $Y_i = Z_i + \theta$ with $i = 1, \ldots, m$ and future data $\tilde{Y}_j = \tilde{Z}_j + \theta$ with $j = 1, \ldots, n$, where θ is unknown and we assume $Z = (Z_1, \ldots, Z_m)$ and $\tilde{Z} = (\tilde{Z}_1, \ldots, \tilde{Z}_n)$ have a known joint density $p_{Z,\tilde{Z}}$. Then the joint density for Y and $\tilde{Y}$ is given by $p(y, \tilde{y} \,|\, \theta) = p_{Z,\tilde{Z}}(y - \theta, \tilde{y} - \theta)$.

3.1.1 Proof for Minimaxity

Definition 4 *A predictive procedure q is called extended Bayes, if there exists a sequence of Bayes procedures $\{p_{w_k}\}$ with proper priors w_k such that their Bayes risk differences go to zero, that is,*

$$R_{w_k}(q) - R_{w_k}(p_{w_k}) \to 0, \text{ as } k \to \infty.$$

Theorem 1 *Assume for the location family that at least one of the $Z_1, \ldots, Z_m$ has finite second moment. Then, under Kullback-Leibler loss, the best invariant predictive procedure*

$$q^*(\tilde{y} \mid y) = \frac{\int p(y, \tilde{y} \mid \theta) d\theta}{\int p(y \mid \theta) d\theta}$$

is minimax for any dimension d.

Proof: To show minimaxity, it suffices to show that q^*, known to have constant risk, is extended Bayes. We take a sequence of priors to be the normal distributions with mean zero and variance k. Recall that the corresponding Bayes predictive procedure p_{w_k} is defined by (1.3).

Examine the Bayes risk difference between q^* and p_{w_k}.

$$\begin{aligned} R_{w_k}(q^*) - R_{w_k}(p_{w_k}) &= \int \left[R(\theta, q^*) - R(\theta, p_{w_k})\right] w_k(\theta) d\theta \\ &= \mathbb{E}_{Y,\tilde{Y}} \log \frac{p_{w_k}(\tilde{Y} \mid Y_1, \ldots, Y_m)}{q^*(\tilde{Y} \mid Y_1, \ldots, Y_m)}, \end{aligned}$$

where in the expectation $\mathbb{E}_{Y,\tilde{Y}}$, the distribution of $(Y, \tilde{Y})$ is a mixture with respect to prior w_k.

By the chain rule of Information Theory, the Bayes risk difference

$$\mathbb{E}_{Y,\tilde{Y}} \log \frac{p_{w_k}(\tilde{Y} \mid Y_1, \ldots, Y_m)}{q^*(\tilde{Y} \mid Y_1, \ldots, Y_m)}$$

is less than or equal to the following total Bayes risk difference (conditioning only on Y_1).

$$\begin{aligned}
& \mathbb{E}_{Y,\tilde{Y}} \log \frac{p_{w_k}(Y_2 \ldots, Y_m, \tilde{Y} \mid Y_1)}{q^*(Y_2, \ldots, Y_m, \tilde{Y} \mid Y_1)} \\
= \; & \mathbb{E}_{Y,\tilde{Y}} \Big[-\log \frac{\int p(Y_1, \ldots, Y_m, \tilde{Y} \mid \theta) w_k(\theta) \frac{1}{w_k(\theta)} d\theta}{\int p(Y_1, \ldots, Y_m, \tilde{Y} \mid \theta) w_k(\theta) d\theta} - \log \frac{\int p(Y_1 \mid \theta') w_k(\theta') d\theta'}{\int p(Y_1 \mid \theta') d\theta'} \Big] \\
= \; & \mathbb{E}_{Y,\tilde{Y}} \Big[-\log \mathbb{E}_{\theta \mid Y,\tilde{Y}} \Big(\frac{1}{w_k(\theta)}\Big) - \log \int p(Y_1 \mid \theta') w_k(\theta') d\theta' \Big]
\end{aligned}$$

where we have used that $\int p(Y_1 \mid \theta') d\theta' = 1$. The variable on which to condition is chosen to be one for which the variance is finite (here Y_1, without loss of generality).

Invoking Jensen's inequality in both terms (using convexity of $-\log$), we get the Bayes risk difference is less than or equal to

$$\begin{aligned}
& \mathbb{E}_\theta \log w_k(\theta) - \mathbb{E}_{Y_1} \int p(Y_1 \mid \theta') \log w_k(\theta') d\theta' \\
= \; & \int w_k(\theta) \log w_k(\theta) d\theta - \int\int w_k(\theta) p(y_1 - \theta) p(y_1 - \theta') \log \frac{1}{w_k(\theta')} d\theta' dy_1 d\theta \quad (3.1)
\end{aligned}$$

where $\int w_k(\theta) p(y_1 - \theta) d\theta$ in the second term is the mixture giving the distribution of Y_1. Next we do a change of variables where for each θ, we replace y_1 and θ' with $z_1 = y_1 - \theta$ and $z_1' = y_1 - \theta'$, which have unit Jacobians. So (3.1) becomes

$$\begin{aligned}
& \int w_k(\theta) \log w_k(\theta) d\theta - \int\int w_k(\theta) p(z_1') p(z_1) \log \frac{1}{w_k(\theta + z_1 - z_1')} dz_1' dz_1 d\theta \\
= \; & \mathbb{E}_{Z_1, Z_1', \theta} \log \frac{w_k(\theta)}{w_k(\theta + Z_1 - Z_1')} \qquad (3.2) \\
= \; & \mathbb{E}_{Z_1, Z_1', \theta} \frac{\|\theta + Z_1 - Z_1'\|^2 - \|\theta\|^2}{2k} \\
= \; & \mathbb{E}_{Z_1, Z_1'} \frac{\|Z_1 - Z_1'\|^2}{2k} = \frac{\mathbb{E}\|Z_1\|^2}{k}.
\end{aligned}$$

Thus $R_{w_k}(q^*) - R_{w_k}(p_{w_k})$ is made arbitrary small for large k. So q^* is extended Bayes, and therefore minimax (as per Lemma 2 of Appendix A). □

Remark: A similar but more involved argument using prior $w_k(\theta)$ with tails that decay at a polynomial rather than exponential rate (e.g. Cauchy priors) shows that finite logarithmic moment (that is, $\mathbb{E}\log(1+|Z_i|)$ finite for some i) is sufficient for minimaxity of the best invariant rule (see Appendix B).

3.1.2 Admissibility and Inadmissibility

The proof for minimaxity already implies the admissibility of q^* in one dimension.

Theorem 2 *Assume for a location family on $\mathbb{R}$ that at least one of the $Z_1, \ldots, Z_m$ has finite second moment. Then*

$$q^*(\tilde{y} \mid y) = \frac{\int p(y, \tilde{y} \mid \theta) d\theta}{\int p(y \mid \theta) d\theta}$$

is admissible under Kullback-Leibler loss.

Proof: Sufficient conditions for admissibility are summarized in Lemma 3 in Appendix C. Choose π_n to be the unnormalized normal density

$$\pi_k = \frac{1}{\sqrt{2\pi}} \exp\{\frac{-\theta^2}{k}\}.$$

Notice that π_k is bounded below by π_1. Therefore for any nondegenerate convex set $C \in \Theta$,

$$\int_C \pi_n(\theta) d\theta \geq \int_C \pi_1(\theta) d\theta = K > 0.$$

So condition (b) in Lemma 3 is satisfied by this choice of π_k. (Note that the standard $N(0, k)$ densities, which are equal to $\pi_k/\sqrt{k}$, would not satisfy this condition.) Some calculation reveals that the Bayes risks $R_{\pi_k}(q^*)$ and $R_{\pi_k}(p_{\pi_k})$ are finite where p_{π_k} are the corresponding Bayes estimators with respect to prior π_k. Therefore condition (a) is also satisfied. In our proof for minimaxity of q^*, we have already showed that the Bayes risk difference is bounded by $\mathbb{E}Z_1^2/k$ using the standard normal prior $w_k = \pi_k/\sqrt{k}$. So

$$R_{\pi_k}(q^*) - R_{\pi_k}(p_{\pi_k}) \leq (\sqrt{k}) \frac{\mathbb{E}Z_1^2}{k},$$

which goes to zero when k goes to ∞. Condition (c) is verified. Thus q^* is admissible in one dimension. $\square$

Remark: Apparently, the same trick (choice of priors π_k) is going to fail when the dimension is bigger than one. Based on the parallel result for point estimation, it might be true that q^* is also admissible in two dimension. But we think the proof will involve a sequence of more delicate priors.

Let us consider a normal location family and focus on the density estimation for only one future observation $\tilde{y}$. As we mentioned before, the minimax estimator (also the best invariant with constant risk) q^* is reduced to normal density with mean $\bar{y}_m$ and a slightly larger variance $\sigma^2(1+\frac{1}{m})$. We are going to show the inadmissibility of q^* when dimension is three or higher ($\mathbf{d \geq 3}$).

Consider a special estimator q which is a normal density with mean $T(y)$ and variance $\sigma^2(1+\frac{1}{m})$, i.e.

$$q(\tilde{y} \mid y_1, \ldots, y_m) = \phi_{\sigma^2(1+\frac{1}{m})}(\tilde{y} - T(y)),$$

where $T(y)$ is a function of the sample $y_1, \ldots, y_m$. For example, if $T(y) = \bar{y}_m$, the mean of the sample, then q is just equal to q^*. We are going to show that estimator q will has smaller risk than q^* by some choices of $T(\cdot)$.

For any θ, the risk difference between q and q^* is given by

$$\begin{aligned} R(\theta, q) - R(\theta, q^*) &= \mathbb{E}_{Y,\tilde{Y}|\theta} \log \frac{p(\tilde{Y} \mid \theta)}{q(\tilde{Y} \mid Y)} - \mathbb{E}_{Y,\tilde{Y}|\theta} \log \frac{p(\tilde{Y} \mid \theta)}{q^*(\tilde{Y} \mid Y)} \\ &= \mathbb{E}_{Y,\tilde{Y}|\theta} \log \frac{q^*(\tilde{Y} \mid Y)}{q(\tilde{Y} \mid Y)}. \end{aligned}$$

Due to the special form of normal density, the risk difference is equal to

$$\begin{aligned} &\frac{1}{2\sigma^2(1+\frac{1}{m})} \mathbb{E}_{Y,\tilde{Y}|\theta}(\|\tilde{Y} - \bar{Y}\|^2 - \|\tilde{Y} - T(Y)\|^2) \\ = &\frac{1}{2\sigma^2(1+\frac{1}{m})} \mathbb{E}_{Y|\theta}(\|\bar{Y} - \theta\|^2 - \|T(Y) - \theta\|^2). \qquad (3.3) \end{aligned}$$

Notice that (3.3) is proportional to the risk difference in parameter estimation under mean squared loss. So if $T(Y)$, as a point estimation for θ, has smaller mean squared risk than the sample mean $\bar{Y}$, then the predictive density estimator q which is $N(T(y), \sigma^2(1 + \frac{1}{m}))$, has smaller Kullback-Leibler risk than q^*. Apparently, when dimension is three or higher, such an estimator $T(Y)$ does exist, such as Stein's shrinkage estimator [21] and Strawderman's proper Bayes estimator [22]. So q^* is inadmissible when dimension $d \geq 3$ for normal location families.

3.2 Other Transformation Families

Next we consider minimaxity for other groups. For linear transformation and affine families, the best invariant procedure uses a prior $1/|\theta|^d$ which is not only improper, but also hard to be approximated by sequences of proper priors when $d > 1$. Nevertheless, the cases of univariate scale (Theorem 3) and multivariate location with univariate scale (Theorem 4) can be handled by our technique.

Theorem 3 *Assume for the scale family (i.e. general linear transformation family with $d = 1$ and $\theta \neq 0$) that there exists $i \in \{1, \ldots, m\}$ such that $\log(|Z_i|)$ is integrable. Then, under the Kullback-Leibler loss, the best invariant predictive procedure*

$$q^*(\tilde{y} \mid y) = \frac{\int \frac{1}{|\theta|} p(y, \tilde{y} \mid \theta) d\theta}{\int \frac{1}{|\theta|} p(y \mid \theta) d\theta}$$

is minimax.

Proof: To show that q^* is extended Bayes, we take a sequence of proper priors to be $w_k(\theta)$ proportional to $\min(|\theta|^{-1-\alpha_k}, |\theta|^{-1+\alpha_k})$, where $\alpha_k > 0$. For α_k small, these priors have behavior close to that of improper prior $w(\theta) = |\theta|^{-1}$.

By the chain rule of Information Theory, the Bayes risk difference $R_{w_k}(q^*) - R_{w_k}(p_{w_k})$ is less than or equal to the Bayes risk difference conditioning only on Y_1.

$$
\begin{aligned}
& \mathbb{E}_{Y,\tilde{Y}} \log \frac{p_{w_k}(Y_2 \dots, Y_m, \tilde{Y} \mid Y_1)}{q^*(Y_2, \dots, Y_m, \tilde{Y} \mid Y_1)} \\
= \; & \mathbb{E}_{Y,\tilde{Y}} \big[-\log \frac{\int p(Y, \tilde{Y} \mid \theta) w_k(\theta) \frac{w(\theta)}{w_k(\theta)} d\theta}{\int p(Y, \tilde{Y} \mid \theta) w_k(\theta) d\theta} - \log \frac{\int p(Y_1 \mid \theta') w(\theta') \frac{w_k(\theta')}{w(\theta')} d\theta'}{\int p(Y_1 \mid \theta') w(\theta') d\theta'} \big] \\
= \; & \mathbb{E}^{w_k}_{Y,\tilde{Y}} \big[-\log \mathbb{E}^{w_k}_{\theta \mid Y,\tilde{Y}} \frac{w(\theta)}{w_k(\theta)} - \log \mathbb{E}^{w}_{\theta' \mid Y_1} \frac{w_k(\theta')}{w(\theta')} \big], \qquad (3.4)
\end{aligned}
$$

where all the superscripts on $\mathbb{E}$ indicate the corresponding priors on θ for those marginal or posterior distributions, for example, $\mathbb{E}^{w_k}_{\theta \mid Y,\tilde{Y}}$ is the posterior expectation when the prior is $w_k(\theta)$ and $\mathbb{E}^{w}_{\theta' \mid Y_1}$ is the posterior expectation (given only Y_1) when the prior is $w(\theta)$. The outer expectation $\mathbb{E}^{w_k}_{Y,\tilde{Y}}$ is taken with respect to the marginal distribution of $(Y, \tilde{Y})$ when θ has prior $w_k(\theta)$. By Jensen's inequality, we have (3.4) is less or equal to

$$
\mathbb{E}^{w_k}_{\theta} \log \frac{w_k(\theta)}{w(\theta)} - \mathbb{E}^{w_k}_{Y_1} \mathbb{E}^{w}_{\theta' \mid Y_1} \log \frac{w_k(\theta')}{w(\theta')}. \qquad (3.5)
$$

For given y_1, the density of θ' is proportional to $\frac{1}{|\theta'|} p(y_1 \mid \theta') = p(y_1 \theta')$. We change variable θ' to $z_1' = y_1 \theta'$ which has Jacobian y_1, then, with y_1 fixed, the density for Z_1' is indeed $p(z_1')$ independent of y_1. Also replace y_1 by z_1 with $z_1 = \theta y_1$, then (3.5) is equal to

$$
\begin{aligned}
& \mathbb{E}_{Z_1, Z_1', \theta} \log \frac{|\theta| w_k(\theta)}{|\theta| \frac{|Z_1'|}{|Z_1|} w_k(\theta \frac{Z_1'}{Z_1})} \\
= \; & \mathbb{E}_{Z_1, Z_1', \theta} \min(-\alpha_k \log |\theta|, \; \alpha_k \log |\theta|) \\
& \quad - \min\big(-\alpha_k \log |\theta| - \alpha_k \log \frac{|Z_1'|}{|Z_1|}, \; \alpha_k \log |\theta| + \alpha_k \log \frac{|Z_1'|}{|Z_1|} \big). \qquad (3.6)
\end{aligned}
$$

Use the inequality: $\min(a, -a) - \min(-a - b, a + b) \leq |b|$, then (3.6) is less than or equal to $\alpha_k \mathbb{E} \; | \log \frac{|Z_1'|}{|Z_1|} |$, which goes to zero when α_k goes to zero by our assumption. □

One can see the same technique is used in deriving the upper bounds for the Bayes risk differences in the proofs for Theorems 1 and 2. This technique turns out to be very useful for Theorems 3, 4 and 5 as well. We summarize a key step in this technique as a more general lemma.

Lemma 1 [Bayes Risk Difference Bound]: *Suppose there is a parametric family* $\{p(y, \tilde{y} \mid \theta) : \theta \in \Theta\}$. *Let* v *and* w *be two priors (*v *proper,* w *possibly improper) on* θ *and let* $u = f(y)$ *be a function of* y *with density* $p_U(u \mid \theta)$ *for which the posterior* $w(\theta \mid u)$ *is proper, that is,* $\int p_U(u \mid \theta) w(\theta) d\theta$ *is finite for all* u. *Then the Bayes risk difference satisfies the following inequality:*

$$R_v(p_w) - R_v(p_v) \leq \mathbb{E}_\theta^v \mathbb{E}_{U|\theta} \mathbb{E}_{\theta'|U}^w \log \frac{v(\theta)/w(\theta)}{v(\theta')/w(\theta')},$$

where $\mathbb{E}_{\theta'|U}^w$ *denotes the expectation with respect to the posterior of* θ' *given* U *when* θ' *has prior* w *and* $\mathbb{E}_\theta^v$ *denotes the expectation with respect to the prior* v *on* θ.

Proof: By definition, the risk difference $R_v(p_w) - R_v(p_v)$ is equal to

$$\begin{aligned} & \mathbb{E}_\theta^v \left[\mathbb{E}_{Y,\tilde{Y}|\theta} \log \frac{p(\tilde{Y} \mid \theta)}{p_w(\tilde{Y} \mid Y)} - \mathbb{E}_{Y,\tilde{Y}|\theta} \log \frac{p(\tilde{Y} \mid \theta)}{p_v(\tilde{Y} \mid Y)}\right] \\ = \; & \mathbb{E}_{Y,\tilde{Y}}^v \log \frac{p_v(\tilde{Y} \mid Y)}{p_w(\tilde{Y} \mid Y)} = \mathbb{E}_{Y,\tilde{Y}}^v \log \frac{p_v(Y, \tilde{Y})}{p_w(Y, \tilde{Y})} - \mathbb{E}_Y^v \log \frac{p_v(Y)}{p_w(Y)}. \end{aligned} \tag{3.7}$$

Similarly to the proof for Theorems 1 and 2, we express the first term of (3.7) as a conditional expectation and then apply Jensen's inequality using the convexity of $-\log$.

$$\begin{aligned} \mathbb{E}_{Y,\tilde{Y}}^v \log \frac{p_v(Y, \tilde{Y})}{p_w(Y, \tilde{Y})} &= \mathbb{E}_{Y,\tilde{Y}}^v \Big(-\log \frac{\int p(Y, \tilde{Y} \mid \theta) v(\theta) \frac{w(\theta)}{v(\theta)} d\theta}{\int p(Y, \tilde{Y} \mid \theta) v(\theta) d\theta} \Big) \\ &\leq \mathbb{E}_{Y,\tilde{Y}}^v \mathbb{E}_{\theta|Y,\tilde{Y}}^v \Big(-\log \frac{w(\theta)}{v(\theta)} \Big) = \mathbb{E}_\theta^v \log \frac{v(\theta)}{w(\theta)}. \end{aligned}$$

The second term of (3.7), $\mathbb{E}_Y^v \log p_v(y)/p_w(y)$, is the Kullback-Leibler divergence between densities p_v and p_w. Recall the following result from Information Theory:

let p_Y, q_Y be two densities and u is a function of y with corresponding densities p_U and q_U, then

$$D(p_Y\|q_Y) \geq D(p_U\|q_U). \tag{3.8}$$

To prove the inequality, consider the Kullback-Leibler divergence between the joint densities $p_{Y,U}$ and $q_{Y,U}$, $D(p_{Y,U}\|q_{Y,U})$, which is equal to $\mathbb{E}_U D(p_{Y|U}\|q_{Y|U}) + D(p_U\|q_U)$. On the other hand,

$$D(p_{Y,U}\|q_{Y,U}) = \mathbb{E}_Y D(p_{U|Y}\|q_{U|Y}) + D(p_Y\|q_Y) = D(p_Y\|q_Y),$$

since U is the function of Y. Therefore $D(p_Y\|q_Y) \geq D(p_U\|q_U)$ by the non-negativity of the Kullback-Leibler divergence.

Let p_v, p_w be the p, q in inequality (3.8) and then

$$\begin{aligned} \mathbb{E}^v_Y \log \frac{p_v(Y)}{p_w(Y)} &\geq \mathbb{E}^v_U \log \frac{p_v(U)}{p_w(U)} = \mathbb{E}^v_U \log \frac{\int p(U \mid \theta')w(\theta')\frac{v(\theta')}{w(\theta')}d\theta'}{\int p(U \mid \theta')w(\theta')d\theta'} \\ &\geq \mathbb{E}^v_U \mathbb{E}^w_{\theta'|U} \log \frac{v(\theta')}{w(\theta')}, \end{aligned}$$

where Jensen's inequality is applied at the last step.

Combining all the steps after equation (3.7), we have the Bayes risk difference is less than or equal to

$$\mathbb{E}^v_\theta \log \frac{v(\theta)}{w(\theta)} - \mathbb{E}^v_U \mathbb{E}^w_{\theta'|U} \log \frac{v(\theta')}{w(\theta')},$$

which completes the proof. □

Theorem 4 *For the multivariate location with univariate scale family, conditioning on at least two observations ($m \geq 2$), assume that there exist $i, j \in \{1, \ldots, m\}$ and $k \in \{1, \ldots, d\}$ such that $\log(|Z_{ik} - Z_{jk}|)$, $\log\left(1 + \left|\frac{Z_{ik}-Z_{jk}}{Z'_{ik}-Z'_{jk}}\right|\right)$ and $\log(1 + \|Z_i\|)$ are integrable, where Z'_i and Z'_j are independent copies of Z_i and Z_j, respectively, and Z_{ik}*

denotes the k^{th} coordinate of the d-dimensional vector Z_i. Then, under the Kullback-Leibler loss, the best invariant predictive procedure

$$q^*(\tilde{y} \mid y) = \frac{\iint \frac{1}{|\theta_2|} p(y, \tilde{y} \mid \theta_1, \theta_2) d\theta_1 d\theta_2}{\iint \frac{1}{|\theta_2|} p(y \mid \theta_1, \theta_2) d\theta_1 d\theta_2}$$

is minimax.

Proof: We take the proper prior $w_k(\theta_1, \theta_2)$ to be the product of priors on θ_1 and θ_2 which we used in the proofs for location families (Appendix B, Theorem 1′) and scale families (Theorem 2). That is, $w_k(\theta_1, \theta_2) = w_k^{(1)}(\theta_1) w_k^{(2)}(\theta_2)$ and

$$w_k^{(1)}(\theta_1) \sim \frac{1}{(1 + \|\theta_1\|/k)^{d+1}}, \quad w_k^{(2)}(\theta_2) \sim \min(|\theta_2|^{-1-\alpha_k}, |\theta_2|^{-1+\alpha_k}). \tag{3.9}$$

This provides our sequence of proper priors with behavior close to that of the improper prior $w(\theta_1, \theta_2) = 1/|\theta_2|$.

Without loss of generality, we assume the indices i, j and k in the assumption are equal to 1, 2 and 1. Apply Lemma 1 with $u = (y_1, y_{21})$ and $v = w_k$, where y_{21} is the first coordinate of y_2. Then the Bayes risk difference $R_{w_k}(p^*) - R_{w_k}(p_{w_k})$ is less than or equal to

$$\mathbf{E}_\theta^{w_k} \mathbf{E}_{Y_1, Y_{21} \mid \theta} \mathbf{E}^{w}_{\theta' \mid Y_1, Y_{21}} \log \frac{w_k(\theta)|\theta_2|}{w_k(\theta')|\theta_2'|}. \tag{3.10}$$

In a manner similar to the previous proofs, for given y_1 and y_{21}, we change variable (θ_1', θ_2') to (z_1', z_{21}') with

$$\left\{ \begin{array}{lcl} z_1' & = & \theta_2'(y_1 - \theta_1') \\ z_{21}' & = & \theta_2'(y_{21} - \theta_{11}') \end{array} \right. \Rightarrow \left\{ \begin{array}{lcl} \theta_1' & = & y_1 - z_1' \frac{y_{11} - y_{21}}{z_{11}' - z_{21}'} \\ \theta_2' & = & \frac{z_{11}' - z_{21}'}{y_{11} - y_{21}} \end{array} \right. .$$

The corresponding Jacobian is equal to $|\theta_1'|^{-d} |y_{21} - y_{11}|^{-1}$. Do a change of variables with (y_1, y_{21}) replaced by $z_1 = \theta_2(y_1 - \theta_1)$ and $z_{21} = \theta_2(y_{21} - \theta_{11})$. We find that the joint density for (Z_1', Z_{21}') is independent of y_1, y_{21} and has the same distribution as

(Z_1, Z_{21}). Now (3.10) is equal to

$$\mathbb{E}_\theta^{w_k} \mathbb{E}_{Z_1, Z_{21}} \mathbb{E}_{Z_1', Z_{21}'} \Big[\log \frac{w_k^{(1)}(\theta_1)}{w_k^{(1)}(\frac{Z_1}{\theta_2} - \frac{Z_1'}{\theta_2} \frac{Z_{11}-Z_{21}}{Z_{11}'-Z_{21}'} + \theta_1)}$$

$$+ \log \frac{w_k^{(2)}(\theta_2)|\theta_2|}{w_k^{(2)}(\frac{Z_{11}'-Z_{21}'}{Z_{11}-Z_{21}}\theta_2)|\theta_2 \frac{Z_{11}'-Z_{21}'}{Z_{11}-Z_{21}}|} \Big].$$

By the proof for Theorem $1'$ (in Appendix B) and Theorem 2, we know that the quantities above go to zero provided that $\log(1 + \|\frac{Z_1}{\theta_2} - \frac{Z_1'}{\theta_2} \frac{Z_{11}-Z_{21}}{Z_{11}'-Z_{21}'}\|)$ and $\log(|Z_{21} - Z_{11}|)$ are integrable. Now

$$\mathbb{E} \log \Big(1 + \|\frac{Z_1}{\theta_2} - \frac{Z_1'}{\theta_2} \frac{Z_{11} - Z_{21}}{Z_{11}' - Z_{21}'}\|\Big) \le 2\mathbb{E} \log(1 + \|\frac{Z_1}{\theta_2}\|) + \mathbb{E} \log \Big(1 + |\frac{Z_{11} - Z_{21}}{Z_{11}' - Z_{21}'}|\Big),$$

where each term is finite by our assumptions. □

Use Theorems 1 and 3, it is easy to check that those best invariant estimators calculated in section 2 for normal families are minimax.

3.3 Minimal Conditioning Size

Next we show that the minimax risk is infinite without conditioning on enough initial observations. Here the minimal number of initial observations required is one for location or scale families, and two for multivariate location with univariate scale families.

Proposition 4 *For the location or scale families, the minimax risk (using Kullback-Leibler loss) is infinity if one does not condition on any observations. That is,*

$$\min_q \max_\theta D(p_{\tilde{Y}|\theta} \| q_{\tilde{Y}}) = \infty.$$

Proof: We first prove the conclusion for location families. Let $q(\tilde{y})$ denote any density estimator with risk

$$D(p_{\tilde{Y}|\theta}\|q_{\tilde{Y}}) = \mathbb{E}_{\tilde{Y}|\theta} \log \frac{p_{\tilde{Z}}(\tilde{Y}-\theta)}{q(\tilde{Y})} = \mathbb{E}_{\tilde{Z}} \log \frac{p_{\tilde{Z}}(\tilde{Z})}{q(\tilde{Z}+\theta)}.$$

Let q_θ denote the shifted density function $q(\cdot+\theta)$, then the risk is equal to $D(p_{\tilde{Z}}\|q_\theta)$.

Since q and $p_{\tilde{Z}}$ both integrate to one, there exists a ball $B(0,r)$ centering at origin with radius r, such that

$$P_{\tilde{Z}}(B) \geq 1-\epsilon \geq 1/2, \text{ and } Q(B) \geq 1-\epsilon.$$

Let $\theta = 2r$, then the shift of this ball $B(0,r)+\theta = B(2r,r)$ is in B^c. Therefore $Q_\theta(B) = Q(B+\theta) \leq \epsilon$. The divergence between distributions is at least as large as the divergence restricted to a partition [23]. Partitioning simply into $\{B, B^c\}$ yields

$$\begin{aligned} D(p_{\tilde{Z}}\|q_\theta) &\geq P_{\tilde{Z}}(B) \log \frac{P_{\tilde{Z}}(B)}{Q(B+\theta)} + P_{\tilde{Z}}(B^c) \log \frac{P_{\tilde{Z}}(B^c)}{Q(B^c+\theta)} \\ &= -\log 2 + \frac{1}{2} \log \frac{1}{\epsilon}. \end{aligned}$$

Letting $\epsilon \to 0$ yields $\sup_\theta D(p_{\tilde{Z}}\|q_\theta) = \infty$. Therefore the minimax risk is equal to ∞.

For scale families, we have $D(p_{\tilde{Y}|\theta}\|q) = D(p_{\tilde{Z}}\|q_\theta)$, where q_θ denotes the scaled density $q_\theta(y) = |\theta|^{-1} q(y/\theta)$. Since q is integrable, for any $\epsilon > 0$, there exits a δ such that for any measurable set A with measure less than δ, $Q(A) \leq \epsilon$. Consider a ball B with $P(B) > 1/2$. Let θ be a sufficiently large positive number such that the Lebesgue measure of B/θ is less than δ, then

$$\begin{aligned} D(p_{\tilde{Z}}\|q_\theta) &\geq P_{\tilde{Z}}(B) \log \frac{P_{\tilde{Z}}(B)}{Q_\theta(B)} + P_{\tilde{Z}}(B^c) \log \frac{P_{\tilde{Z}}(B^c)}{Q_\theta(B^c)} \\ &= P_{\tilde{Z}}(B) \log \frac{P_{\tilde{Z}}(B)\theta}{Q(B/\theta)} + P_{\tilde{Z}}(B^c) \log \frac{P_{\tilde{Z}}(B^c)\theta}{Q(B^c/\theta)} \\ &= -\log 2 + \frac{1}{2} \log \frac{1}{\epsilon} + \log \theta, \end{aligned}$$

which, as it shows for location families, means the minimax risk is infinity. □

Proposition 5 *For multivariate location with univariate scale families, the minimax risk (using Kullback-Leibler loss) is infinity if conditioning on less than two observations.*

Proof: When conditioning on no observations, the conclusion is a consequence of Proposition 4. Now we condition on only one observation. Suppose the minimax risk is finite, then there exists $q(\tilde{y} \,|\, y_1)$, such that for any $\theta = (\theta_1, \theta_2)$, the risk $\mathbb{E}_{\tilde{Y}, Y_1 | \theta} \log \left[p(\tilde{Y} \,|\, \theta) / q(\tilde{Y} \,|\, Y_1) \right]$ is bounded by some positive number M. Therefore for any θ, there exists a y_1, such that

$$\mathbb{E}_{\tilde{Y} | \theta} \log \frac{p(\tilde{Y} \,|\, \theta)}{q(\tilde{Y} \,|\, y_1)} \leq M. \tag{3.11}$$

Fixing y_1, we define a new variable $\tilde{X} = \tilde{Y} - y_1$. Its density is given by $|\theta_2|^d p(\theta_2(\tilde{x} + y_1 - \theta_1)) = |\theta_2|^d p(\theta_2(\tilde{x} + z_1))$ which only depends on the scale factor θ_2. The function $q(\tilde{y} \,|\, y_1)$ produces a predictive density for $\tilde{X}$ with $g_{y_1}(\tilde{x}) = q(\tilde{x} + y_1 \,|\, y_1)$. By changing variables, we can find that the risk $R(\theta_2, g_{y_1})$ is equal to the left side of (3.11) and hence bounded by M for any θ_2. But by Proposition 4, we know that $\max_{\theta_2} R(\theta_2, g_{y_1})$ is infinity, so the minimax risk is infinity when conditioning on only one observation. □

Remark: The minimal requirement for the conditioning size is the same as the one for the minimal training set in Berger and Pericchi's intrinsic Bayes factor [7][8] for the transformation groups discussed in this Chapter. In [8], the minimal training set is used for the convenience in computation of the Bayes factor.

3.4 Minimax Rule For Regression

We consider a linear regression model

$$\tilde{y}_i = \tilde{x}_{i1}\theta_1 + \cdots + \tilde{x}_{id}\theta_d + \tilde{z}_i = \tilde{x}_i^t \theta + \tilde{z}_i,$$

where $\tilde{x}_i = (\tilde{x}_{i1}, \dots, \tilde{x}_{id})$ is a d-dimensional input vector, and $\tilde{z}_i$ is the random error. Our interest is in finding the exact minimax coding strategy (or predictive density estimation) for linear regression models. We use $Y = (Y_1, \dots, Y_m)$ for the initial data, $\tilde{Y} = (\tilde{Y}_1, \dots, \tilde{Y}_n)$ for the data for which we want to predict the distribution, and $\tilde{Z}$, Z for the corresponding errors. Let $\tilde{x}$ denote the $d \times n$ matrix with $\tilde{x}_i$ as its i^{th} column. Same for x_i and x.

Assume $(\tilde{Z}, Z)$ is modeled by a distribution P with density p. Then the density for $(\tilde{Y}, Y)$ is given by

$$p_{\tilde{Y},Y|\theta}(\tilde{y}, y \,|\, \theta) = p(\tilde{y} - \tilde{x}^t\theta, y - x^t\theta), \quad \theta \in \mathbb{R}^d, \tag{3.12}$$

which is different from the ordinary location families we studied before, but similar analysis can be applied and it reveals that the exact minimax strategy is the Bayes procedure with uniform prior over the parameter space $\mathbb{R}^d$, conditioning on at least $m \geq d$ observations.

Theorem 5 *Assume that for the parametric family given in (3.12) with $m \geq d$ there exists a d-element subset from $(1, \dots, m)$, denoted by $(i_1, \dots, i_d)$, such that the d errors $(Z_{i_1}, \dots, Z_{i_d})$ have finite second moments and that the $d \times d$ matrix composed by the d vectors $x_{i_1}, \dots, x_{i_d}$ is non-singular. Then*

$$q^*(\tilde{y} \,|\, y) = \frac{\int p(\tilde{y} - \tilde{x}^t\theta, y - x^t\theta)d\theta}{\int p(y - x^t\theta)d\theta}$$

is minimax under the Kullback-Leibler loss.

Proof: First show that q^* has constant risk.

$$\begin{aligned} R(\theta, q^*) &= \mathbb{E}_{\tilde{Y},Y|\theta} \log \frac{p(\tilde{Y} - \tilde{x}^t\theta \,|\, Y - x^t\theta)}{q^*(\tilde{Y} - \tilde{x}^t\theta \,|\, Y - x^t\theta)} \\ &= \mathbb{E}_{\tilde{Z},Z} \log \frac{p(\tilde{Z} \,|\, Z)}{q^*(\tilde{Z} \,|\, Z)} \end{aligned} \tag{3.13}$$

where (3.13) is because

$$\begin{aligned} q^*(\tilde{y} - \tilde{x}^t\theta \,|\, y - x^t\theta) &= \frac{\int p(\tilde{y} - \tilde{x}^t\theta - \tilde{x}^t\alpha, y - x^t\theta - x^t\alpha)d\alpha}{\int p(y - x^t\theta - x^t\alpha)d\alpha} \\ &= \frac{\int p(\tilde{y} - \tilde{x}^t\theta', y - x^t\theta')d\theta'}{\int p(y - x^t\theta')d\theta'}, \quad \theta' = \alpha + \theta \\ &= q^*(\tilde{y} \,|\, y). \end{aligned}$$

That is, q^* is invariant to shift of y by $x^t\theta$ if $\tilde{y}$ is correspondingly shifted by $\tilde{x}^t\theta$.

Next we show that q^* is extended Bayes. Take normal priors $w_k(\theta)$ as in the proof for Theorem 1. Let $w(\theta) = 1$ and $u = (y_{i_1}, \ldots, y_{i_d})$, then by Lemma 1,

$$R_{w_k}(q^*) - R_{w_k}(p_{w_k}) \leq \mathbb{E}_\theta^{w_k} \mathbb{E}_{U|\theta} \mathbb{E}_{\theta'|U}^{w} \log \frac{w_k(\theta)}{w_k(\theta')}. \tag{3.14}$$

Let x denote the $d \times d$ matrix $(x_{i_1}, \ldots, x_{i_d})$ which is non-singular by our assumption. Change variables with $z' = u - x^t\theta'$ and $z = u - x^t\theta$. We find the posterior distribution of Z' given u is independent of u and has the same distribution as $Z = (Z_{i_1}, \ldots, Z_{i_d})$. So the right side of inequality (3.14) is equal to

$$\begin{aligned} \mathbb{E}_\theta^{w_k} \mathbb{E}_Z \mathbb{E}_{Z'} \log \frac{w_k(\theta)}{w_k(\theta')} &= \mathbb{E}_{Z,Z',\theta} \log \frac{w_k(\theta)}{w_k(\theta + (x^t)^{-1}(Z - Z'))} \\ &= \mathbb{E}_{Z,Z',\theta} \frac{\|\theta + (x^t)^{-1}(Z - Z')\|^2 - \|\theta\|^2}{2k} \\ &= \frac{\text{Trace}[(x^{-1})(x^{-1})^t \; \mathbb{E}ZZ^t]}{k}, \end{aligned}$$

which goes to zero when k goes to infinity provided that x is non-singular and Z has finite second moment which are implied in our assumption. Thus q^* is extended Bayes with constant risk, hence minimax. □

In ordinary regression models, we often assume that the errors $\tilde{Z}_i$'s and Z_i's are distributed as independent Normal$(0, \sigma^2)$. The minimax predictive density q^* for future n observations $\tilde{Y} = (\tilde{Y}_1, \ldots, \tilde{Y}_n)$ based on the past observations $Y = (Y_1, \ldots, Y_m)$ is

$$q^*(\tilde{y} \,|\, y) = \frac{\int \phi_{\sigma^2}(\tilde{y} - \tilde{x}^t\theta)\phi_{\sigma^2}(y - x^t\theta)d\theta}{\int \phi_{\sigma^2}(y - x^t\theta)d\theta}. \tag{3.15}$$

We note that

$$\int \phi_{\sigma^2}(y - x^t\theta)d\theta = \Big(\frac{1}{\sqrt{2\pi\sigma^2}}\Big)^{m-d} \frac{1}{|S_m|^{1/2}} \exp(-\frac{1}{2\sigma^2}\mathrm{RSS}_m)$$

where $S_m = \sum_{i=1}^m x_i x_i^t$ is the information matrix and $\mathrm{RSS}_m = ||y - x^t\hat{\theta}_m||^2$ is the residual sum of squares (RSS) from the least squares regression, where $\hat{\theta}_m = (x^tx)^{-1}x^ty$ is the least squares estimate of θ based on the m observations y. Similarly simplifying the numerator of (3.15), we have the following expression for the log predictive density and MDL code length.

$$\log\frac{1}{q^*(\tilde{y}\,|\,y)} = \frac{n}{2}\log 2\pi\sigma^2 + \frac{1}{2\sigma^2}(\mathrm{RSS}_{m+n} - \mathrm{RSS}_m) + \frac{1}{2}\log\frac{|S_{m+n}|}{|S_m|}, \tag{3.16}$$

where $S_{m+n} = S_m + \sum_{i=1}^n \tilde{x}_i\tilde{x}_i^t$ and $\mathrm{RSS}_{m+n} = ||y - x^t\hat{\theta}_{m+n}||^2 + ||\tilde{y} - \tilde{x}^t\hat{\theta}_{m+n}||^2$, respectively, are the information matrix and the residual sum of squares using all $N = m + n$ observations.

For regression model selection, we are looking for the optimal subset of $\tilde{x}$ to predict $\tilde{y}$. Here, the "optimal" means the resulting model has the shortest description length. The code length for the minimax coding strategy q^* given in (3.16) can be used as the criterion for model selection. Since the first term $(n/2)\log 2\pi\sigma^2$ is shared by all models, we omit it from the final MDL criterion:

$$\frac{1}{2\sigma^2}(\mathrm{RSS}_{m+n} - \mathrm{RSS}_m) + \frac{1}{2}\log\frac{|S_{m+n}|}{|S_m|}.$$

When σ^2 is unknown, we find that the minimax procedure q^* is the generalized Bayes procedure with a uniform prior on the location and log-scale parameters (Theorem 6).

$$\begin{aligned} q^*(\tilde{y}\,|\,y) &= \frac{\iint \frac{1}{\sigma}\phi_{\sigma^2}(\tilde{y} - \tilde{x}^t\theta)\phi_{\sigma^2}(y - x^t\theta)d\theta d\sigma}{\iint \frac{1}{\sigma}\phi_{\sigma^2}(y - x^t\theta)d\theta d\sigma} \\ &= \frac{\Gamma(\frac{m+n-d}{2})}{\Gamma(\frac{m-d}{2})} \frac{1}{(\pi)^{n/2}} \frac{|S_m|^{1/2}}{|S_{m+n}|^{1/2}} \frac{(\mathrm{RSS}_m)^{(m-d)/2}}{(\mathrm{RSS}_{m+n})^{(m+n-d)/2}}, \end{aligned}$$

which leads to the following MDL criterion

$$\frac{m+n-d}{2}\log \mathrm{RSS}_{m+n} - \frac{m-d}{2}\log \mathrm{RSS}_m + \frac{1}{2}\log\frac{|S_{m+n}|}{|S_m|} - \log\frac{\Gamma(\frac{m+n-d}{2})}{\Gamma(\frac{m-d}{2})}.$$

Theorem 6 *For the regression model with $m \geq d+1$, assume $(\tilde{Y}, Y)$ is modeled by normal with mean $(\tilde{x}^t\theta, x^t\theta)$ and unknown variance σ^2. Then*

$$q^*(\tilde{y}\,|\,y) = \frac{\iint \frac{1}{\sigma}\phi_{\sigma^2}(\tilde{y}-\tilde{x}^t\theta)\phi_{\sigma^2}(y-x^t\theta)d\theta d\sigma}{\iint \frac{1}{\sigma}\phi_{\sigma^2}(y-x^t\theta)d\theta d\sigma}$$

is minimax under the Kullback-Leibler loss.

Proof: Similarly to the proof for Theorem 5, we can show that q^* has constant risk. To show q^* is extended Bayes, we take the priors $w_k(\theta,\sigma) = w_k^{(1)}(\theta)w_k^{(2)}(\sigma)$ where $w_k^{(1)}$ and $w_k^{(2)}$ are defined in (3.9). The limiting (improper) prior is denoted by $w(\theta,\sigma) = 1/\sigma$. Let $u = (y_1,\ldots,y_{d+1})$ and then by Lemma 1,

$$R_{w_k}(q^*) - R_{w_k}(p_{w_k}) \leq \mathbb{E}^{w_k}_{\theta,\sigma}\mathbb{E}_{U|\theta,\sigma}\mathbb{E}^{w}_{\theta',\sigma'|U}\log\frac{w_k(\theta,\sigma)|\sigma|}{w_k(\theta',\sigma')|\sigma'|}.$$

Change variables from (θ',σ') to $z' = (z'_1,\ldots,z'_{d+1})$ with $z'_i = (y_i - x_i^t\theta')/\sigma'$ and from y_i's to z_i's with $z_i = (y_i - x_i^t\theta)/\sigma$, $i = 1,\ldots,d+1$. We find that the posterior distribution of Z' given U is independent of U and has the same distribution as $Z = (Z_1,\ldots,Z_d)$. So,

$$R_{w_k}(q^*) - R_{w_k}(p_{w_k}) \leq \mathbb{E}^{w_k}_{\theta,\sigma}\mathbb{E}_Z\mathbb{E}_{Z'}\left[\log\frac{w_k^{(1)}(\theta)}{w_k^{(1)}(\theta')} + \log\frac{w_k^{(2)}(\sigma)|\sigma|}{w_k^{(2)}(\sigma')|\sigma'|}\right].$$

From the proof for Theorem 4, we know that the risk difference will go to zero if $\mathbb{E}\log(1+\|\theta-\theta'\|)$ and $\mathbb{E}|\log(|\sigma/\sigma'|)|$ are finite.

Solve (θ',σ') in terms of (θ,σ), z_i's and z'_i's.

$$\begin{aligned}\begin{pmatrix}\theta'-\theta\\ \sigma'\end{pmatrix} &= \begin{pmatrix} x_{11} & \ldots & x_{1d} & z'_1\\ \ldots & & & \ldots\\ x_{d+1,1} & \ldots & x_{d+1,d} & z'_{d+1}\end{pmatrix}^{-1}\begin{pmatrix}\sigma z_1\\ \ldots\\ \sigma z_{d+1}\end{pmatrix} = A^{-1}\begin{pmatrix}\sigma z_1\\ \ldots\\ \sigma z_{d+1}\end{pmatrix}\\ &= \frac{1}{det(A)}\begin{pmatrix} a^{11} & \ldots & a^{d+1,1}\\ \ldots & & \ldots\\ a^{1,d+1} & \ldots & a^{d+1,d+1}\end{pmatrix}\begin{pmatrix}\sigma z_1\\ \ldots\\ \sigma z_{d+1}\end{pmatrix}, \qquad (3.17)\end{aligned}$$

where a^{ij} is the cofactor for the (i,j) element in matrix A. Note that $\{a^{i,d+1}\}_{i=1}^{d+1}$ only involve x's and all other a^{ij} are linear combinations of z_i''s.

$$\begin{aligned}\log(1+\|\theta-\theta'\|) &\leq \sum_{i=1}^{d}\log(1+|\theta_i-\theta_i'|)\\ &= \sum_{i=1}^{d}\log(1+\frac{|\sum_{j=1}^{d+1}\sigma a^{ji}z_j|}{|det(A)|})\\ &\leq \sum_{i=1}^{d}\log(1+\frac{|\sum_j a^{ji}|}{|det(A)|})+\log(1+|\sum_j \sigma z_j|), \qquad (3.18)\end{aligned}$$

where each term is integrable since $\sum_j a^{ji}$, $det(A)=\sum_i a^{i,d+1}z_i'$ and $\sum_j z_j$'s are all normal distributed by our assumption.

$$\begin{aligned}\log\frac{|\sigma|}{|\sigma'|} &= \log\frac{|det(A)|}{|\sum a^{j,d+1}z_j|}\\ &= \log(|\sum a^{j,d+1}z_j'|\,)-\log(|\sum a^{j,d+1}z_j|\,),\end{aligned}$$

which is integrable due to the normality of $\sum_j a^{j,d+1}z_j$ and $\sum_j a^{j,d+1}z_j'$.

So we proved that q^* is extended Bayes and therefore it is minimax for regression model with normal errors whose variance is unknown. □

3.5 Appendix

Appendix A

First for completeness we give a standard fact from statistical decision theory (cf. Ferguson[10], pp. 91, Theorem 3)

Lemma 2 *If procedure q is extended Bayes and has constant finite risk, then q is minimax.*

Proof: Suppose not, then there exists a procedure q' and a positive constant c such that $\max_\theta R(\theta,q') < \max_\theta R(\theta,q)-c$. Since $R(\theta,q)$ is constant for all θ, we have $R(\theta, q') - R(\theta, q) \leq -c$ for all θ. Since the Bayes procedure p_{w_k} minimizes the Bayes risk, we have

$$R_{w_k}(q') - R_{w_k}(p_{w_k}) \geq 0. \tag{3.19}$$

The left side of (3.19) is equal to

$$\begin{aligned} & R_{w_k}(q') - R_{w_k}(q) + R_{w_k}(q) - R_{w_k}(p_{w_k}) \\ = & \int w_k(\theta)[R(\theta, q') - R(\theta, q)]d\theta + [R_{w_k}(q) - R_{w_k}(p_{w_k})] \\ \leq & -c + R_{w_k}(q) - R_{w_k}(p_{w_k}), \end{aligned}$$

which is strictly less than zero when k goes to infinity because of q being extended Bayes. Then it contradicts the condition (3.19) and hence q is minimax. □

Appendix B

Here we relax the moment assumption in Theorem 1.

Theorem 1′ Assume for the location family that at least one of the $Z_1, \ldots, Z_m$ has finite expectation of $\log(1 + |Z_i|)$. Then, under Kullback-Leibler loss, the best invariant predictive procedure

$$q^*(\tilde{y} \mid y) = \frac{\int p(y, \tilde{y} \mid \theta)d\theta}{\int p(y \mid \theta)d\theta}$$

is minimax for any dimension d.

Proof: We use the following priors with a polynomial tails:

$$w_k(\theta) \sim \frac{1}{(1 + ||\theta||/k)^{d+1}}.$$

Continuing the calculation from equation (3.2).

$$\begin{aligned} & \mathbb{E}_{Z,Z',\theta}(d+1)\big[\log(1 + \frac{||\theta + Z_1 - Z_1'||}{k}) - \log(1 + \frac{||\theta||}{k})\big] \\ \leq & \ \mathbb{E}_{Z,Z',\theta}(d+1)\log\big(1 + \frac{||Z_1 - Z_1'||}{k}\big) \\ \leq & \ \mathbb{E}_Z 2(d+1)\log\big(1 + \frac{||Z_1||}{k}\big), \end{aligned} \tag{3.20}$$

where we use $\log(1+||a+b||) \leq \log(1+||a||)+\log(1+||b||)$ at the two inequalities.

Since $\log(1+||Z_1||/k)$ is monotone decreasing with k and it is integrable when $k=1$ by our assumption, the right side of (3.20) goes to zero when k goes to infinity, as a result of Monotone Convergence Theorem. □

Appendix C

The following lemma states the sufficient conditions for admissibility from Berger [6] (page 386). This version is summarized from Farrell (1964) and Brown (1971).

Lemma 3 *Consider a decision problem in which Θ is a nondegenerate convex subset of Euclidean space (i.e., Θ has positive Lebesgue measure), and in which the decision rules with continuous risk functions form a complete class. Then an estimator δ_0 (with a continuous risk function) is admissible if there exists a sequence $\{\pi_k\}$ of (generalized) priors such that*

(a) the Bayes risks $R_{\pi_k}(\delta_0)$ and $R_{\pi_k}(\delta_{\pi_k})$ are finite for all k, where δ_{π_k} is the Bayes rule with respect to π_k;

(b) for any nondegenerate convex set $C \in \Theta$, there exists a $K > 0$ and an integer N such that, for $n \geq N$,

$$\int_C dF^{\pi_k}(\theta) \geq K;$$

(c) $\lim_{k\to\infty}[R_{\pi_k}(\delta_0) - R_{\pi_k}(\delta_{\pi_k})] = 0.$

Proof: Suppose δ_0 is not admissible. Then there exists a decision rule δ' such that $R(\theta,\delta') \leq R(\theta,\delta_0)$, with strict inequality for some θ, say θ_0. Since the rules with continuous risk function form a complete class, it can be assumed that δ' has continuous risk function. Since $R(\theta,\delta_0)$ is also continuous, it follows that there exist constants $\epsilon_1,\epsilon_2 > 0$ such that $R(\theta,\delta') < R(\theta,\delta_0) - \epsilon_1$ for $\theta \in C = \{\theta \in \Theta : |\theta - \theta_0| \leq$

$\epsilon_2\}$. Using this, conditions (a) and (b), and the fact that $B_{\pi_k}(\delta_k) \leq B_{\pi_k}(\delta')$, it can be concluded that for $n \geq N$,

$$\begin{aligned} B_{\pi_k}(\delta_0) - B_{\pi_k}(\delta_k) &\geq B_{\pi_k}(\delta_0) - B_{\pi_k}(\delta') \\ &= \int_\Theta \pi_k(\theta)[R(\theta, \delta_0) - R(\theta, \delta')]d\theta \\ &\geq \int_C \pi_k(\theta)[R(\theta, \delta_0) - R(\theta, \delta')]d\theta \\ &\geq \epsilon_1 K. \end{aligned}$$

This contradicts condition (c) in the assumption. Hence δ_0 must be admissible.

Chapter 4

A Proper Bayes Minimax Estimator

4.1 Introduction

Assume we have data $Y_1, \dots, Y_m$ in $\mathbb{R}^d$ from a Gaussian family $N(\theta, \sigma^2 I)$ with density $\prod_{i=1}^m \phi_{\sigma^2}(y_i - \theta)$ where θ is the unknown location parameter and $\phi_{\sigma^2}(\cdot - \theta)$ denotes the density function for $N(\theta, \sigma^2)$. Let $q(\tilde{y} \,|\, Y_1, \dots, Y_m)$ denote the predictive density estimator for future observations $\tilde{Y} = (Y_{m+1}, \dots, Y_N)$ given the previous m observations. Define the loss to be the Kullback divergence between the density functions $\phi(\tilde{y} - \theta)$ and $q(\tilde{y} \,|\, Y_1, \dots, Y_m)$. The corresponding risk is given by

$$R(\theta, q) = \mathbb{E}_{Y_1, \dots, Y_m, \tilde{Y} | \theta} \log \frac{\phi(\tilde{Y} - \theta)}{q(\tilde{Y} \,|\, Y_1, \dots, Y_m)}.$$

In Chapter 3, we give a minimax estimator q^* which is the best invariant estimator and therefore has constant risk. It takes the form

$$q^*(\tilde{y} \,|\, y_1, \dots, y_m) = \frac{\int \phi(\tilde{y} - \theta) \prod_{i=1}^m \phi(y_i - \theta) d\theta}{\int \prod_{i=1}^m \phi(y_i - \theta) d\theta} \tag{4.1}$$

For instance when $N = m + 1$ this reduces to $q^*(\tilde{y} \,|\, y_1, \dots, y_m) = \phi_{\sigma^2(1+\frac{1}{m})}(\tilde{y} - \bar{y}_m)$ where $\bar{y}_m$ denotes the mean of $y_1, \dots, y_m$. Note that q^* is a generalized Bayes procedure with the improper uniform prior on $\mathbb{R}^d$. In this Chapter we will give a proper

Bayes estimator which is also minimax. It is admissible and beats q^* everywhere provided that the dimension is bigger than four.

4.2 Main Result and Proof

Let $p_w(\tilde{y} \mid y_1, \ldots, y_m)$ denote the Bayes estimator with prior w. Consider the following two-stage prior:

$$\begin{aligned} \theta &\sim N(0, 1/a) \\ p_r(a) &= C\frac{(a\sigma^2/m_0)^{r-\frac{d}{2}-1}}{(1+a\sigma^2/m_0)^{r-\frac{d}{2}+1}}, \quad r > d/2. \end{aligned} \tag{4.2}$$

It is essentially Strawderman's prior [22] except that (rather than having the prior depend on the size m of the sample on which we condition instead) we now have a fixed m_0 and allow all conditioning size $m \geq m_0$.

Theorem 7 *The Bayes procedure p_w using the above two-stage prior for the multivariate normal location family $N(\theta, \sigma^2 I)$ is minimax using Kullback loss, with risk that is everywhere strictly smaller than what is achieved by q^*, for every conditioning size $m \geq m_0$ and all predictive horizons $N > m$.*

Proof: We are to show that the risk difference $R(\theta_0, p_w) - R(\theta_0, q^*)$ is less than zero for any θ_0, by the following steps.

1. Recall that the risk difference is equal to

$$\begin{aligned} & \mathbb{E}\left[\log\frac{p(\tilde{Y} \mid \theta)}{p_w(\tilde{Y} \mid Y)} - \log\frac{p(\tilde{Y} \mid \theta)}{q^*(\tilde{Y} \mid Y)}\right] \\ = \; & \mathbb{E}\log\frac{p_w(Y)/q^*(Y)}{p_w(Y,\tilde{Y})/q^*(Y,\tilde{Y})}. \end{aligned} \tag{4.3}$$

   For the normal distribution, the marginal density $p_w(y)$ has the following decomposition (as used in factorization of the likelihood in accordance with the sufficiency of $\bar{Y}_m$):

$$\begin{aligned} p_w(y) &= \int \prod_{i=1}^{m} \phi_{\sigma^2}(y_i - \theta) w(\theta) d\theta \\ &= \frac{1/\sqrt{m}}{(\sqrt{2\pi\sigma^2})^{m-1}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{m} \|y_i - \bar{y}_m\|^2} \int \phi_{\sigma^2/m}(\bar{y}_m - \theta) w(\theta) d\theta. \quad (4.4) \end{aligned}$$

The above equality holds true for any Bayes mixture p_w including q^* which has the improper prior $w = 1$. Therefore the terms outside the integral in (4.4) are shared by both p_w and q^*. Moreover the integral in (4.4) for q^* is equal to 1 since $w = 1$. So the risk difference (4.3) can be simplified to be

$$\mathbb{E} \log \frac{\int \phi_{\frac{\sigma^2}{m}}(\bar{Y}_m - \theta) w(\theta) d\theta}{\int \phi_{\frac{\sigma^2}{N}}(\bar{Y}_N - \theta) w(\theta) d\theta}, \quad (4.5)$$

where $\bar{Y}_m \sim N(\theta_0, \frac{\sigma^2}{m})$ and $\bar{Y}_{m+1} \sim N(\theta_0, \frac{\sigma^2}{m+1})$.

Notice that the risk difference (4.5) involves normal random variables which only differ in variance. If we define

$$D(t) = \mathbb{E}_Z \log \int \phi_{t^2}(tZ + \theta_0 - \theta) w(\theta) d\theta,$$

where $Z \sim \text{Normal}(0, I)$, then the risk difference (4.5) is equal to $D(\frac{t}{\sqrt{m}}) - D(\frac{t}{\sqrt{N}})$. To show that the risk difference is less than or equal to zero, it suffices to show that $D(t)$ is a decreasing function of t when t is less than $t_0 = \sigma/\sqrt{m_0}$.

2. Next we are to show that the derivative of $D(t)$ using our two-stage prior w is negative. Let $g(t)$ denote the integral inside the log. Using $w(\theta) = \int \phi_{\frac{1}{a}}(\theta) p(a) da$, we have

$$g(t) = \int \phi_{t^2}(tZ + \theta_0 - \theta) \phi_{\frac{1}{a}}(\theta) p_r(a) d\theta da. \quad (4.6)$$

Changing the variable θ to $\tilde{\theta} = \theta/t$ and then integrating $\tilde{\theta}$ out, we obtain the following:

$$\begin{aligned} g(t) &= \frac{1}{t^d}\int \phi(Z+\frac{\theta_0}{t}-\tilde{\theta})\ \phi_{\frac{1}{at^2}}(\tilde{\theta})p(a)d\tilde{\theta}da \\ &= \frac{1}{t^d}\int \phi_{(1+\frac{1}{at^2})}(Z+\frac{\theta_0}{t})\ \phi_{\frac{1}{1+at^2}}(\tilde{\theta}-\frac{Z+\frac{\theta}{t})}{1+at^2})\ p(a)d\tilde{\theta}da \\ &= \frac{1}{t^d}\int \phi_{1+\frac{1}{at^2}}(Z+\frac{\theta_0}{t})\ p(a)da. \end{aligned}$$

Use λ to denote $at^2/(1+at^2)$ which is between 0 and 1, and also use $p_\lambda(\lambda)$ to denote the corresponding density for λ induced from the density $p(a)$. Then the derivative of $g(t)$ with respect to t is given by

$$\frac{1}{t^{d+1}}\int \phi_{\frac{1}{\lambda}}(X)p_\lambda(\lambda)\big[(\tilde{\theta}_0\cdot X-d)\lambda-\lambda(1-\lambda)\|X\|^2\big]d\lambda,$$

where $\tilde{\theta}_0=\theta_0/\sigma$ and $X=Z+\tilde{\theta}_0$ is distributed as $N(\tilde{\theta}_0,I)$. With our choice of $p(a)$ as given in (4.2), the induced prior on λ is

$$p_\lambda(\lambda)=\frac{C\lambda^{r-\frac{d}{2}-1}}{[1+\lambda(t_0^2/t^2-1)]^{r-\frac{d}{2}+1}}\frac{t_0^{2r-d-2}}{t^{2r-d+2}}.$$

Take the derivative of $D(t)=\mathbb{E}\log g(t)$ with respect to t to obtain

$$\mathbb{E}\frac{g'(t)}{g(t)}=\frac{1}{t}\mathbb{E}_{X|\tilde{\theta}_0}\mathbb{E}_{\lambda|X}\big[(\tilde{\theta}_0\cdot X-d)\lambda-\lambda(1-\lambda)\|X\|^2\big], \tag{4.7}$$

where the conditional distribution of λ given X is given by

$$\begin{aligned} p(\lambda\,|\,X) &= \frac{\phi_{\frac{1}{\lambda}}(X)p_\lambda(\lambda)}{\int\phi_{\frac{1}{\lambda}}(X)p_\lambda(\lambda)d\lambda} \\ &= \frac{\lambda^{r-1}e^{-\lambda\|X\|^2/2}h_t(\lambda)}{\int_0^1\lambda^{r-1}e^{-\lambda\|X\|^2/2}h_t(\lambda)d\lambda}, \end{aligned} \tag{4.8}$$

with $h_t(\lambda)=[1+\lambda(t_0^2/t^2-1)]^{-(r-\frac{d}{2}+1)}$.

Using the fact that the noncentraled chi-squared distribution is a Poisson mixture of central chi-squares and some results from [1] which are summarized in Lemma 7 in the Appendix, we have that expression (4.7) is equal to

$$\frac{1}{t}\mathbb{E}_K\mathbb{E}_{V|K}[(2K-d-V)\mathbb{E}(\lambda\,|\,V)+V\mathbb{E}(\lambda^2\,|\,V)],$$

where K is a Poisson random variable with mean $||\tilde{\theta}_0||^2/2t^2$ and given K, the random variable V is chi-squared distributed with $d + 2K$ degrees of freedom. The density of λ given V is given by (4.8) with all the $||X||^2$ replaced by V.

3. We are going to show that

$$\mathbb{E}_{V|K=k}[(2k - d - V)\mathbb{E}(\lambda \mid V) + V\mathbb{E}(\lambda^2 \mid V)] \tag{4.9}$$

is negative for any integer k.

It is shown in Lemma 4 that $V\mathbb{E}[\lambda^2 \mid V] \leq 2(r+1)\mathbb{E}[\lambda \mid V]$. Thus expression (4.9) is less than or equal to

$$\mathbb{E}_{V|K=k}V\mathbb{E}(\lambda \mid V)(\frac{*}{V} - 1)$$

where $*$ is used to denote the expression $2k - d + 2r + 2$. Lemma 6 shows that $f(V) = V\mathbb{E}[\lambda \mid V]$ is increasing in V. Therefore considering the expectation separately over the parts where $V \leq *$ and $V > *$, we obtain the bound

$$f(*)\mathbb{E}_{V|K=k}(\frac{*}{V} - 1) = f(2k - d + 2r + 2)\frac{2r + 4 - 2d}{d + 2k - 2}. \tag{4.10}$$

Here we used the fact that $\mathbb{E}(1/V) = 1/(d + 2k - 2)$ when V is distributed as Chi-squared with degrees of freedom $d + 2k$. The term (4.10) is negative if $r \leq d - 2$. Recall that r is also required to be bigger than $d/2$. So we have the desired terms are negative if $d > 4$. □

Lemma 4 *We have that*

$$V\mathbb{E}(\lambda^2 \mid V) \leq 2(r+1)\mathbb{E}(\lambda \mid V)$$

where the density function of λ given V is given by

$$p(\lambda \mid V) = \frac{\lambda^{r-1}e^{-\lambda V/2}h_t(\lambda)}{\int_0^1 \lambda^{r-1}e^{-\lambda V/2}h_t(\lambda)d\lambda}, \tag{4.11}$$

with $h_t(\lambda) = [1 + \lambda(t_0^2/t^2 - 1)]^{-(r - \frac{d}{2} + 1)}$.

Proof: For each V, change the variable from λ to $\tilde{\lambda} = \lambda V$, then

$$V\mathbb{E}[\lambda^2 \mid V] = \frac{1}{V}\mathbb{E}[\tilde{\lambda}^2 \mid V], \tag{4.12}$$

where the conditional density function of $\tilde{\lambda}$ given V is given by

$$\frac{\tilde{\lambda}^{r-1}e^{-\tilde{\lambda}/2}h_t(\tilde{\lambda}/V)}{\int_0^V \tilde{\lambda}^{r-1}e^{-\tilde{\lambda}/2}h_t(\tilde{\lambda}/V)d\tilde{\lambda}}, \quad V \geq \tilde{\lambda} \geq 0.$$

Let

$$G_r(V) = \int_0^V \frac{1}{\Gamma(r)2^r}\tilde{\lambda}^{r-1}e^{-\tilde{\lambda}/2}h_t(\tilde{\lambda}/V)d\tilde{\lambda},$$

which can be regarded as the expectation of the function $h_t(\tilde{\lambda}/V)1_{V\geq\tilde{\lambda}\geq 0}$ with $\tilde{\lambda}$ distributed as Gamma$(r, 2)$.

The Gamma$(r, 2)$ random variables are stochastically increasing in r, which implies that expectations of increasing functions are also increasing in r. Likewise expectation of decreasing functions are decreasing in r. Since $h_t(\tilde{\lambda}/V)1_{V\geq X\geq 0}$ is non-negative, decreasing with respect to $\tilde{\lambda}$ (when $t \leq t_0$), we have

$$G_{r+1}(V) \leq G_r(V). \tag{4.13}$$

Notice that $\mathbb{E}[\tilde{\lambda}^2 \mid V] = 4r(r+1)\frac{G_{r+2}(V)}{G_r(V)}$. Using (4.13), we obtain

$$\mathbb{E}[\tilde{\lambda}^2 \mid V] \leq 4r(r+1)\frac{G_{r+1}(V)}{G_r(V)} = \frac{2(r+1)}{V}\mathbb{E}[\tilde{\lambda} \mid V].$$

Incorporating this bound into expression (4.12), and reexpressing it in terms of λ and V, we finally get

$$V\mathbb{E}(\lambda^2 \mid V) \leq 2(r+1)\mathbb{E}(\lambda \mid V).$$

□

Lemma 5 *Suppose $g(x)$ and $f(x)$ are two positive functions on $\mathbb{R}$ and assume $xf(x)/f'(x)$ is a monotone decreasing function. Then for a random variable $X \in \mathbb{R}$ with density proportional to $g(x)f(x/v)$ where $v > 0$, its mean is increasing in v.*

Proof: We want to prove that

$$\frac{\int_{-\infty}^{\infty} xg(x)f(x/v)dx}{\int_{-\infty}^{\infty} g(x)f(x/v)dx} \tag{4.14}$$

is an increasing function of v. Taking derivative of (4.14) with respect to v, we obtain

$$\frac{1/v^2}{[\int g(x)f(x/v)dx]^2}\big[-\int x^2 g(x)f'(x/v)dx \int g(x)f(x/v)dx$$
$$+\int xg(x)f(x/v)dx \int xg(x)f'(x/v)dx\big].$$

Writing the product of two integrals as a double integral, we have the expression within the bracket is equal to

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\big[xy\frac{f'(y/v)}{f(y/v)} - x^2\frac{f'(x/v)}{f(x/v)}\big]g(x)f(x/v)g(y)f(y/v)dxdy$$
$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\big[y\frac{f'(y/v)}{f(y/v)} - x\frac{f'(x/v)}{f(x/v)}\big]S(x,y)dxdy, \tag{4.15}$$

where we use $S(x,y)$ denote the symmetric expression $g(x)f(x/v)g(y)f(y/v)$ at the last step.

Recall that

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots dxdy = \int_{-\infty}^{\infty}\int_{-\infty}^{y}\cdots dxdy + \int_{-\infty}^{\infty}\int_{-\infty}^{x}\cdots dydx.$$

So we have expression (4.15) is equal to

$$\int_{-\infty}^{\infty}\int_{-\infty}^{y} x\big[y\frac{f'(y/v)}{f(y/v)} - x\frac{f'(x/v)}{f(x/v)}\big]S(x,y)dxdy$$
$$+ \int_{-\infty}^{\infty}\int_{-\infty}^{x} x\big[y\frac{f'(y/v)}{f(y/v)} - x\frac{f'(x/v)}{f(x/v)}\big]S(x,y)dydx. \tag{4.16}$$

Switching the symbols x and y in (4.16) and using the symmetry of $S(x,y)$, we have the last expression equal to

$$\int_{-\infty}^{\infty}\int_{-\infty}^{y}(x-y)\big[y\frac{f'(y/v)}{f(y/v)} - x\frac{f'(x/v)}{f(x/v)}\big]S(x,y)dxdy,$$

which is always positive if $xf'(x)/f(x)$ is decreasing in x. So we just show that the derivative of the mean of X (with respect to v) is non-negative, hence it is a monotone increasing function of v. □

Lemma 6 *Show that $V\mathbb{E}[\lambda \mid V]$ increases in V where the conditional density of λ given V is given by (4.11).*

Proof: Recall the changing variable we did in the proof for Lemma 4, and we have $V\mathbb{E}[\lambda \mid V] = \mathbb{E}[\tilde{\lambda} \mid V]$, which is the conditional mean of the random variable $\tilde{\lambda}$ which has the density proportional to $\tilde{\lambda}^{r-1}e^{-\tilde{\lambda}/2}\ 1_{V \geq \tilde{\lambda} \geq 0} h_t(\tilde{\lambda}/V)$. The function $h_t(\cdot)$ takes the form of $[1 + xa]^{-k}$ where a and k are both positive. It is easy to check that

$$x\frac{h'(x)}{h(x)} = x\frac{-ka}{1+xa}$$

is a decreasing function. So the monotonicity of $V\mathbb{E}[\lambda \mid V]$ follows by Lemma 5. □

4.3 Implications

Implication for Data Compression

In universal data compression [5][23] each choice of proper probability distribution for $Y_1, \ldots, Y_N$ provides a strategy for compression of (arbitrary discretizations of) these variables. The total description length corresponds to the sum of the description length for an initial segment $Y = (Y_1, \ldots, Y_m)$ based on a distribution $q(y)$ and the description length for the rest $\tilde{Y} = (Y_{m+1}, \ldots, Y_N)$ based on a conditional distribution $q(\tilde{y} \mid y)$. If θ were known the optional strategy would use $p(y, \tilde{y} \mid \theta)$. Performance is measured using the redundancy (expected excess codelength), which for the conditional descriptions is given by

$$R(\theta, q) = \mathbb{E}_{Y,\tilde{Y}|\theta}\left[\log\frac{1}{q(\tilde{Y} \mid Y)} - \log\frac{1}{p(\tilde{Y} \mid Y, \theta)}\right].$$

In Chapter 3, it is shown that for location families the redundancy of the total description length (without conditioning) has infinite supremum over θ for every code strategy q (i.e., the minimax total redundancy is infinite). Fortunately, with conditioning on at least one observation ($m \geq 1$) the minimax value is finite and is achieved by a Bayes predictive distribution (the best invariant rule) $q^*(\tilde{y}\,|\,y)$ based on the uniform prior on $\mathbb{R}^d$. This predictive distribution is made proper by the conditioning on the initial observation(s) y. However, it does not correspond to a proper Bayes distribution $q(\tilde{y}\,|\,y) = \int p(y, \tilde{y}\,|\,\theta)w(\theta)d\theta$ for the description of the total sequence. This motivates our search for proper Bayes minimax strategies, as we have determined in the special case of Gaussian location families. Indeed with the prior given above we have a valid Bayes optimal description for the entire sequence which is simultaneously minimax optimal for the conditional description for all conditioning sizes $m \geq m_0$. Serindipidously, compared to the best invariant rule, it provides everywhere smaller (conditional) redundancy $R(\theta, q)$.

Implication for MDL Criteria in Regression

I believe the above theory will extend to the problem of assigning an optimal description length criterion for model selection in linear regression. If data $Y_1, \ldots, Y_N$ given explanatory variables $x_1, \ldots, x_N$ are modeled as $Y_i = x_i^t\theta + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2)$ with σ^2 known and unknown $\theta \in \mathbb{R}^d$, a minimax optimal description length criterion (for selection among choices of the explanatory variables of dimension $d \leq m$) is

$$\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - x_i^t\hat{\theta}_N)^2 + \frac{1}{2}\log|\sum_{i=1}^{N} x_i x_i^t| - c_m,$$

where $c_m = \frac{1}{2\sigma^2}\sum_{i=1}^{m}(y_i - x_i^t\hat{\theta}_m)^2 + \frac{1}{2}\log|\sum_{i=1}^{m} x_i x_i^t|$ and we conditioned the descriptions on the first m observations. Here $\hat{\theta}_m$, $\hat{\theta}_N$ denote the least squares estimates based on m, N observations, respectively. This minimax criterion is the Bayes strategy using a uniform (improper) prior for θ in $\mathbb{R}^d$.

It is under current investigation whether certain proper Bayes rules are also minimax for this regression problem. The problem is similar to that studied above except for the fact that the covariance matrix of the sufficient statistics based on all the data $Y_1, \ldots, Y_N$ is $\sigma^2(\sum_{i=1}^N x_i x_i^t)^{-1}$ (rather than $(\sigma^2/N)I$) which is not simply a scalar multiple of the corresponding covariance $\sigma^2(\sum_{i=1}^m x_i x_i^t)^{-1}$ based on $Y_1, \ldots, Y_m$. So the calculations are somewhat more delicate.

4.4 Appendix

Lemma 7 *Assume $Y \sim N(\theta, \sigma^2 I)$ and $g(||Y||^2)$ is any function of the norm of Y, then*

$$\begin{aligned} \mathbb{E}g(||Y||^2) &= e^{-||\theta||^2/2\sigma^2} \sum_{k=0}^{\infty} \frac{(||\theta||^2/2\sigma^2)^k}{k!} \mathbb{E}[g(\sigma^2 \chi^2_{d+2k})] \\ &= \mathbb{E}_K \; \mathbb{E}[g(\sigma^2 \chi^2_{d+2K}) \mid K] \end{aligned} \tag{4.17}$$

where K is distributed as Poisson with mean $||\theta||^2/2\sigma^2$ and

$$\theta' \mathbb{E}[Y g(||Y||^2)] = \mathbb{E}_K \; \mathbb{E}[2\sigma^2 K g(\sigma^2 \chi^2_{d+2K}) \mid K]. \tag{4.18}$$

Proof: The proof uses the idea from [12] and [1].

First we make an orthogonal transformation mapping Y to another random variable with the same norm and θ to $(||\theta||, 0, \ldots, 0)$. So $||Y||^2 = U + V$ where $U \sim \sigma^2 \chi^2_{d-1}$ and $V = (\sigma Z + ||\theta||)^2$ with $Z \sim N(0,1)$. The density functions for U and V are

$$\begin{aligned} p_U(u) &= \frac{1}{\Gamma(\frac{d-1}{2})(2\sigma^2)^{(d-1)/2}} u^{\frac{d-1}{2}-1} e^{-\frac{u}{2\sigma^2}} \\ p_V(v) &= \frac{1}{\sqrt{2\pi\sigma^2}} \frac{1}{2\sqrt{v}} [e^{-\frac{(\sqrt{v}+||\theta||)^2}{2\sigma^2}} + e^{-\frac{(\sqrt{v}-||\theta||)^2}{2\sigma^2}}]. \end{aligned}$$

The density for $||Y||^2 = U + V$ is equal to the convolution of p_U and p_V,

$$p_{||Y||^2}(r) = \int_0^r p_U(u) \frac{1}{\sqrt{2\pi\sigma^2}} \frac{1}{2\sqrt{r-u}} [e^{-\frac{(\sqrt{r-u}+||\theta||)^2}{2\sigma^2}} + e^{-\frac{(\sqrt{r-u}-||\theta||)^2}{2\sigma^2}}].$$

Noticing that

$$\begin{aligned} & e^{-(\sqrt{r-u}+||\theta||)^2/2\sigma^2} + e^{-(\sqrt{r-u}-||\theta||)^2/2\sigma^2} \\ = \; & e^{-\frac{r-u}{2\sigma^2}} e^{-||\theta||^2/2\sigma^2} 2 \sum_{k=0}^{\infty} \frac{(||\theta||/\sigma^2)^{2k}(r-u)^{k-\frac{1}{2}}}{(2k)!}, \end{aligned}$$

and reorganizing the expression, we have the density for $||Y||^2$ is equal to

$$\frac{e^{-\frac{r}{2\sigma^2}} e^{-||\theta||^2/2\sigma^2}}{\Gamma(\frac{d-1}{2})(2\sigma^2)^{(d/2)}\sqrt{\pi}} \sum_{k=0}^{\infty} \frac{(||\theta||/\sigma^2)^{2k}}{(2k)!} \int_0^r u^{\frac{d-1}{2}-1}(r-u)^{k+1-\frac{1}{2}} du. \tag{4.19}$$

The integration in last expression is equal to $r^{(d+2k)/2-1} B(\frac{d-1}{2}, k+\frac{1}{2})$ where $B(\cdot,\cdot)$ is the Beta function. Plugging in the result back into expression (4.19), finally we get the density for $||Y||^2$ expressed as

$$\sum_{k=0}^{\infty} \frac{r^{\frac{d+2k}{2}-1} e^{-r/2\sigma^2}}{\Gamma(\frac{d+2k}{2})(2\sigma^2)^{\frac{d+2k}{2}}} e^{-||\theta||^2/2\sigma^2} (\frac{||\theta||^2}{2\sigma^2})^k \frac{1}{k!},$$

which is a mixture of a Poisson (with mean$||\theta||^2/2\sigma^2$) random variable $K$and a scaled (with factor σ^2) central chi-squared distribution with degrees of freedom $d + 2K$. Then equality (4.17) is straightforward.

After the transformation, θ' $\mathbb{E}[Y g(||Y||^2)]$ is equal to $||\theta||$ $\mathbb{E}[Y_1 g(||Y||^2)]$ where Y_1 is the first component of the vector Y. Observe that

$$\mathbb{E} g(||Y||^2) = \int \cdots \int g(\sum y_i^2) \frac{1}{(2\pi\sigma^2)^{d/2}} e^{-\frac{1}{2\sigma^2}\sum y_i^2 - \frac{y_1||\theta||}{\sigma^2}} e^{-||\theta||^2/2\sigma^2}.$$

So $||\theta||$ $\mathbb{E}[Y_1 g(||Y||^2)$ is equal to

$$||\theta||\sigma^2 e^{-||\theta||^2/2\sigma^2} \frac{d}{d||\theta||} \int \cdots \int g(\sum y_i^2) \frac{1}{(2\pi\sigma^2)^{d/2}} e^{-\frac{1}{2\sigma^2}\sum y_i^2 + \frac{y_1||\theta||}{\sigma^2}},$$

which is equal to

$$\begin{aligned}
& ||\theta||\sigma^2 e^{-||\theta||^2/2\sigma^2} \frac{d}{d||\theta||} e^{||\theta||^2/2\sigma^2} \mathbb{E}\big[g(\sigma^2 \chi_{d+2K})\big], \\
= \; & ||\theta||\sigma^2 e^{-||\theta||^2/2\sigma^2} \frac{d}{d||\theta||} \sum_{k=0}^{\infty} \big(\frac{||\theta||^2}{2\sigma^2}\big)^k \; \frac{1}{k!} \; \mathbb{E}\big[g(\sigma^2 \chi_{d+2k})\big], \\
= \; & 2\sigma^2 \sum_{k=0}^{\infty} \big(\frac{||\theta||^2}{2\sigma^2}\big)^k \; \frac{1}{k!} \; k \; \mathbb{E}\big[g(\sigma^2 \chi_{d+2k})\big] \\
= \; & \mathbb{E}_K \mathbb{E}\big[2\sigma^2 K g(\sigma^2 \chi_{d+2K}) \mid K\big].
\end{aligned}$$

So equality (4.18) is true. □

Bibliography

[1] A. J. Baranchik. A family of minimax estimators of the mean of a multivariate normal distribution. *Ann. Math. Stat.*, 41:642–645, 1970.

[2] A. R. Barron. *Locally smooth density estimation.* Ph.D. dissertation, Stanford University, 1985.

[3] A. R. Barron and B. S. Clarke. Information-theoretic asymptotics of bayes methods. *IEEE Trans. Inform. Theory*, 36:453–471, May 1990.

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

[5] A. R. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Trans. Inform. Theory*, 44:2743–2760, October 1998.

[6] J. O. Berger. *Statistical decision theory, foundations, concepts, and methods.* Springer-Verlag New York, 1980.

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

[8] J. O. Berger, L. R. Pericchi, and J. A. Varshavsky. Bayes factors and marginal distributions in invariant situations. *Sankhya A*, 60:307–321, 1998.

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

[10] M. L. Eaton. *Group Invariance Applications in Statistics.* IMS Lecture Notes-Monograph Series, 1989.

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

[12] J. Hartigan. *Bayes Theory.* New York: Springer-Verlag, 1983.

[13] W James and C. Stein. Estimation with quadratic loss. In *Proc. Fourth Berkeley Symp. Math. Statist. Prob.*, pages 1:361–379. Univ. of California Press, 1960.

[14] H. Jeffreys. *Theory of Probability.* New York: Oxford Univ. Press, 1961.

[15] J. Kiefer. Invariance, minimax sequential estimation, and continuous time processes. *Ann. Math. Stat.*, 28:537–601, 1957.

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

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

[18] J. Rissanen. Stochastic complexity and modeling. *Ann. Statist.*, 14:1080–1100, 1986.

[19] J. Rissanen. Stochastic complexity. *Journal of the Royal Statistical Society*, 49(3):223–239, 1987.

[20] J. Rissanen. *Stochastic complexity and statistical inquiry.* Singapore: World Scientific, 1989.

[21] J. Rissanen. Fisher information and stochastic complexity. *IEEE Trans. Inform. Theory*, 42:40–47, 1996.

[22] C. Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proc. Third Berkeley Symp. Math. Statist. Prob.*, pages 1:197–206. Univ. of California Press, 1955.

[23] W. E. Strawderman. Proper bayes minimax estimators of the multivariate normal mean. *Ann. Math. Stat.*, 42:385–388, 1971.

[24] O. Wesler. Invariance theory and a modified minimax principle. *Ann. Math. Stat.*, 30:1–20, 1959.

[25] R. A. Wijsman. *Invariant Measures on Groups and Their Use in Statistics.* IMS Lecture Notes-Monograph Series, 1990.