Bayesian Predictive Density Estimation

Edward I. George and Xinyi Xu [*]

Abstract

The richest form of a prediction is a predictive density over the space of all possible outcomes, a density which is obtained naturally by the Bayesian approach. In this chapter, we describe a variety of recent results that use a decision theoretic framework based on expected Kullback-Leibler loss to evaluate the long run performance of Bayesian predictive estimators. In particular, we focus on high dimensional prediction for the multivariate normal distribution and extensions to the normal linear regression model. General conditions for minimaxity and admissibility, as well as a complete class theorem, are described.

Keywords: Admissibility; Bayes Rules; Kullback-Leibler Loss; Minimaxity; Multiple Shrinkage; Prior choice; Shrinkage Estimation.

[*] *Edward I. George is Professor, Statistics Department, The Wharton School, 3730 Walnut Street 400 JMHH, Philadelphia, PA 19104-6340, edgeorge@wharton.upenn.edu. Xinyi Xu is Assistant Professor, Department of Statistics, The Ohio State University, Columbus, OH 43210-1247, xinyi@stat.ohio-state.edu. This work was supported by NSF grants DMS-0605102 and DMS-0907070. .*

1 Introduction

Predictive analysis, which extracts information from historical and current data to predict future trends and behavior patterns, is one of the most fundamental and important areas in statistics. Of the many possible forms a prediction can take, the richest is a predictive density, a probability distribution over all possible outcomes. Such a comprehensive description of future uncertainty opens the door to sharper risk assessment and better decision making. The statistical challenge of course is how to estimate an unknown predictive density from historical or current data. For this purpose, the Bayesian approach of introducing a prior on the unknowns provides a natural and immediate answer. For example, suppose we observe data $X \sim p(x \mid \theta)$ with unknown parameter θ and wish to predict $Y \sim p(y \mid \theta)$. Given a prior π on θ, it follows from purely probabilistic considerations that a natural estimate of $p(y \mid \theta)$ is the predictive density

$$\hat{p}_\pi(y \mid x) = \int p(y \mid \theta)\pi(\theta \mid x)d\theta, \tag{1}$$

where $\pi(\theta \,|\, x)$ is the posterior distribution of θ. The sheer generality of this formulation provides a systematic approach to estimating $p(y|\theta)$ in a wide variety of setups. For instance, in subsequent sections we will illustrate how such predictive density estimates can borrow strength by combining information across dimensions in a multivariate setting and how they can adapt under model uncertainty in a regression setup. Furthermore, modern developments in numerical and simulation methods, such as Markov Chain Monte Carlo, and the rapid growth in computing power have unleashed the potential of these Bayesian predictive methods even in rather complicated settings.

Although a subjective Bayesian would find the predictive formulation above to be compelling, a skeptical frequentist might wonder how one should go about selecting a "good" prior or, for that matter, why should one even restrict attention to a Bayesian predictive density in the first place. At it turns out, these questions can be answered within a statistical decision theory framework, at least for certain formulations. In such a framework, the performance potential of a density estimator $\hat{p}(y \mid x)$ of $p(y \mid \theta)$ is evaluated by a loss $L(p, \hat{p})$ which is typically averaged over x or θ or both (Berger, 1985). An appealing loss function here is the Kullback-Leibler (KL) or entropy loss,

$$L(p, \hat{p}) = \int p(y \mid \theta) \log \frac{p(y \mid \theta)}{\hat{p}(y \mid x)} \, dy, \tag{2}$$

which when averaged with respect to $p(x \mid \theta)$ leads to a measure of average long run performance, the KL risk criterion

$$R_{KL}(p, \hat{p}) = \int p(x \mid \theta) L(p, \hat{p}) dx. \tag{3}$$

Aitchison (1990) noted that the KL loss is coherent here in the sense that for a given $\pi(\theta)$, the Bayes rule under $R_{KL}(p, \hat{p})$ is $\hat{p}_\pi(y \mid x)$, a property not shared for example by the symmetrized KL loss. For further discussion of the many attractive properties of KL loss, including considerations of information theory, proper local scoring and invariance, see Bernardo and Smith (1994) and the references therein. A more general class of loss functions, the divergence losses, have been considered for prediction in Ghosh *et al.* (2008).

A traditional approach to predictive density estimation has been to substitute an estimator $\hat{\theta}$ for θ and then use $\hat{p}(y \mid x) = p(y \mid \hat{\theta})$. Although appealing in its simplicity, this commonly used "plug-in" approach has been shown by many to often lead to inferior predictive density estimators (Aitchison 1975, Levy and Perng 1986, Geisser 1993, Komaki 1996, Barberis 2000, Tanaka and Komaki 2005, Tanaka 2006). In particular, Aitchison (1975) showed that maximum likelihood plug-in density estimators for Gamma models and for normal models are uniformly dominated under $R_{KL}(p, \hat{p})$ by Bayesian predictive estimators based on flat priors ($\pi(\theta) \equiv 1$). Intuitively, the problem with plug-in estimators is that they ignore the uncertainty about θ by treating it as if were known and equal to $\hat{\theta}$. In contrast, the Bayesian approach directly addresses this parameter uncertainty by margining out θ with respect to a prior distribution, thereby incorporating it into the density estimator.

We note in passing that for plug-in estimators, KL predictive risk is closely related to squared error estimation risk since by a Taylor expansion

$$R_{KL}(p(y \mid \theta), p(y \mid \hat{\theta})) \approx \frac{I(\theta)}{2} E(\theta - \hat{\theta})^2, \tag{4}$$

where $I(\theta)$ is the Fisher information. However, for Bayesian predictive estimators, this simple relationship does not hold. In fact, a Bayes rule does not necessarily belong to the class $\{p(y \mid \theta) : \theta \in R^p\}$, i.e., $\hat{p}_\pi(y \mid x)$ does not correspond to a "plug-in" estimator for θ, although under suitable conditions on π, $\hat{p}_\pi(y \mid x) \to p(y \mid \theta)$ as the sample size $n \to \infty$. Interestingly, as will be described in the next section, for Bayesian predictive densities under the multivariate normal model, there is a direct relationship between the KL predictive risk and the squared error estimation risk, a connection that was established using Stein's unbiased estimate of risk in George *et al.* (2006).

The main challenge for the implementation of the Bayesian predictive approach is the choice of an appropriate prior π. Ideally, such a choice would be guided by meaningful subjective information. However, such information is often not available, especially in complicated problems with many unknown parameters. As noted by Liang *et al.* (2008), "Subjective elicitation of priors for model-specific coefficients is often precluded, particularly in high-dimensional model spaces, such as in nonparametric regression using spline and wavelet bases. Thus, it is often necessary to resort to specification of priors using some formal method (Berger and Pericchi 2001; Kass and Wasserman 1996)." Perhaps the simplest such "objective" approach is to attempt to reduce prior influence by using a diffuse prior such as a flat prior. Although such priors may yield reasonable procedures in low dimensional settings, such priors can also lead to inadequate predictive estimators, especially in high dimensional settings (see, e.g., Jeffreys, 1961 and Berger and Bernardo, 1989).

Ultimately, a criterion such as the KL risk function described above provides a statistical decision theory framework in which the performance properties of Bayesian predictive densities can be compared and evaluated. Recent work using this approach has been fruitful for a number a high dimensional problems. In particular, work by Komaki (2001), Liang and Barron (2004), George *et al.* (2006) and Brown *et al.* (2007) has established conditions for minimaxity and admissibility as well as complete class results for Bayesian predictive density estimators in the fundamental multivariate normal setup. For distributions beyond the normal, new KL risk results for Bayesian predictive densities have been developed by Aslan (2006), Hartigan (1999), Komaki (1996, 2004, 2006) and Sweeting *et al.* (2004). In the following sections, we begin by describing the multivariate normal results in more detail, showing how they lead to uniformly improved Bayesian predictive density estimators over those based on uniform priors. We then proceed to describe how these results can be extended to the linear regression setting. After a simulated illustration of the potential of some of these Bayesian predictive estimators, we conclude with a discussion of directions for future research in this area.

2 Prediction for the Multivariate Normal Distribution

We now focus exclusively on predictive density estimation for the multivariate normal distribution, the centerpiece of parametric models. For this setup, we observe $X \mid \mu \sim N_p(\mu, v_x I)$ and wish to predict $Y \mid \mu \sim N_p(\mu, v_y I)$, two independent p-dimensional multivariate normal vectors with common unknown mean μ. Here $v_x > 0$ and $v_y > 0$ are assumed to be known. By a sufficiency and transformation reduction, this problem is equivalent to estimating the predictive density of X_{n+1} based on observing $X_1, \cdots, X_n$ where $X_1, \cdots, X_n \mid \theta$ i.i.d. $\sim N_p(\theta, \Sigma)$ with unknown θ and known Σ.

The Bayesian predictive density $\hat{p}_U$ under the uniform prior $\pi_U(\theta) \equiv 1$, namely

$$\hat{p}_U(y \mid x) = \frac{1}{\{2\pi(v_x + v_y)\}^{\frac{p}{2}}} \exp\left\{-\frac{\|y - x\|^2}{2(v_x + v_y)}\right\}, \tag{5}$$

dominates the plug-in rule $p(y \mid \hat{\theta}_{MLE})$, which substitutes the maximum likelihood estimate $\hat{\theta}_{MLE} = x$ for θ (Aitchison 1975). Moreover, it is best invariant and minimax with constant risk (Murray 1977, Ng 1980, Liang and Barron 2004), and is admissible when the model dimension $p = 1$ or 2 (Liang and Barron, 2004, Brown *et al.*, 2008). However, when $p \geq 3$, it turns out that $\hat{p}_U(y \mid x)$ can be further dominated by other predictive estimators. Indeed, Komaki (2001) showed that $\hat{p}_H$, the Bayesian predictive density under the Harmonic prior $\pi_H(\beta) \propto \|\beta\|^{-(p-2)}$ dominates $\hat{p}_U$ when the number of potential predictors $p \geq 3$. Similarly, Liang and Barron (2004) showed that proper Bayes rules $\hat{p}_a$ under Strawderman priors $\pi_a(\beta)$, which are defined hierarchically as $\beta \mid s \sim N_p(0, s v_0 I), s \sim (1+s)^{a-2}$, also dominate $\hat{p}_U$ when $p \geq 5$.

It is interesting to note that these results closely parallel some key developments concerning minimax estimation of a multivariate normal mean under quadratic loss.

Based on observing $X \mid \theta \sim N_p(\theta, I)$, that problem is to estimate θ under

$$R_Q(\theta, \hat{\theta}) = E\|\hat{\theta} - \theta\|^2. \tag{6}$$

The maximum likelihood estimator $\hat{\theta}_{MLE}$, which is best invariant, minimax and admissible when $p = 1$ or 2, is dominated by the Bayes rules $\hat{\theta}_\pi = \int \theta \, \pi(\theta \,|\, x) d\theta$ under the Harmonic prior (Stein, 1974) and under the Strawderman prior (Strawderman, 1971) in high dimensions. Note that in the predictive density estimation problem, $\hat{p}_U$ plays the same "straw man" role as $\hat{\theta}_{MLE}$ in the point estimation problem. A further connection between $\hat{\theta}_{MLE}$ and $\hat{p}_U$ is revealed by the fact that $\hat{\theta}_{MLE}$ can also be motivated as the Bayes rule under the uniform prior $\pi_U(\theta) \equiv 1$.

George *et al.* (2006) drew out these parallels by establishing a unifying theory that not only subsumes the specialized results of Komaki (2001) and Liang and Barron (2004), but can also be used to construct large new classes of improved minimax Bayesian predictive densities. Their developments began by showing that any Bayes predictive density $\hat{p}_\pi$ can be represented in terms of the uniform prior estimator $\hat{p}_U$ and the corresponding marginal m_π, namely

$$\hat{p}_\pi(y \mid x) = \frac{m_\pi(w; v_w)}{m_\pi(x; v_x)} \, \hat{p}_U(y \mid x), \tag{7}$$

where $W = \frac{v_y X + v_x Y}{v_x + v_y}$ is a weighted average of X and Y. The principal benefit of the representation (7) is that it reduces the KL risk difference between $\hat{p}_\pi$ and $\hat{p}_U$ to a simple functional of the marginal $m_\pi(z; v)$

$$\begin{aligned} R_{KL}(\theta, \hat{p}_U) - R_{KL}(\theta, \hat{p}_\pi) &= E_{\theta, v_w} \log m_\pi(W; v_w) - E_{\theta, v_x} \log m_\pi(X; v_x) \\ &= \int_{v_w}^{v_x} \frac{\partial}{\partial v} E_{\theta, v} \log m_\pi(Z; v) dv. \end{aligned} \tag{8}$$

Using the heat equation, Brown's representation (Brown 1971) and Stein's identity (Stein 1981), this risk difference can be represented by

$$\begin{aligned} R_{KL}(\theta, \hat{p}_U) - R_{KL}(\theta, \hat{p}_\pi) &= \int_{v_w}^{v_x} E_{\theta, v} \left(\frac{\nabla^2 m_\pi(Z; v)}{m_\pi(Z; v)} - \frac{1}{2} \|\nabla \log m_\pi(Z; v)\|^2 \right) dv \quad (9) \\ &= \int_{v_w}^{v_x} E_{\theta, v} \left[2\nabla^2 \sqrt{m_\pi(Z; v)} / \sqrt{m_\pi(Z; v)} \right] dv. \quad (10) \end{aligned}$$

It is easy to see from (9) and (10) that a sufficient condition for a Bayes predictive density $\hat{p}_\pi$ to be minimax is that $m_\pi(z;v)$ or $\sqrt{m_\pi(z;v)}$ is superharmonic, or as a direct collary, that the prior π is superharmonic. These conditions are essentially the same as the minimax condition for the quadratic risk estimation problem. In both problems, that the Bayes rules under the harmonic prior and the Strawderman prior are minimax in high dimensions now follows easily from the fact that their corresponding marginals or square rooted marginals are superharmonic.

Comparing (9) and (10) with Stein's unbiased estimate of risk (Stein 1974, 1981), George *et al.* (2006) reveals a fascinating identity that provides a connection between KL risk reduction to quadratic risk reduction

$$R_{KL}(\theta, \hat{p}_U) - R_{KL}(\theta, \hat{p}_\pi) = \frac{1}{2}\int_{v_w}^{v_x} \frac{1}{v^2}\left[R_Q^v(\theta, \hat{\theta}_U^v) - R_Q^v(\theta, \hat{\theta}_\pi^v)\right] dv. \tag{11}$$

Ultimately, it is this connection identity that yields similar sufficient conditions for minimaxity and domination in these two problems.

Brown *et al.* (2006) used the connection identity (11) to investigate the admissibility of Bayesian predictive density estimators. As proper Bayes rules are easily shown to be admissible in the KL risk setting, see Berger (1985), the focus was on formal Bayes rules. They showed that under essentially the same tail conditions for π as in Brown and Hwang (1982), there exists a sequence of densities $\{\pi_n\}$ such that $\int_{\|\theta\|\leq 1} \pi_n(\theta)d\theta = \int_{\|\theta\|\leq 1} \pi(\theta)d\theta > 0$ and that $B_Q(\pi_n, \hat{\theta}) - B_Q(\pi_n, \hat{\theta}_{\pi_n}) \to 0$, which using (11) leads to

$$B_{KL}(\pi_n, \hat{p}_\pi) - B_{KL}(\pi_n, \hat{p}_{\pi_n}) = \frac{1}{2}\int_{v_w}^{v_x} \frac{1}{v^2}\left[B_Q^v(\pi_n, \hat{\theta}_\pi) - B_Q^v(\pi_n, \hat{\theta}_{\pi_n})\right] dv \to 0.$$

Then by a variant of Blyth's method, the corresponding Bayes predictive estimator $\hat{p}_\pi$ is admissible. The admissibility of $\hat{p}_U$ when $p = 1$ or 2, and the admissibility of the Bayes rule under the harmonic prior when $p \geq 3$ follow directly from these tail conditions.

Going beyond obtaining prior tail conditions for admissibility, Brown *et al.* (2008) established a compelling justification for restricting attention to Bayesian predictive density estimators for the multivariate normal setup. They showed that for this setup, the class of all generalized Bayes rules forms a complete class under the KL risk criterion. Thus, any predictive estimator, including any plug-in estimator, can at least be matched if not dominated in risk, by some Bayesian predictive density estimator.

These recent results for the multivariate normal model have laid the foundations for the development of new predictive methods for more complicated settings. In particular, the connection identity (11) provides a bridge between the predictive density estimation problem and the classic point estimation problem, providing a tool to borrow strength from some important, beautiful and fundamental results in the latter area.

3 Predictive Density Estimation for Linear Regression

Linear regression models are the mainstay of statistical modeling, in many scenarios at least providing useful approximations to the relationship between explanatory variables and the future outcome of interest (Gelman *et al.* 2003). George and Xu (2008) and Kobayashi and Komaki (2008) both independently studied the problem of predictive density estimation under KL loss in a linear regression setting where they successfully extended a variety of the results discussed in the previous section.

The predictive density estimation problem in this context begins with the canonical normal linear model

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \varepsilon_{n\times 1}, \tag{12}$$

where $\varepsilon \sim N_n(0, \sigma^2 I)$ and X is a full rank, fixed $n \times p$ matrix of p potential predictors where $n \geq p$. Based on observing $X = x$, the goal is to estimate the density of a future vector $\tilde{Y}$ where

$$\tilde{Y}_{m\times 1} = \tilde{X}_{m\times p}\beta_{p\times 1} + \tau_{m\times 1}.$$

Here $\tau \sim N_m(0, \sigma^2 I)$ is independent of ε and $\tilde{X}$ is a fixed $m \times p$ matrix of the same p potential predictors in X with possibly different values. Assume that σ^2 is known, and without loss of generality set $\sigma^2 = 1$ throughout.

Letting $\hat{\beta}_y$ be the traditional maximum likelihood estimate of β based on the observed data, it is tempting to consider the plug-in predictive estimate $\hat{p}_{plug-in}(\tilde{y} \mid \hat{\beta}_y)$, which simply substitutes $\hat{\beta}_y$ for β in $p(\tilde{y} \mid \beta)$. However, as shown by George and Xu (2008), it can be dominated by the Bayesian predictive density $\hat{p}_U(\tilde{y} \mid y)$ under the uniform prior $\pi(\beta) \equiv 1$, namely,

$$\hat{p}_U^L(\tilde{y} \mid y) = \frac{1}{(2\pi\sigma^2)^{\frac{p}{2}}|\Psi|} \exp\left\{ \frac{(\tilde{y} - \tilde{X}\hat{\beta}_y)'\Psi^{-1}(\tilde{y} - \tilde{X}\hat{\beta}_y)}{2\sigma^2} \right\}, \tag{13}$$

where $\Psi = I + \tilde{X}(X'X)^{-1}\tilde{X}'$. Moreover, $\hat{p}_U^L$ has constant risk and is minimax under the KL loss (Liang and Barron 2004). Thus, like $\hat{p}_U$ in (5), it plays the role of straw man in this linear regression setup and is a good default predictive estimator. But not surprisingly, it can be improved upon by other Bayesian predictive densities when $p \geq 3$.

Analogous to the development in the multivariate normal case, the key marginal representation for Bayesian predictive estimator $\hat{p}_\pi^L$ in linear regression can be expressed as

$$\hat{p}_\pi^L(\tilde{y} \mid y) = \frac{m_\pi(\hat{\beta}_{y,\tilde{y}}, (W'W)^{-1})}{m_\pi(\hat{\beta}_y, (X'X)^{-1})} \hat{p}_U^L(\tilde{y} \mid y), \tag{14}$$

where $W = (X', \tilde{X}')'$ and

$$\begin{aligned} \hat{\beta}_y = (X'X)^{-1}X'y \quad &\sim \quad N_p(\beta, (X'X)^{-1}) \\ \hat{\beta}_{y,\tilde{y}} = (W'W)^{-1}W'(x', y')' \quad &\sim \quad N_p(\beta, (W'W)^{-1}). \end{aligned}$$

The representation (14) facilitates the the KL risk comparison of $\hat{p}_U^L$ and $\hat{p}_\pi^L$, where the difference takes the form

$$\begin{aligned} & R_{KL}(\beta, \hat{p}_U^L) - R_{KL}(\beta, \hat{p}_\pi^L) \\ = \;& E_{\beta,(W'W)^{-1}} \log m_\pi(\hat{\beta}_{y,\tilde{y}}; (W'W)^{-1}) - E_{\beta,(X'X)^{-1}} \log m_\pi(\hat{\beta}_y; (X'X)^{-1}). \end{aligned}$$

Since $(W'W)^{-1}$ and $(X'X)^{-1}$ are both symmetric and positive definite, there exists an invertible $p \times p$ matrix P such that

$$(X'X)^{-1} = PP' \qquad \text{and} \qquad (W'W)^{-1} = P\Sigma_D P', \tag{15}$$

where $\Sigma_D = diag(d_1, \ldots, d_p)$. Moreover, $d_i \in (0, 1]$ for all $1 \leq i \leq p$ with at least one $d_i < 1$, because $(W'W)^{-1} = (X'X + \tilde{X}'\tilde{X})^{-1}$ and $\tilde{X}'\tilde{X}$ is nonnegative definite.

Therefore, the KL risk difference between $\hat{p}_U$ and $\hat{p}_\pi$ can then be represented by

$$R_{KL}(\beta, \hat{p}_U^L) - R_{KL}(\beta, \hat{p}_\pi^L) = \sum_{i=1}^{p} (1 - d_i) \int_{d_i}^{1} \frac{\partial}{\partial v_i} E_{\beta,V} \log m_{\pi_P}(Z, V) dv_i, \qquad (16)$$

where $\pi_P(\beta) = \pi(P\beta)$ and $V = diag(v_1, \cdots, v_p)$. Paralleling the development of (9) and (10), unbiased estimates of the components in (16) can be obtained. By combining the above results, George and Xu (2008) established that a sufficient condition for $\hat{p}_\pi^L$ to be minimax is $trace\left\{H(m_\pi(z; PV_wP'))[(X'X)^{-1} - (W'W)^{-1}]\right\} \leq 0$ or $trace\left\{H(\sqrt{m_\pi(z; PV_wP')})[(X'X)^{-1} - (W'W)^{-1}]\right\} \leq 0$ for all $0 \leq w \leq 1$, where $H(f(z_1, \cdots, z_p))$ is the Hessian matrix of a function $f(z_1, \cdots, z_p)$. These results provide substantial generalizations of those in George *et al.* (2008), and can be used to construct improved predictive predictive estimators for linear regression models using scaled harmonic priors, shifted inverted gamma priors and generalized t-priors, following the development in Fourdrinier *et al.* (1998).

4 Multiple Shrinkage Predictive Density Estimation

As will be illustrated in the simulation examples of the next section, Bayesian predictive density estimators can achieve dramatic risk reduction, but only in relatively small neighborhoods of prior modes. Thus, a desirable prior will not only satisfy the minimax and domination conditions above, but will also concentrate prior probability in a neighborhood of β. Now although β will almost always be unknown, there will sometimes be good reason to believe that β may be close to a particular subspace. For example, in large regression problems, it will often be suspected that at least some subset of the predictors is irrelevant in the sense that their coefficients, the corresponding components of β, are very small or zero. In this case, this suspicion would translate into the belief that β might be close to a subspace of β values for which a subset of components is identically zero. To exploit this possibility, George and Xu (2008) proposed the following minimax multiple shrinkage predictive estimators that adaptively shrink β towards the subspace most favored by the data.

First consider the construction of a predictive density estimator that shrinks a particular subset of the β components towards 0. Let S be the subset of $\{1, \ldots, p\}$ corresponding to the indices of the irrelevant predictors, and let β_S be the subvector of β corresponding to the columns of X indexed by S. If the components of β_S were in fact small or zero, it would be have been effective to have used a prior, such as the harmonic prior, that was centered around 0 on β_S and was uniform on $\beta_{\bar{S}}$, where $\bar{S}$ denotes the complement of S. Denoting such a prior by π_S and letting π_S^* be the restriction of π_S to β_S, i.e., $\pi_S^*(\beta_S) = \pi_S(\beta)$ is a function of β_S only, the Bayesian predictive density $\hat{p}^L_{\pi_S}(y \mid x)$ can be expressed as

$$\hat{p}_{\pi_S^*}(\tilde{y} \mid y) = \frac{m_{\pi_S^*}(\hat{\beta}_{S,y,\tilde{y}}, (W_S'W_S)^{-1})}{m_{\pi_S^*}(\hat{\beta}_{S,y}, (X_S'X_S)^{-1})}\, \hat{p}_U(\tilde{y} \mid y).$$

This shrinkage predictive density estimator offers substantial risk reduction when the components of β_S are all very small or zero by shrinking the posterior on the corresponding coefficients of β towards 0.

As was mentioned above, there will typically be uncertainty about which subset of the p predictors in X should be included in the model. Rather than arbitrarily selecting S, an attractive alternative is to use a multiple shrinkage predictive estimator which uses the data to emulate the most effective $\hat{p}_{\pi_S}$. Let Ω be the set of all potentially irrelevant subsets S, possibly even the set of all possible subsets. For each $S \in \Omega$, let π_S be a shrinkage prior constructed as above, and assign it probability $w_S \in [0, 1]$ such that $\sum_{S \in \Omega} w_S = 1$. Then the mixture prior

$$\pi^*(\beta) = \sum_{S \in \Omega} w_S\, \pi_S(\beta)$$

will yield a multiple shrinkage predictive estimator

$$\hat{p}^*(\tilde{y} \mid y) = \sum_{S \in \Omega} \hat{p}(S \mid y) \hat{p}_{\pi_S}(\tilde{y} \mid y), \tag{17}$$

where $\hat{p}(S \mid y)$ is the model posterior probability of the form

$$\hat{p}(S \mid y) = \frac{w_S\, m_{\pi_S^*}(\hat{\beta}_{S,y}, (X_S'X_S)^{-1})}{\sum_{S \in \Omega} w_S\, m_{\pi_S^*}(\hat{\beta}_{S,y}, (X_S'X_S)^{-1})}.$$

The expression (17) shows that $\hat{p}^*(\tilde{y} \mid y)$ is an adaptive convex combination of the individual shrinkage predictive estimates $\hat{p}_{\pi_S}$. Note that through $\hat{p}(S \mid y)$, $\hat{p}^*$ doubly shrinks $\hat{p}_U(\tilde{y} \mid y)$ by putting more weight on the $\hat{p}_{\pi_S}$ for which $m_{\pi_S^*}$ is largest and $\hat{p}_{\pi_S}$ shrinks most. Thus $\hat{p}^*$ is adaptive in the sense that it automatically adjusts to the subset index S for which β_S corresponds exactly to the zero or very small components of β. We expect $\hat{p}^*$ to offer meaningful risk reduction whenever any β_S is small for $S \in \Omega$, and so the potential for risk reduction using $\hat{p}^*$ is far greater than the risk reduction obtained by using an arbitrarily chosen $\hat{p}_{\pi_S}$.

It should be pointed out that the allocation of risk reduction by $\hat{p}^*$ is in part determined by the w_S weights in $\hat{p}(S \mid x)$. Because each $\hat{p}(S \mid y)$ is so sensitive, through $m_{\pi_S^*}$, to the value of $\hat{\beta}_{S,y}$, choosing the weights to be uniform should be adequate. However, one may also want to consider some of the more refined suggestions in George (1986b) for choosing such weights.

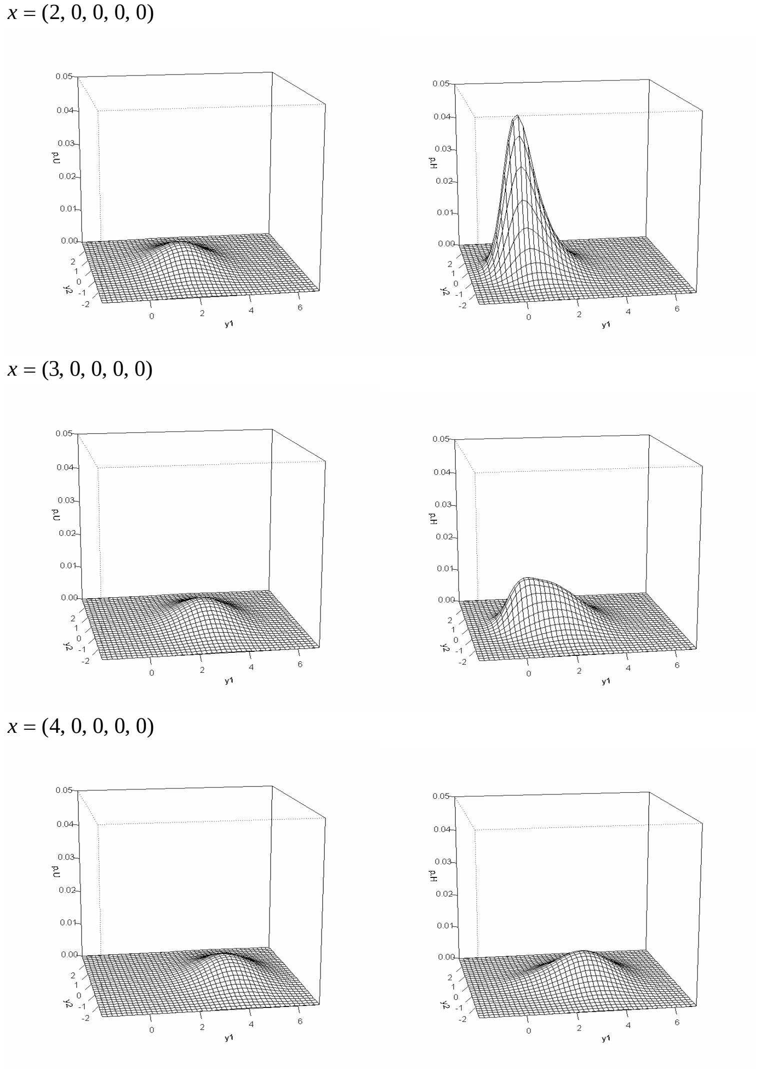
5 Simulation Studies

In this section, we demonstrate the shrinkage properties of some Bayesian predictive densities and their risk improvements over the default procedure under the uniform prior. To make the illustration simple and easy to understand, we use the multivariate normal setup from Section 2 for our simulations. Similar results can be obtained for linear regression models through direct extensions.

Figure 1 illustrates the shrinkage property of the Bayesian predictive density $\hat{p}_H(y|x)$ under the harmonic prior when $v_x = 1, v_y = 0.2$ and $p = 5$. Analogous to Bayes estimators $E_\pi(\theta \mid x)$ of θ that "shrink" $\hat{\theta}_{MLE} = x$, the marginal representation (7) reveals that Bayes predictive densities $\hat{p}_\pi(y \mid x)$ "shrink" $\hat{p}_U(y \mid x)$ by a multiplicative factor of the form $m_\pi(w; v_w)/m_\pi(x; v_x)$. However, the nature of the shrinkage by $\hat{p}_\pi(y \mid x)$ is different than that by $E_\pi(\theta \mid x)$. To insure that $\hat{p}_\pi(y \mid x)$ remains a proper probability distribution, the factor cannot be strictly less than 1. In contrast to simply shifting $\hat{\theta}_{MLE} = x$ towards the mean of π, $\hat{p}_\pi(y \mid x)$ adjusts $\hat{p}_U(y \mid x)$ to concentrate more on the higher probability regions of π.

To study the potential risk improvements provided by Bayesian predictive densities, we illustrate the risk differences of $\hat{p}_U(y \mid x)$ with the Bayes rules under the harmonic prior π_H or the Strawderman's prior π_a with $a = 0.5$. Because $\hat{p}_H$ and $\hat{p}_a$ are unimodal at 0, it intuitively seems that the risk functions $R_{KL}(\theta, \hat{p}_H)$ and $R_{KL}(\theta, \hat{p}_a)$ should take on their minima at $\theta = 0$, and then asymptote up to $R_{KL}(\theta, \hat{p}_U)$ as $\|\theta\| \to \infty$. That

Figure 1: **Shrinkage of $\hat{p}_U(y \mid x)$ to obtain $\hat{p}_H(y \mid x)$ when $v_x = 1, v_y = 0.2$ and $p = 5$. Here $y = (y_1, y_2, 0, 0, 0)$.**



$x = (2, 0, 0, 0, 0)$

$x = (3, 0, 0, 0, 0)$

$x = (4, 0, 0, 0, 0)$

this is exactly what happens for these priors is illustrated in Figure 2 and Figure 3, which display the difference at $\theta = (c, \dots, c)'$, $0 \leq c \leq 4$ when $v_x = 1$ and $v_y = 0.2$ for dimensions $p = 3, 5, 7, 9$. The largest risk reduction in all cases occurs close to $\theta = 0$ and decreases rapidly to 0 as $\|\theta\|$ increases. (Recall that $R_{KL}(\theta, \hat{p}_U)$ is constant as a function of θ). At the same time, risk reduction by $\hat{p}_H$ and $\hat{p}_a$ is larger for larger p at each fixed $\|\theta\|$. Note that $\hat{p}_a$ offers more risk reduction than $\hat{p}_H$, apparently because it more sharply "shrinks $\hat{p}_U(y \mid x)$ towards 0". Note also that when $p = 3$, $[R_{KL}(\theta, \hat{p}_U) - R_{KL}(\theta, \hat{p}_a)]$ is negative for large θ, a manifestation of the non minimaxity of p_a when $a = 0.5$ and $p = 3$.

Figure 2: **The risk difference between $\hat{p}_U$ and $\hat{p}_H$ when $\theta = (c, \cdots, c), v_x = 1, v_y = 0.2$.**

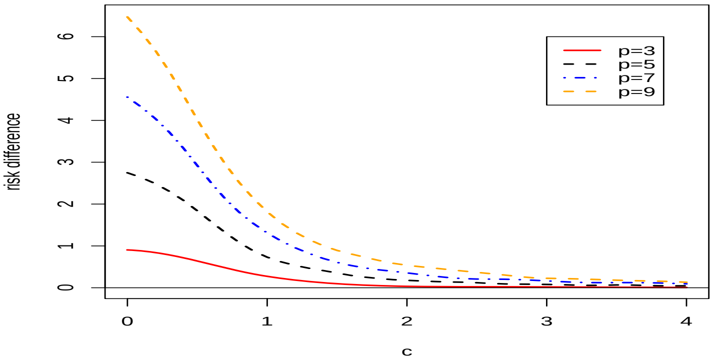
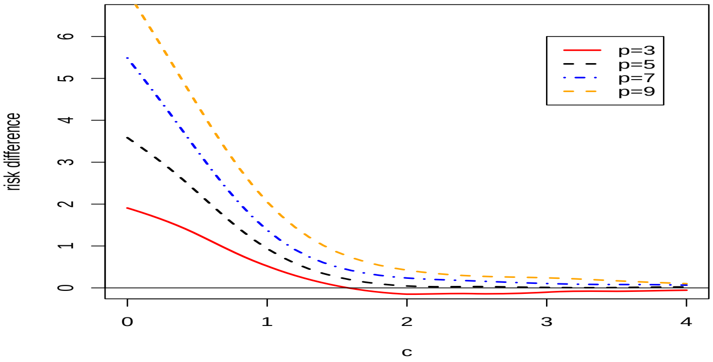


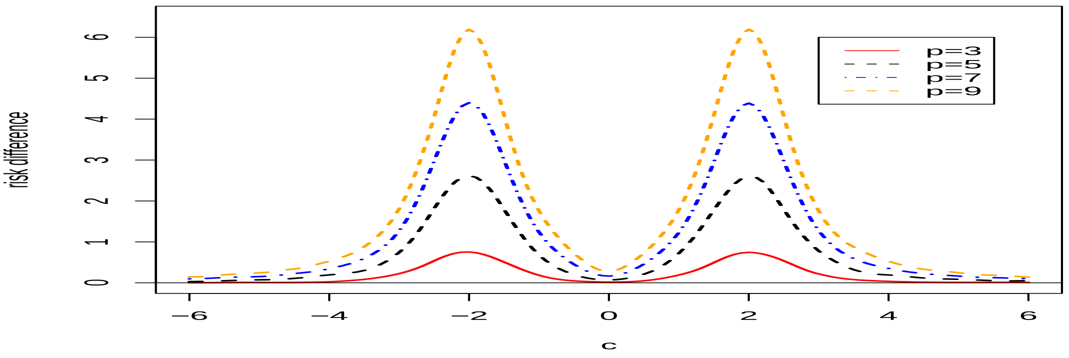
Figure 3: **The risk difference between $\hat{p}_U$ and $\hat{p}_a$ with $a = 0.5$, $v_x = 1, v_y = 0.2$, and $\theta = (c, \cdots, c)$.**



As we have seen in Section 4, the underlying priors and marginals of the Bayesian predictive densities can be readily modified to obtain minimax shrinkage towards subspaces, and linear combinations of superharmonic priors and marginals can be constructed to obtain minimax multiple shrinkage predictive densities $\hat{p}^*$ as in (17), which are analogues of the minimax multiple shrinkage estimators of George (1986abc). As a result of the shrinkage behavior of $\hat{p}^*$, we would expect the risk reduction of $R_{KL}(\theta, \hat{p}^*)$ over $R_{KL}(\theta, \hat{p}_U)$ to be greatest there wherever any β_S is small for $S \in \Omega$.

To see that this is precisely what would happen with $\hat{p}_{H^*}$, a multiple shrinkage version of $\hat{p}_H$ in the multivariate normal setting of Section 2, we consider $\hat{p}_{H^*}$ obtained analogously to (17) but using harmonic priors recentered at $s_1, s_2 \in R^p$, namely $\pi_{H_1}(\beta) \propto \|\beta - s_1\|^{-(p-2)}$ and $\pi_{H_2}(\beta) \propto \|\beta - s_2\|^{-(p-2)}$. Figure 4 illustrates the risk reduction $[R_{KL}(\theta, \hat{p}_U) - R_{KL}(\theta, \hat{p}_{H^*})]$ at various $\theta = (c, \ldots, c)'$ obtained by $\hat{p}_{H^*}$, which adaptively shrinks $\hat{p}_U(y \mid x)$ towards the closer of the two points $s_1 = (2, \ldots, 2)'$ and $s_2 = (-2, \ldots, -2)'$ using equal weights $w_1 = w_2 = 0.5$. As in Figure 2 and 3, we considered the case $v_x = 1, v_y = 0.2$ for $p = 3, 5, 7, 9$. As the plot shows, maximum risk reduction occurs when θ is close to either s_1 or s_2, and goes to 0 when θ moves away from these points. At the same time, for each fixed $\|\theta\|$, risk reduction by $\hat{p}_{H^*}$ is larger for larger p. It is impressive that the size of the risk improvement offered by $\hat{p}_{H^*}$ is nearly the same as each of its single target counterparts. The cost of multiple shrinkage enhancement seems negligible, especially compared to the benefits.

Figure 4: **The risk difference between p_U and multiple shrinkage p_{H^*}, with $\theta = (c, \cdots, c), v_x = 1, v_y = 0.2, a_1 = 2, a_2 = -2$, and $w_1 = w_2 = 0.5$.**



6 Concluding Remarks

Bayesian predictive densities have been widely used in many research areas. Besides predicting future trends and behavior patterns (Taylor and Buizza 2004, Lewis and Whiteman 2006, Weinberg *et al.* 2007), they have also been used in model checking and model diagnostics (Pardoe 2001, Gelman *et al.* 2004, Sinharay *et al.* 2006), missing data analysis (Rubin 1996, Gelman *et al.* 1998, Schafer 1999, Gelman and Raghunathan 2001, Little and Rubin 2002), and data compression and information theory (Barron *et al.* 1998, Clarke and Yuan 1999, Liang and Barron 2004).

Recent developments in Bayesian predictive density estimation for high-dimensional models provide valuable guidance for the construction of predictive estimators for particular setups. However, there are many open directions with much more to be done, especially for more general model setups. In this vein, Kato (2008) considered the predictive density estimation problem for a multivariate normal distribution where both the means and the variances are unknown. The Bayesian predictive estimator under an improper shrinkage prior was shown to dominate the default one under the right invariant prior when $p \geq 3$ and therefore be minimax. In another new direction, Xu and Liang (2009) explored the problem of estimating the predictive density of future observations from a nonparametric regression model. To evaluate the exact asymptotics of the minimax risk, they derived the convergence rate and constant for minimax risk among Bayesian predictive densities under Gaussian priors, and then showed that this minimax risk is asymptotically equivalent to that among all the density estimators. Such results provide not only powerful theoretical tools, but also easily-implementable prior selection strategies for predictive analysis.

REFERENCES

Aitchison, J. (1975). Goodness of Prediction Fit. *Biometrika.* 62, 547-554.

Aitchison, J. (1990). On Coherence in Parametric Density Estimation. *Biometrika.* 77, 905-908.

Aslan, M. (2006). Asymptotically Minimax Bayes Predictive Densities. *Annals of Statistics*, 34, 2921–2938.

Barberis, N. (2000). Investing for the Long Run when Returns are Predictable. *Journal of Finance*, 55, 225-264.

Barron, A.R., Rissanen, J., and Yu, B. (1998). The minimum description length principle in coding and modelling. *IEEE Transaction on Information Theory*, 44, 2743-2760.

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis, Second Edition.* Springer, New York.

Berger, J.O. and Bernardo, J.M. (1989). Estimating a product of means: Bayesian analysis with reference priors. *Journal of the American Statistical Association*, 84, 200-207.

Berger, J. O. and Pericchi, L. (2001). Objective Bayesian methods for model selection: Introduction and comparison. Institute of Mathematical Statistics, Lecture Notes-Monograph Series, Volume 38 (*Model Selection*), 135-193.

Bernardo, J.M. and Smith, A.F.M. (1994). *Bayesian Theory.* New York: Wiley.

Brown, L.D. (1971). Admissible Estimators, Recurrent Diffusions, and Insoluble Boundary Value Problems. *Annals of Mathematical Statistics*, 42, 855-903.

Brown, L.D., George, E.I. and Xu, X. (2008). Admissible Predictive Density Estimation. *Annals of Statistics*, 36, 1156–1170.

Clarke, B. and Yuan, A. (1999). An Informative Criterion for Likelihood Selections. *IEEE Transaction on Information Theory*, 45, 562-571.

Fourdrinier, D., Strawderman, W.E., and Wells, M.T. (1998). On the Construction of Bayes Minimax Estimators. *Annals of Statistics.* 26, 660-671.

Geisser, S. (1993). *Predictive Inference: An Introduction.* CRC Press.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2003). *Bayesian Data Analysis, Second Edition.* Boca Raton, FL: Chapman & Hall / CRC Press.

Gelman, A., King, G. and Liu, C. (1998). Multiple Imputation for Multiple Surveys. *Journal of the American Statistical Association*, 93, 846-874.

Gelman, A. and Raghunathan, T.E. (2001). Using Conditional Distributions for Missing-Data Imputation. *Statistical Science*, 15, 268-269.

George, E.I. (1986a). Minimax Multiple Shrinkage Estimation. *Annals of Statistics*, 14, 188-205.

George, E.I. (1986b). Combining Minimax Shrinkage Estimators. *Journal of the American Statistical Association.* 81, 437-445.

George, E.I. (1986c). A Formal Bayes Multiple Shrinkage Estimator. *Communications in Statistics: Part A - Theory and Methods (Special issue "Stein-type Multivariate Estimation")*, 15, 7, 2099-2114.

George, E.I., Liang, F. and Xu, X. (2006). Improved Minimax Prediction under Kullback-Leibler Loss. *The Annals of Statistics*, 34, 78-91.

George, E.I. and Xu, X. (2008). Predictive Density Estimation for Multiple Regression. *Econometric Theory,* 24, 528-544.

Ghosh, M., Mergel, V., and Datta, G.S. (2008). Estimation, prediction and the Stein phenomenon under divergence loss. *Journal of Multivariate Analysis*, 99, 1941-1961.

Harris, I.R. (1989). Predictive Fit for Natural Exponential Families. *Biometrika* 74, 675-684.

Hartigan, J.A. (1998). The Maximum Likelihood Prior. *Annals of Statistics*, 26, 6, 2083-2103.

Jeffreys, H. (1961). *Theory of Probability, Third Edition.* Oxford University Press.

Kass, R. E. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, 90, 928-934.

Kato, K. (2009). Improved Prediction for a Multivariate Normal Distribution with Unknown Mean and Variance. *Annals of the Institute of Statistical Mathematics*, 61, 531-542.

Kobayashi and Komaki (2008). Bayesian shrinkage prediction for the regression problem. *Journal of Multivariate Analysis*, 99, 1888-1905.

Komaki, F. (1996). On Asymptotic Properties of Predictive Distributions. *Biometrika* 83, 299-313.

Komaki, F. (2001). A Shrinkage Predictive Distribution for Multivariate Normal Observations, *Biometrika.* 88, 859-864.

Komaki, F. (2004). Simultaneous prediction of independent Poisson observables. *Annals of Statistics*, (to appear).

Lehmann, E.L., and Casella, G. (1998). *Theory of Point Estimation, Second Edition* Springer, New York.

Levy, M.S. and Perng, S.K. (1986). An optimal Prediction Function for the Normal Linear Model. *Journal of the American Statistical Association*, 81, 196-198.

Lewis, K.F. and Whiteman, C.H. (2006). Empirical Bayesian Density Forecasting in Iowa and Shrinkage for the Monte Carlo Era. *Working Paper.*

Liang, F. (2002). *Exact Minimax Procedures for Predictive Density Estimation and Data Compression.* Ph.D. dissertation, Department of Statistics, Yale University.

Liang, F. and Barron, A. (2004). Exact Minimax Strategies for Predictive Density Estimation, Data Compression and Model Selection. *IEEE Information Theory Transactions.* 50, 2708-2726.

Liang, F., Paulo, R., Molina, G., Clyde, M., and Berger, J. (2008). Mixtures of g-priors for Bayesian Variable Selection. *Journal of the American Statistical Association*, 103, 410-423.

Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data, Second Edition.* New York: Wiley.

Murray, G.D. (1977), A Note on the Estimation of Probability Density Functions. *Biometrika.* 64, 150-152.

Ng, V.M. (1980). On the Estimation of Parametric Density Functions. *Biometrika.* 67, 505-506.

Pardoe, I. (2001). A Bayesian Sampling Approach to Regression Model Checking. *Journal of Computational and Graphical Statistics*, 10, 617-627.

Rubin, D.B. (1996). Multiple Imputation after 18+ years, *Journal of the American Statistical Association*, 91, 473-489.

Schafer, J.L. (1999). Multiple Imputation: A primer. *Statistical Methods in Medical Research*, 8, 3-15.

Sinharay, S., Johnson, M.S. and Stern, H.S. (2006). Posterior Predictive Assessment of Item Response Theory Models. *Applied Psychological Measurement*, 30, 298-321.

Steele, J.M. (2001). *Stochastic Calculus and Financial Applications.* Springer, New York.

Stein, C. (1974). Estimation of the Mean of a Multivariate Normal Distribution. In *Proceedings of the Prague Symposium on Asymptotic Statistics*, Ed. J. Hájek, pp. 345-81. Prague: Universita Karlova.

Stein, C. (1981). Estimation of a Multivariate Normal Mean. *Ann. Statist.* **9**, 1135-51.

Strawderman, W.E. (1971). Proper Bayes Minimax Estimators of the Multivariate Normal Mean. *Annals of Mathematical Statistics.* 42, 385-388.

Sweeting, T.J., Datta, G.S. and Ghosh, M. (2004). Nonsubjective Priors Via Predictive Relative Entropy Regret. *Annals of Statistics*, (to appear).

Tanaka, F. (2006) Generalized Bayesian predictive density operators. *The 14th Quantum Information Technology Symposium*, 107-110.

Tanaka, F. and Komaki, F. (2005). Bayesian predictive density operators for exchangeable quantum-statistical models. *Physical Review A*, American Institute of Physics, 71, 052323.

Taylor, J.W. and Buizza, R. (2004). Comparing Temperature Density Forecasts from GARCH and Atmospheric Models. *Journal of Forecasting*, 23, 337C355.

Weinberg, J., Brown, L.D., and Stroud, J.R. (2007). Bayesian Forecasting of an Inhomogeneous Poisson Process with Applications to Call Center Data. *Journal of the American Statistical Association.*, 102, 1185-1198.

Xu, X. and Liang, F. (2009). Asymptotic Minimax Risk of Predictive Density Estimation for Nonparametric Regression. *Appearing in Bernoulli.*