Asymptotic Comparison of Predictive Densities for Dependent Observations

Xuanyao He, Zhengyuan Zhu, and Richard L. Smith *

Abstract

This paper studies Bayesian predictive densities based on different priors and frequentist plug-in type predictive densities when the predicted variables are dependent on the observations. Average Kullback-Leibler divergence to the true predictive density is used to measure the performance of different inference procedures. The notion of second-order KL dominance is introduced, and an explicit condition for a prior to be second-order KL dominant is given using an asymptotic expansion. As an example, we show theoretically that for mixed effects models, the Bayesian predictive density with prior from a particular improper prior family dominates the performance of REML plug-in density, while the Jeffreys prior is not always superior to the REML approach. Simulation studies are included which show good agreement with the asymptotic results for moderate sample size.

Some key words: Mixed effect models; Kullback-Leibler divergence; Jeffreys prior; Predictive density; Prediction fit.

*Xuanyao He is Ph.D. student, Department of Statistics and Operations Research, the University of North Carolina at Chapel Hill, NC 27599 (E-mail: xoyo@unc.edu). Richard L. Smith is Professor, Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599 (E-mail: rls@email.unc.edu). Zhengyuan Zhu is Assistant Professor, Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599 (E-mail: zhuz@email.unc.edu).

1 Introduction

Prediction is of great importance in statistics. The general prediction problem can be described as follows. Let $Y = (Y_1, Y_2, \ldots, Y_n)$ be the observation from the distribution $f(\mathbf{y}; \theta)$, where θ is the parameter, and Z be another random variable with distribution also parameterized by θ. Y and Z may be dependent for time series or spatial data, and we would like to predict Z based on observation Y. In principle, we would like to know the distribution of Z conditional on Y. This is usually characterized by a point predictor and a prediction interval in the frequentist framework, and good prediction means both an accurate point predictor and a narrow prediction interval with correct coverage probability. Alternatively, one can take a decision-theoretic approach and define a loss function between the true conditional density $g(z|y, \theta)$ and the predictive density $\hat{g}(z|y)$. A common measure of discrepancy between two density functions g and $\hat{g}$ is the Kullback-Leibler (KL) divergence:

$$D(g(z|y,\theta), \hat{g}(z|y)) = \int g(z|y,\theta) \log \tfrac{g(z|y,\theta)}{\hat{g}(z|y)} dz.$$

To compare two predictive densities $\tilde{g}_1$ and $\tilde{g}_2$, one can look at the expected difference of KL divergence

$$\int \{D(g, \tilde{g}_1) - D(g, \tilde{g}_2)\} f(y; \theta) dy. \tag{1}$$

A number of authors have considered the comparison of different prediction methods using (1) when Z is independent of Y. In terms of this criterion, Aitchison (1975) claims that in general the Bayesian predictive density based on a vague prior is better than the predictive density $g(Z; \hat{\theta})$, where $\hat{\theta}$ is the maximum likelihood estimate (MLE) (Aitchison refers to it as the estimative density), and proved that Bayesian predictive density is better in the sense of (1) for all θ for $g(y; \theta)$ in the gamma and multivariate normal families. These

are apparently the only two known cases when it is proven that such claim holds exactly, though it is widely suspected that the same is true for a much wider class of distributions. Murray (1977) gives a stronger result in his note, using the information measure together with the idea of invariance to derive the vague predictive density as the optimum in a wide class of possible estimates of $p(z|\theta)$. Hartigan (1998) shows that, frequently, in more than one dimension the maximum likelihood estimate plug-in density is asymptotically inadmissible and may be improved upon by using the predictive density corresponding to a least favorable prior. He also provides solutions (if they exist) to certain differential equations as the answer to admissibility questions for the "near ML" estimates. Komaki (1996) considers optimal adjustments of estimative to predictive estimators for exponential families, and confirms Aitchison's conjecture from the viewpoint of asymptotic theory. An expression of the average Kullback-Leibler divergence from the true distribution to a predictive distribution is obtained under the assumption that Z independent of Y. Ren et al. (2006) investigated the estimation and prediction for exponential distributions with unknown rate θ and compared the performances of MLE with Bayesian estimates under several loss functions. They developed second-order asymptotic expressions for the Bayes estimates under several loss functions, one of which is KL divergence. They also assumed that Z is independent of Y.

In this paper, we consider the comparison of predictive density using (1) when Z and Y are dependent. Instead of considering the MLE-based estimative density, we compare the restricted maximum likelihood (REML) estimator-based estimative density and Bayesian predictive densities with some objective priors. One reason for using the KL divergence is that it has historically been the principal device for developing noninformative priors (Hartigan, 1998). Since (1) is in general intractable, we use a higher order Laplace expansion to approximate (1), and introduce the notion of "second-order KL REML-dominant" to compare predictive distributions using the Laplace approximation. A prior on θ is second-order KL REML-dominant if the corresponding predictive distribution under this prior is

better than the REML based estimative distribution for all θ in the sense that the leading term of the second order Laplace expansion of (1) is greater than zero. We derive explicit conditions for second-order KL REML-dominance, and show for one specific mixed effect model that the Jeffreys prior is not second-order KL REML-dominant, while an alternative family of improper prior is. To our knowledge this is the first of such result for the predictive distribution of Z which is dependent on Y. Simulation studies are conducted which show good match to the asymptotic results for moderately large sample size.

The paper is organized as follows. Section 2 provides notation and the model under which we derive our results. Section 3 derives the Laplace expansion of the expected difference between the KL divergence of the REML estimative density and the Bayesian predictive density, and gives explicit conditions for a prior to be second-order KL REML-dominant. Section 4 applies the results to a specific mixed effect model to show that Jeffreys prior is not second-order KL REML-dominant, while a family of improper priors is. Simulation results are also included there. We conclude in Section 5 with some discussion and future work.

2 Notation and Preliminaries

Let $Y = (Y_1, ..., Y_n)'$ be an arbitrary $n \times 1$ observation vector. We are interested in predicting some unobserved value Z, which are dependent on Y. In this paper we only consider univariate Z for simplicity. We assume that Y and Z have joint Guassian distribution of the form

$$\begin{pmatrix} Y \\ Z \end{pmatrix} \sim N\left[\begin{pmatrix} X\eta \\ x_0\eta \end{pmatrix}, \begin{pmatrix} V(\theta) & w^T(\theta) \\ w(\theta) & v_0(\theta) \end{pmatrix} \right], \tag{2}$$

where X and x_0 are assumed to be known matrices of regressors of dimensions $n \times q$ and $q \times 1$ respectively, and η is an unknown $q \times 1$ vector of regression coefficients. Special cases of (2) include various linear mixed effect models, and spatial linear models where the errors

are assumed to be a realization from a Gaussian random field (GRF). We further assume that the covariance elements $V(\theta), w(\theta)$ and $v(\theta)$ are all known functions of an unknown p -dimensional parameter vector θ. In Bayesian analysis we denote the prior density function by $\pi(\theta)$. To simplify the notation, we write V, w and v without indicating the dependence on θ.

The restricted log likelihood function of Y is given by

$$\ell_n(\theta) = -\tfrac{n-q}{2}(\log 2 + \log \pi) + \tfrac{1}{2}\log|X^T X| - \tfrac{1}{2}\log|V| - \tfrac{1}{2}\log|X^T V^{-1} X| - \tfrac{G^2}{2}, \qquad (3)$$

where G is the generalized residual sum of squares given by

$$G^2 = G^2(\theta) = Y^T\{V^{-1} - V^{-1}X(X^T V^{-1} X)^{-1} X^T V^{-1}\}Y. \qquad (4)$$

See for example, Stein (1999) for more discussion on REML and its advantage over regular maximum likelihood (ML) estimators for covariance parameters.

If θ is known, then the Best Linear Unbiased Predictor (BLUP) of z is given by $\hat{z} = \lambda^T Y$, where

$$\lambda = V^{-1}w + V^{-1}X(X^T V^{-1} X)^{-1}(x_0 - X^T V^{-1} w), \qquad (5)$$

and the corresponding prediction error is given by

$$\sigma_0 = v_0 - w^T V^{-1} w + (x_0^T - w^T V^{-1} X)(X^T V^{-1} X)^{-1}(x_0 - X^T V^{-1} w). \qquad (6)$$

Equation (5) and (6) are also known as the universal kriging formula in geostatistics where Y are observations from a GRF.

If we assume θ is known and η unknown with a uniform prior density, the Bayes rule under least squares loss coincides with the Best Linear Unbiased Predictor (BLUP). In this

simple case, the predictive distributions for the frequentist and Bayesian inference are the same, since both reduce to

$$\psi(z; Y, \theta) = \Phi(\tfrac{z-\lambda^T Y}{\sigma_0}), \tag{7}$$

where Φ is the standard normal distribution function. This is also the conditional distribution of Z given Y in a Bayesian framework.

In this work, we study more a complicated case, where θ is assumed to be unknown. Furthermore, under the Bayesian framework, we assume that θ has a prior density $\pi(\theta)$ which is differentiable over a region that includes the true value of θ, while the prior of η continues to be assumed uniform (independent of θ). The posterior predictive distribution function of z given Y is given by

$$\widetilde{\psi}(z; Y) = \tfrac{\int e^{l_n(\theta)+Q(\theta)}\psi(z;Y,\theta)d\theta}{\int e^{l_n(\theta)+Q(\theta)}d\theta}, \tag{8}$$

with $Q(\theta) = \log \pi(\theta)$.

In contrast to (7), we also consider the estimative distribution under the frequentist framework

$$\widehat{\psi}(z; Y) = \psi(z; Y, \widehat{\theta}), \tag{9}$$

where $\widehat{\theta}$ is the REML estimator to maximize $\ell_n(\theta)$.

In this paper we use superscripts to denote components of θ, e.g. θ^i for the ith component. For scalar functions of θ, such as $\psi(z; Y, \theta)$ (with z and Y fixed for the time being) or $Q(\theta)$, subscripts will indicate derivative with respect to the components of θ, i.e. $Q_i = \frac{\partial Q}{\partial \theta^i}, \psi_{ij} = \frac{\partial^2 \psi}{\partial \theta^i \partial \theta^j}$, etc. Also we define $U_i = \frac{\partial \ell_n(\theta)}{\partial \theta^i}, U_{ij} = \frac{\partial^2 \ell_n(\theta)}{\partial \theta^i \partial \theta^j}, U_{ijk} = \frac{\partial^3 \ell_n(\theta)}{\partial \theta^i \partial \theta^j \partial \theta^k}$.

All the above quantities are functions of a particular θ, and could be evaluated at the REML estimator $\hat{\theta}$, which we denote with a hat, such as $\hat{U}_i, \hat{\psi}_{ij}$, and so on. By definition, we have $\{\hat{U}_i\} = 0$, and $\{-\hat{U}_{ij}\}$ is the observed information matrix. If the latter matrix is invertible, we denote its inverse matrix with superscripts, i.e., if A is the $p \times p$ matrix with

(i,j) entry as $\hat{U}_{ij}$, then if A^{-1} exist, $\hat{U}^{ij}$ is its (i,j)s entry.

We also use the summation convention, where a repeated index appearing as both a subscript and a superscript in the same formula implicitly indicates a summation over that index. For simplicity, it is not explicitly indicated that all the quantities depend on dimension n. The following assumptions are made for the rest of the paper.

Assumption 1: $\ell_n(\theta)$ and all of its derivatives are of $O_p(n)$.

Assumption 2: Q and ψ and their derivatives are of $O_p(1)$.

Assumption 3: $\hat{U}_{ij}$ is invertible.

Assumption 1 and 2 are made for consistency with regular maximum likelihood theory for i.i.d. observations, and are satisfied by many linear mixed effect models. For spatial linear models, these two assumptions imply that we are working with the framework of "increasing domain asymptotics", given by K.V. and R.J. (1984), instead of the alternative "infill asymptotics" by Stein (1999).

With the above notations, and also applications of formulae (8.3.50) - (8.3.55) in Chapter 8 of Bleistein and Handelsman (1986), to the numerator and denominator of (7), we get

$$\tilde{\psi} - \hat{\psi} = \tfrac{1}{2}\hat{U}_{ijk}\hat{\psi}_\ell \hat{U}^{ij}\hat{U}^{k\ell} - \tfrac{1}{2}(\hat{\psi}_{ij} + 2\hat{\psi}_i\hat{Q}_j)\hat{U}^{ij} + O_p(n^{-2}). \tag{10}$$

We can write $U_i = n^{1/2}Z_i, U_{ij} = n\kappa_{ij} + n^{1/2}Z_{ij}, U_{ijk} = n\kappa_{ijk} + n^{1/2}Z_{ijk}$, where $\kappa_{ij}, \kappa_{ijk}$ are non-random and Z_i, Z_{ij}, Z_{ijk} are random variables with mean 0, and we assume that all these quantities are of $O(1)$ or $O_p(1)$ as $n \to \infty$.

Also, let $\kappa_{i,j} = E(Z_iZ_j) = -\kappa_{ij}, \kappa_{ij,k} = E(Z_{ij}Z_k)$. By standard identity, $\kappa_{i,j}$ is the (i,j) entry of the normalized Fisher information matrix, we assume this matrix to be invertible with inverse entries $\kappa^{i,j}$. Explicit formulae exist for calculating these quantities, which can be found in Section 8.2 of Smith and Zhu (2004).

In this notation, by a standard Taylor expansion of ℓ_n we get the approximation

$$\hat{\psi} = \psi + n^{-1/2}\kappa^{i,j}Z_i\psi_j + n^{-1}(\kappa^{i,j}\kappa^{k,\ell}Z_{ik}Z_j\psi_\ell + \tfrac{1}{2}\kappa^{i,r}\kappa^{j,s}\kappa^{k,t}\kappa_{ijk}Z_rZ_s\psi_t + \tfrac{1}{2}\kappa^{i,j}\kappa^{k,\ell}Z_iZ_k\psi_{j\ell}) + O_p(n^{-3/2}), \quad (11)$$

and by (10) we have

$$\tilde{\psi} = \hat{\psi} + \tfrac{1}{2}n^{-1}\{\kappa_{ijk}\kappa^{i,j}\kappa^{k,\ell}\psi_\ell + (\psi_{ij} + 2\psi_iQ_j)\kappa^{i,j}\} + O_p(n^{-3/2}). \quad (12)$$

3 Asymptotic Expression of KL divergences

3.1 Kullback-Leibler divergence

Suppose we have some observations, which can be modeled by (1), and we want to predict Z given Y. In this work we will compare predictive and estimative densities using Kullback-Leibler divergence (KL divergence) as the criterion. Let $\psi(z;Y,\theta)$ be the cumulative distribution function (CDF) of z given Y and θ, then the probability density function (PDF) of z given Y is of the form $\varphi(z;Y,\theta) = \frac{\partial\psi}{\partial z}$, and similarly, for any predictive distribution function $\psi^*(z;Y)$, we let $\varphi^*(z;Y) = \frac{\partial\psi^*}{\partial z}$ be the predictive density function. The KL divergence from $\varphi^*(z;Y)$ to $\varphi(z;Y,\theta)$ (simply written as φ below) is given by

$$D(\varphi(z;Y,\theta),\varphi^*(z;Y)) = \int \varphi(z;Y,\theta)\log\tfrac{\varphi(z;Y,\theta)}{\varphi^*(z;Y)}dz \quad (13)$$

We say the predictive density function $\varphi^{*(1)}$ is KL dominated by $\varphi^{*(2)}$ if for all $\theta \in R^p$,

$$
\begin{aligned}
&E_{Y|\theta}[D(\varphi(z;Y,\theta),\varphi^{*(1)}) - D(\varphi(z;Y,\theta),\varphi^{*(2)})] \\
&= \int [D(\varphi(z;Y,\theta),\varphi^{*(1)}) - D(\varphi(z;Y,\theta),\varphi^{*(2)})]\varphi(Y;\theta)dY \geq 0. \qquad (14)
\end{aligned}
$$

Equation (14) can be used to compare two prediction procedures or two priors under the Bayesian framework. In particular, we can compare the estimative and predictive procedure by taking $\widehat{\varphi}(z;Y)$ as $\varphi^{*(1)}$, and $\widetilde{\varphi}(z;Y)$ as $\varphi^{*(2)}$, respectively. It will be interesting if in the Bayesian framework, there exists some prior such that the Bayesian predictive density function is superior to the REML estimative density function in terms of the KL measure for all θ, and we call such prior "KL REML-dominant prior". It is in general difficult to prove exact KL REML-dominance. In what follows, we derive the Laplace expansion of (14), and use the leading terms to define second-order KL REML-dominance.

3.2 Asymptotic approximation to the KL divergences and their difference

Using Taylor expansion of logarithm, we can approximate the Kullback-Leibler divergence from φ^* to φ by

$$
\begin{aligned}
D(\varphi,\tilde{\varphi}) &= -\int \log(\tfrac{\tilde{\varphi}}{\varphi})\varphi dz = -\int \log(\tfrac{\tilde{\varphi}-\varphi}{\varphi}+1)\varphi dz \\
&= -\int [(\tfrac{\tilde{\varphi}-\varphi}{\varphi}) - \tfrac{1}{2}(\tfrac{\tilde{\varphi}-\varphi}{\varphi})^2 + \tfrac{1}{3}(\tfrac{\tilde{\varphi}-\varphi}{\varphi})^3 + O(n^{-3}))\varphi dz \\
&= \tfrac{1}{2}\int \tfrac{(\tilde{\varphi}-\varphi)^2}{\varphi} dz + O(n^{-2}). \qquad (15)
\end{aligned}
$$

Similarly, we can get $D(\varphi,\hat{\varphi}) = \frac{1}{2}\int \frac{(\hat{\varphi}-\varphi)^2}{\varphi} dz + O(n^{-2})$, and

$$
D(\varphi,\hat{\varphi}) - D(\varphi,\tilde{\varphi}) = \tfrac{1}{2}\int \tfrac{(\hat{\varphi}-\varphi)^2-(\tilde{\varphi}-\varphi)^2}{\varphi} dz + O(n^{-2}). \qquad (16)
$$

The above quantity could be expressed explicitly using (11) and (12). Let

$$\psi^*(z;Y) = \psi(z;Y,\theta) + n^{-1/2}R(z,Y) + n^{-1}S(z,Y) + o_p(n^{-1}),$$

where ψ^* denotes the predictive distribution function of Z given Y, and could be either $\widehat{\psi}$ or $\widetilde{\psi}$. We have

$$\varphi^*(z;Y) = \varphi(z;Y,\theta) + n^{-1/2}R^{'}(z,Y) + n^{-1}S'(z,Y) + o_p(n^{-1}),$$

where φ^* denotes the predictive density function of Z given Y, and could be either $\widehat{\varphi}$ or $\widetilde{\varphi}$. By (11) and (12), $R = \kappa^{i,j}Z_i\psi_j, S_1 = \kappa^{i,j}\kappa^{k,l}Z_{ik}Z_j\psi_l + \frac{1}{2}\kappa^{i,r}\kappa^{j,s}\kappa^{k,t}\kappa_{ijk}Z_rZ_s\psi_t + \frac{1}{2}\kappa^{i,j}\kappa^{k,l}Z_iZ_k\psi_{jl}$ for $\hat{\psi}$, and $S_2 = S_1 + \frac{1}{2}\kappa^{i,j}\kappa^{k,l}\kappa_{ijk}\psi_l + (\frac{1}{2}\psi_{ij} + \psi_iQ_j)\kappa^{i,j}$ for $\tilde{\psi}$. Thus

$$\begin{aligned}
R^{'} &= \tfrac{\partial R}{\partial z} = \kappa^{i,j}Z_i\varphi_j,\\
S_1' &= \tfrac{\partial S_1}{\partial z} = \kappa^{i,j}\kappa^{k,l}Z_{ik}Z_j\varphi_l + \tfrac{1}{2}\kappa^{i,r}\kappa^{j,s}\kappa^{k,t}\kappa_{ijk}Z_rZ_s\varphi_t + \tfrac{1}{2}\kappa^{i,j}\kappa^{k,l}Z_iZ_k\varphi_{jl},\\
S_2' &= \tfrac{\partial S_2}{\partial z} = S_1' + \tfrac{1}{2}\kappa^{i,j}\kappa^{k,l}\kappa_{ijk}\varphi_l + (\tfrac{1}{2}\varphi_{ij} + \varphi_iQ_j)\kappa^{i,j}.
\end{aligned}$$

We have

$$\begin{aligned}
\hat{\varphi}(z;Y) &= \varphi(z;Y,\theta) + n^{-1/2}R^{'}(z,Y) + n^{-1}S_1'(z,Y) + o_p(n^{-1}),\\
\tilde{\varphi}(z;Y) &= \varphi(z;Y,\theta) + n^{-1/2}R^{'}(z,Y) + n^{-1}S_2'(z,Y) + o_p(n^{-1}),
\end{aligned}$$

and

$$\begin{aligned}
D(\varphi,\widehat{\varphi}) - D(\varphi,\widetilde{\varphi}) &= \tfrac{1}{2}\int \tfrac{n^{-2}(S_1'^2 - S_2'^2) + 2n^{-3/2}R'(S_1' - S_2')}{\varphi}dz + O(n^{-2})\\
&= n^{-3/2}\int \tfrac{R'(S_1' - S_2')}{\varphi}dz + O(n^{-2}). \qquad (17)
\end{aligned}$$

From (7), we get

$$\varphi_j = \frac{\partial \varphi}{\partial \theta^j} = [-\frac{\sigma_{0j}}{\sigma_0} + (\frac{z-\lambda^T Y}{\sigma_0})(\frac{\lambda_j^T Y}{\sigma_0} + \frac{\sigma_{0j}(z-\lambda^T Y)}{\sigma_0^2})]\varphi, \tag{18}$$

and

$$\varphi_{jl} = \frac{\partial^2 \varphi}{\partial \theta^j \partial \theta^l} = \frac{\partial \varphi_j}{\partial \theta^l} = \frac{\partial\{[-\frac{\sigma_{0j}}{\sigma_0} + (\frac{z-\lambda^T Y}{\sigma_0})(\frac{\lambda_j^T Y}{\sigma_0} + \frac{\sigma_{0j}(z-\lambda^T Y)}{\sigma_0^2})]\varphi\}}{\partial \theta^l}. \tag{19}$$

Applying the above expressions, and the fact that $\frac{z-\lambda^T Y}{\sigma_0} \sim N(0,1)$, we have

$$\int \frac{\varphi_d \varphi_l}{\varphi} dz = \frac{\lambda_d^T Y \lambda_l^T Y}{\sigma_0^2} + 2\frac{\sigma_{0d}\sigma_{0l}}{\sigma_0^2}, \tag{20}$$

$$\int \frac{\varphi_{ij} \varphi_d}{\varphi} dz = \frac{\lambda_{ij}^T Y \lambda_d^T Y + 2\frac{\sigma_{0d}}{\sigma_0}\lambda_i^T Y \lambda_j^T Y}{\sigma_0^2} + 2\frac{\sigma_{0d}\sigma_{0ij}}{\sigma_0^2} + 2\frac{\sigma_{0i}\sigma_{0j}\sigma_{0d}}{\sigma^3}, \tag{21}$$

$$\begin{aligned}\int \frac{\varphi_{ij} \varphi_{bd}}{\varphi} dz = {} & \frac{\lambda_{ij}^T Y \lambda_{bd}^T Y + 2\sigma_{0bd}\sigma_{0ij}}{\sigma_0^2} + 26\frac{\sigma_{0i}\sigma_{0j}\sigma_{0b}\sigma_{0d}}{\sigma^4} \\ & + 2\frac{(\sigma_{0b}\sigma_{0d}\sigma_{0ij} + \sigma_{0i}\sigma_{0j}\sigma_{0bd} + \sigma_{0ij}\lambda_b^T Y \lambda_d^T Y + \sigma_{0bd}\lambda_i^T Y \lambda_j^T Y)}{\sigma_0^3} \\ & + 2\frac{(\lambda_i^T Y \lambda_j^T Y \lambda_b^T Y \lambda_d^T Y + \sigma_{0i}\sigma_{0j}\lambda_b^T Y \lambda_d^T Y + \sigma_{0b}\sigma_{0d}\lambda_i^T Y \lambda_j^T Y)}{\sigma_0^4} \\ & + 6\frac{(\sigma_{0i}\lambda_j^T Y + \sigma_{0j}\lambda_i^T Y)(\sigma_{0b}\lambda_d^T Y + \sigma_{0d}\lambda_b^T Y)}{\sigma_0^4}.\end{aligned} \tag{22}$$

By substituting the above quantities into the leading term on the right hand side of equation (17), we get

$$\begin{aligned} n^{-3/2}\int \frac{R'(S_1'-S_2')}{\varphi} dz &= -n^{-3/2}\int \frac{\kappa^{a,b} Z_a \varphi_b [\frac{1}{2}\kappa_{ijk}\kappa^{i,j}\kappa^{k,l}\varphi_l + (\frac{1}{2}\varphi_{ij} + \varphi_i Q_j)\kappa^{i,j}]}{\varphi} dz \\ &= -n^{-3/2}[\frac{1}{2}\kappa^{a,b} Z_a \kappa_{ijk}\kappa^{i,j}\kappa^{k,l}\varphi_l(\frac{\lambda_b^T Y \lambda_l^T Y}{\sigma_0^2} + 2\frac{\sigma_{0b}\sigma_{0l}}{\sigma_0^2}) \\ &\quad + \frac{1}{2}\kappa^{a,b} Z_a \kappa^{i,j}(\frac{\lambda_{ij}^T Y \lambda_b^T Y + 2\frac{\sigma_{0b}}{\sigma_0}\lambda_i^T Y \lambda_j^T Y}{\sigma_0^2} + 2\frac{\sigma_{0b}\sigma_{0ij}}{\sigma_0^2} + 2\frac{\sigma_{0i}\sigma_{0j}\sigma_{0b}}{\sigma^3}) \\ &\quad + Q_j \kappa^{a,b} Z_a \kappa^{i,j}(\frac{\lambda_b^T Y \lambda_i^T Y}{\sigma_0^2} + 2\frac{\sigma_{0b}\sigma_{0i}}{\sigma_0^2})]. \end{aligned} \tag{23}$$

3.3 Integration over Y given θ

To compute the leading term in the expression of (14), we need to integrate (23), the leading term of the difference between two KL divergences, over Y given θ. Let $Y_\varepsilon = (X\eta)_\varepsilon + e_\varepsilon$, we have $E\{Z_i Y_\varepsilon\} = n^{-1/2} E\{U_i Y_\epsilon\}$. Note that

$$U_i = \tfrac{1}{2}(\upsilon_{\alpha\beta}\tfrac{\partial\omega^{\alpha\beta}}{\partial\theta^j} - e_\alpha e_\beta \tfrac{\partial\omega^{\alpha\beta}}{\partial\theta^j}), \tag{24}$$

where $\{\omega^{\alpha\beta}\} = W = V^{-1} - V^{-1}X(X^TV^{-1}X)^{-1}X^TV^{-1}$, and also $Y^TWY = e^TWe$, we can get $E\{Z_iY_\varepsilon\} = 0$. Furthermore,

$$\begin{aligned}
E\{Z_eZ_fY_\varepsilon Y_\xi\} &= \kappa_{e,f}(X\eta)_\varepsilon(X\eta)_\xi + \kappa_{e,f}\upsilon_{\varepsilon\xi} + \tfrac{1}{n}\{V\tfrac{\partial W}{\partial\theta^e}V\tfrac{\partial W}{\partial\theta^f}V\}_{\varepsilon\xi} + \tfrac{1}{n}\{V\tfrac{\partial W}{\partial\theta^f}V\tfrac{\partial W}{\partial\theta^e}V\}_{\varepsilon\xi},\\
E\{\lambda_j^\xi Y_\xi \lambda_d^\varepsilon Y_\varepsilon\} &= \lambda_j^\xi\lambda_d^\varepsilon[(X\eta)_\varepsilon(X\eta)_\xi + \upsilon_{\varepsilon\xi}] = \lambda_j^\xi\lambda_d^\varepsilon\upsilon_{\varepsilon\xi},\\
E\{Z_a\lambda_b^\xi Y_\xi\lambda_l^\varepsilon Y_\varepsilon\} &= \lambda_b^\xi\lambda_l^\varepsilon E(Z_a e_\varepsilon e_\xi) = -\tfrac{1}{2}n^{-1/2}\lambda_b^\beta\lambda_l^\varepsilon\tfrac{\partial\omega^{\alpha\beta}}{\partial\theta^a}(\upsilon_{\alpha\varepsilon}\upsilon_{\beta\xi} + \upsilon_{\alpha\xi}\upsilon_{\beta\varepsilon}).
\end{aligned}$$

Integration of (23) over Y can be evaluated using the above equations, which leads to the following theorem.

Theorem 1 *Under Assumption 1-3,*

$$E_{Y|\theta}[D(\varphi,\widehat{\varphi}) - D(\varphi,\tilde{\varphi})] = g_1 - g_2 + O(n^{-2}), \tag{25}$$

where

$$g_1 = n^{-2}\{Q_j\kappa^{a,b}\kappa^{i,j}\lambda_b^\xi\lambda_i^\epsilon\tfrac{\partial\omega^{\alpha\beta}}{\partial\theta^a}(\upsilon_{\alpha\epsilon}\upsilon_{\beta\xi} + \upsilon_{\alpha\xi}\upsilon_{\beta\epsilon})\}, \tag{26}$$

$$\begin{aligned}
g_2 = -\tfrac{1}{2}n^{-2}\{&\kappa^{a,b}\kappa^{i,j}\kappa^{k,l}\kappa_{kij}\lambda_b^\xi\lambda_l^\epsilon\tfrac{\partial\omega^{\alpha\beta}}{\partial\theta^a}(\upsilon_{\alpha\epsilon}\upsilon_{\beta\xi} + \upsilon_{\alpha\xi}\upsilon_{\beta\epsilon})\\
&+ \kappa^{a,b}\kappa^{i,j}[\lambda_b^\xi\lambda_{ij}^\epsilon\tfrac{\partial\omega^{\alpha\beta}}{\partial\theta^a}(\upsilon_{\alpha\epsilon}\upsilon_{\beta\xi} + \upsilon_{\alpha\xi}\upsilon_{\beta\epsilon})\\
&+ 2\tfrac{\sigma_{0b}}{\sigma_0}\lambda_i^\xi\lambda_j^\epsilon\tfrac{\partial\omega^{\alpha\beta}}{\partial\theta^a}(\upsilon_{\alpha\epsilon}\upsilon_{\beta\xi} + \upsilon_{\alpha\xi}\upsilon_{\beta\epsilon})]\}.
\end{aligned} \tag{27}$$

Remarks: We say a prior $\pi(\theta)$ is 'second-order KL REML-dominant', if under such prior $g_1 \geq g_2$ for all θ, i.e., the leading term of $E_{Y|\theta}[D(\varphi, \widehat{\varphi}) - D(\varphi, \tilde{\varphi})]$ is greater than zero.

4 Example: Mixed Effect Model

In this section we compare the plug-in predictive density and the Bayesian predictive density in terms of their KL divergences to the true conditional density function $\varphi(z; Y, \theta)$, for a simple mixed effect model as an illustration. We consider the Jeffreys prior and a family of improper priors, and show theoretically that the Jeffreys prior is not second-order KL REML-dominant, while there exist α^* such that the improper prior family $\pi(\beta, s_1, s_2) \propto (s_1 s_2)^{-\alpha}$ is second-order KL REML-dominant for $\alpha \in [\alpha^*, 1)$. Simulation studies are conducted which have great agreement with the theoretical results based on asymptotic expansion for moderate sample sizes.

4.1 Model and Notation

Consider the simple mixed effect model

$$y_{i,j} = \beta + \mu_i + \epsilon_{ij}, \qquad i = 1, ..., n, j = 1, ..., m, \tag{28}$$

where $\mu_i \sim N(0, s_1), \epsilon_{ij} \sim N(0, s_2)$, and $\pi(\beta) \propto$ constant. Without loss of generality, we study the predictive density of μ_1. Using the vector notation in equation (2), we have $X = \mathbf{1}_{mn}, x_0 = 1, V = Diag\{V_1, V_2, ..., V_n\}$ with $V_i = s_1 \cdot J_m + s_2 \cdot I_m$, $w = (0_{m(n-1)}, s_1 \cdot 1_m)^T$, and $v_0 = s_1$, where $\mathbf{1}_{mn} = (1, ..., 1)^T, I_m$ the identity matrix, $J_m = \mathbf{1}_m \mathbf{1}_m^T, i = 1, ..., n$. By

standard computation, we get

$$\begin{aligned}
|X^TX|^{1/2} &= (mn)^{1/2}, \\
|V|^{-1/2} &= s_2^{-(m-1)n/2}(s_2+ms_1)^{-n/2}, \\
|X^TV^{-1}X|^{-1/2} &= (s_2+ms_1)^{1/2}(mn)^{-1/2}, \\
\lambda &= \tfrac{s_2}{mn(s_2+ms_1)}1_{mn} + \tfrac{s_1}{s_2+ms_1}(0_{m(n-1)}, 1_m)^T, \\
\sigma^2 &= \tfrac{s_2(mns_1+s_2)}{mn(s_2+ms_1)}, \\
W &= \{\omega^{\alpha,\beta}\} = V^{-1} - V^{-1}X(X^TV^{-1}X)^{-1}X^TV^{-1} \\
&= Diag\{V^1, ..., V^n\} - \tfrac{1}{nm(s_2+ms_1)}J_{mn},
\end{aligned}$$

$$\begin{aligned}
V^i &= \tfrac{1}{s_2}I_{m\times m} - \tfrac{s_1}{s_2(s_2+ms_1)}J_m, i = 1, ..., n, \\
K &= \{\kappa^{i,j}\}_{2\times 2} = \begin{pmatrix} \frac{2n(s_2+ms_1)^2}{(n-1)m} + \frac{2s_2^2}{m(m-1)} & -\frac{2s_2^2}{m-1} \\ -\frac{2s_2^2}{m-1} & \frac{2ms_2^2}{m-1} \end{pmatrix}.
\end{aligned}$$

From the matrix form of K, we can derive that the Jeffreys prior of $\theta = (s_1, s_2)$ in this model is

$$\pi_J(\theta) \propto \tfrac{1}{s_2(ms_1+s_2)},$$

with

$$\begin{aligned}
Q_{1J} &= \tfrac{\partial \log \pi}{\partial s_1} = -\tfrac{m}{ms_1+s_2}, \\
Q_{2J} &= -\tfrac{ms_1+2s_2}{s_2(ms_1+s_2)}.
\end{aligned}$$

We add the subscript "J" to denote the Jeffreys prior and relevant functions. A family of

improper priors is also considered, which we refer to as

$$\pi_I(\beta, s_1, s_2) \propto (s_1 s_2)^{-\alpha}, \tag{29}$$

with $\alpha \in (0, 1)$ to guarantee proper posterior distributions (Hobert and Casella, 1996). Correspondingly, we have

$$\begin{aligned} Q_{1I} &= \tfrac{\partial \log \pi_I}{\partial s_1} = -\tfrac{\alpha}{s_1}, \\ Q_{2I} &= -\tfrac{\alpha}{s_2}, \end{aligned}$$

where we add "I" as a subscript to relevant functions under the improper priors.

For general n and m, equation(26) becomes

$$\begin{aligned} g_1 = n^{-2}[Q_1(&\tfrac{8ns_2^2}{1-n} - \tfrac{4(4-m)s_2^3}{(m-1)(s_2+ms_1)} + \tfrac{4s_2^4(m-2)(n-1)}{n(m-1)^2(s_2+ms_1)^2}) \\ &+ Q_2(\tfrac{4m^2s_2^3}{(m-1)^2(s_2+ms_1)} + \tfrac{4ms_2^3(s_2+mns_1)(m-2)}{n(m-1)^2(s_2+ms_1)^2})], \end{aligned} \tag{30}$$

and equation (27) becomes

$$\begin{aligned} g_2 = n^{-2}[&\tfrac{-2m^2s_2^2(20s_1^2s_2+10s_1s_2^2-s_2^3+4ms_1^3-8ms_1^2s_2-6ms_1s_2^2)}{(m-1)^2s_1(s_2+ms_1)^2} + \tfrac{8ms_2^2}{(n-1)(s_2+ms_1)} \\ &- \tfrac{2s_2^3(m-2)m(-11s_2-4ms_1+8ms_2+4m^2s_1)}{(m-1)^2n(s_2+ms_1)^3} \\ &- \tfrac{2(-2m^2s_1^2s_2^3-m^2s_1s_2^4+m^2s_2^5+2m^3s_1^2s_2^3+2m^3s_1s_2^4)}{(m-1)^2s_1(s_2+ms_1)^2(s_2+mns_1)}]. \end{aligned} \tag{31}$$

4.2 Theoretical results for the mixed effect model

In this section, we show in Theorem 2 and 3 that for the mixed effect model (28), there exists α^* such that the family of improper priors π_I with $\alpha \in [\alpha^*, 1)$ are second-order KL REML

dominant, while the Jeffreys prior is not. In the remarks we give conditions under which Bayesian predictive density with Jeffreys prior outperforms the REML plug-in estimator, and more explicit results on α^*. All the results are derived under Assumptions 1 - 3. We first consider the Jeffreys prior. Substituting Q_1, Q_2 into equation (30) and compare that with (31), we have

$$\begin{aligned} n^2(g_1 - g_2) = \tfrac{2ms_2^2}{(m-1)^2(s_2+ms_1)^3 n(s_2+mns_1)}[&4m^3s_1^3(m^2-3m+3)n^2 + 2m^2ns_1^2s_2((3n+4)m^2 \\ &- 12m(n+1) + 14n + 11) - s_2^3((n^2+n-8)m^2 - m(5n-29) - 26) \\ &+ 2ms_1s_2^2((3n^2-26n-7)m + m^2(-2n^2+8n+2) + 26n + 6)]. \end{aligned} \tag{32}$$

Theorem 2 *For the mixed effect model (28), There exists r such that $g_1 \geq g_2$ iff $\frac{s_1}{s_2} \geq r$ as $n \to \infty$, i.e., the Jeffreys prior is not second-order KL REML-dominant for this specific model.*

Proof: For any fixed n and m, the square bracket section of (32) divided by s_2^3 is a function of the ratio s_1/s_2. Let $x = s_1/s_2$, we have

$$g_1 - g_2 = C(m, n, s_1, s_2)(x^3 + ax^2 + bx + c),$$

where $C(m, n, s_1, s_2)$ is positive for all $m > 1, n > 0, s_1 > 0, s_2 > 0$, and

$$\begin{aligned} a_n &= \tfrac{2m^2n[(3n+4)m^2-12m(n+1)+14n+11]}{4m^3n^2(m^2-3m+3)} \\ b_n &= \tfrac{2m[m^2(-2n^2+8n+2)+(3n^2-26n-7)m+26n+6]}{4m^3n^2(m^2-3m+3)} \\ c_n &= \tfrac{-[(n^2+n-8)m^2-m(5n-29)-26]}{4m^3n^2(m^2-3m+3)}. \end{aligned}$$

Let $f(x) = x^3 + a_n x^2 + b_n x + c_n$, it is easy to see that a prior is second-order KL REML-dominant if and only if $f(x) \geq 0$ for all $x > 0$. When $n \to \infty$,

$$a = \lim_{n\to\infty} a_n = \tfrac{3m^2-12m+14}{2m(m^2-3m+3)},$$
$$b = \lim_{n\to\infty} b_n = \tfrac{3-2m}{2m(m^2-3m+3)},$$
$$c = \lim_{n\to\infty} c_n = \tfrac{-1}{4m(m^2-3m+3)}.$$

$f(x)$ is a convex function of x for $x > 0$ as $f''(x) = 6x+2a \geq 0$, for $x > 0$. Since $f(0) = c < 0$ and $f(x)$ goes to infinite as $x \to \infty$, there exist $r > 0$ such that $f(x) < 0$ for $x \in (0, r)$ and $f(x) \geq 0$ for $x \in [r, \infty)$, which proves the claim. $\square$

Remarks:

- It can be shown that for $m > 1$ and $m \neq 3$, $r \in [1/(2m), 1/m]$ as $n \to \infty$. For $m = 3$, $r \approx 0.372$ as $n \to \infty$, a little bigger than but still very close to $1/m$.
- For $m = 2$, the root of $f(x)$ for any given n is $\frac{2n-3}{4n}$, approaching $\frac{1}{2}$ as $n \to \infty$.

Next, we consider the improper prior, which has the property given by the following theorem.

Theorem 3 *There exists $\alpha^* \in (0, 1)$ such that as $n \to \infty$, under the improper prior $\pi_I(\beta, s_1, s_2) \propto (s_1 s_2)^{-\alpha}$ with $\alpha \in [\alpha^*, 1)$, $g_1 \geq g_2$ for all s_1 and s_2, i.e., this improper prior family is second-order KL REML-dominant for $\alpha \in [\alpha^*, 1)$.*

*P*roof: Substituting Q_{1I}, Q_{2I} into equation (30), we get

$$\begin{aligned} g_1 =& n^{-2}\alpha \tfrac{4s_2^2}{ns_1(n-1)(m-1)^2(ms_1+s_2)^2}\{2s_1^2(m-1)m^2n[(m-2)n+1] \\ &+ s_2^2[m^2n(n+1) - m(3n+1) + 2] + ms_1s_2[m^2n(3n+1) + m(1-5n-4n^2) + 6n - 2]\}. \end{aligned}$$

It is enough to show that there exists α^* such that

$$\lim_{n\to\infty} \frac{g_2}{g_1} = \frac{-4m^3x^3+2m^2(4m-10)x^2+2m^2(3m-5)x+m^2}{2\alpha(xm+1)[2m^2(m-1)(m-2)x^2+m^2(3m-4)x+m^2]} \leq 1$$

for $\alpha \in [\alpha^*, 1)$ and $m > 1, m \in \mathcal{Z}$, where $x = s_1/s_2$. This is equivalent to showing

$$f(x) = ax^3 + bx^2 + cx + d \geq 0,$$

where

$$\begin{aligned} a &= 4m(\alpha(m-1)(m-2)+1), \\ b &= 10\alpha m^2 - (20\alpha+8)m + 20, \\ c &= (8\alpha-6)m + 2\alpha + 2, \\ d &= 2\alpha - 1. \end{aligned}$$

It is easy to show that $a > 0$ for all α, $b > 0$ for $\alpha \in [3/5 - 1/\sqrt{5}, 1)$, $c > 0$ for $\alpha > 3/4$, and $d > 0$ for $\alpha > 1/2$. Let $\alpha^* = 3/4$, we have that $f(x) \geq 0$ for any $\alpha \in [\alpha^*, 1)$. $\square$

Remarks:

- It can be shown numerically that the minimum α^* for π_I to be second order KL REML dominant for all $m > 1, m \in \mathcal{Z}$ is approximately 0.5532.
- For $m = 2$, it can be shown that π_I is second order KL REML dominant for $\alpha \in [1/2, 1)$.

In Figure 1, we plot $(g_1 - g_2)$ as a function of s_1 with $s_2 = 1$ for $n = 10, 20, 50, 100, 1000$ and $m = 2, 5, 10$ respectively. The plots on the left is for the Jeffreys prior, and the plots on the right are for the improper prior with $\alpha = 0.75$. These plots numerically show that the asymptotic results in Theorem 1 and 2 are also valid for finite sample size with n as small as 10: The Jeffreys prior are not second order KL REML dominant while the improper

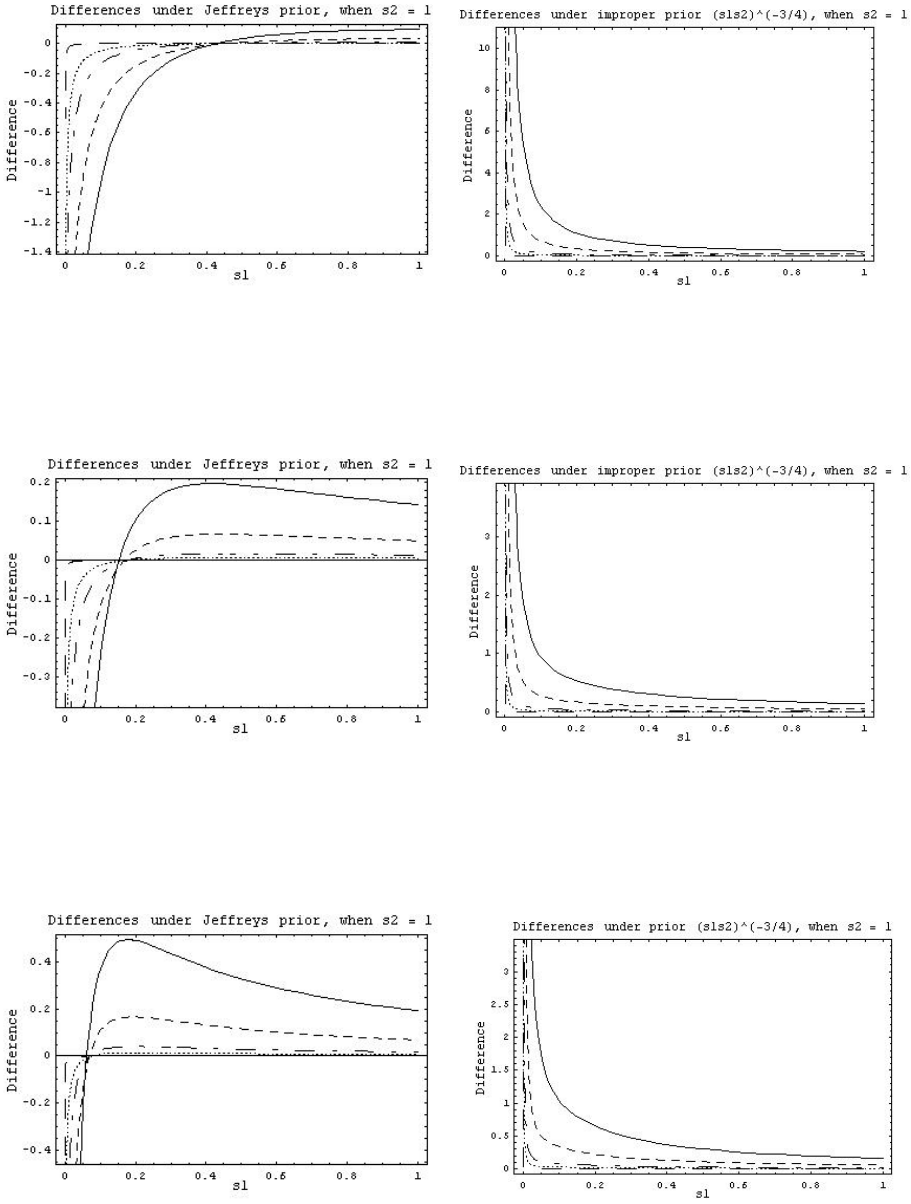

Figure 1: Plots of $g_1 - g_2$ against s_1 with $s_2 = 1$ for n =10 (thin solid line), 20 (broken line), 50 (broken/dotted line), 100 (dotted line),and 1000 (wider broken line). The plots on the left are under the Jeffreys prior, and those on the right are under the improper prior, with $\alpha = 0.75$. The three rows from top to bottom correspond to $m = 2, 5$, and 10 respectively.

prior with $\alpha = 0.75$ are for all the m and n combinations considered. For the Jeffreys prior, the threshold for the Bayesian predictive distribution to be better than the REML based estimative distribution is approximately $1/m$, which is also consistent with the asymptotic results.

4.3 Simulation Studies

In this section we compare both the Jeffreys prior and the proposed improper prior with the REML estimative density in terms of $E_{Y|\theta}(D(\varphi, \widehat{\varphi}) - D(\varphi, \tilde{\varphi}))$ using simulation. Section 4.3.1-4.3.3 gives the implementation details of the numerical procedure for conducting the simulation experiments, and the simulation results are summarized in 4.3.4.

4.3.1 REML estimator for (s_1, s_2)

To evaluate $E_{Y|\theta}(D(\varphi, \widehat{\varphi}))$, we need to plug in the REML estimator from every given data vector and specific $\theta(\theta = (s_1, s_2))$ in each iteration. For the simple random effect model, we can explicitly calculate the REML estimator for s_1, s_2. The restricted log-likelihood for $\theta = (s_1, s_2)$ is

$$\ell_n(\theta) = \tfrac{(1-m)n}{2} \log s_2 + \tfrac{1-n}{2} \log(s_2 + m s_1) - \tfrac{G^2}{2}, \tag{33}$$

with

$$\begin{aligned} G^2 = G^2(\theta) &= Y^T \{V^{-1} - V^{-1} X (X^T V X)^{-1} X^T V^{-1}\} Y \\ &= \tfrac{1}{s_2} \sum_{i,j} y_{i,j}^2 - \tfrac{1}{mn(s_2 + m s_1)} (\sum_{i,j} y_{i,j})^2 - \tfrac{s_1}{s_2(s_2 + m s_1)} \sum_i (\sum_j y_{i,j})^2 \quad . \end{aligned} \tag{34}$$

For general n and m,

$$\hat{s_1} = \tfrac{n \sum_{i=1}^n (\sum_{j=1}^m y_{i,j})^2 - (\sum_{i=1}^n \sum_{j=1}^m y_{i,j})^2}{m^2 n(n-1)} - \tfrac{m \sum_{i=1}^n \sum_{j=1}^m y_{i,j}^2 - \sum_{i=1}^n (\sum_{j=1}^m y_{i,j})^2}{m^2 n(m-1)}, \tag{35}$$

and

$$\hat{s_2} = \frac{m\sum_{i=1}^{n}\sum_{j=1}^{m} y_{i,j}^2 - \sum_{i=1}^{n}(\sum_{j=1}^{m} y_{i,j})^2}{(m-1)mn}. \tag{36}$$

It is easy to check that $\frac{\partial^2 \ell_n(\theta)}{\partial s_1^2}, \frac{\partial^2 \ell_n(\theta)}{\partial s_2^2} < 0$, and the standard deviation for these above REML estimators goes asymptotically to 0 as n goes into infinity.

4.3.2 Sampling for (β, s_1, s_2)

To calculate the predictive density function, we need to generate s_1, s_2 from the posterior distributions. For both the Jeffreys and improper priors, the marginal posterior is complicated. However, Metropolis-Hastings algorithm can be used to generate the posterior distributions as follows:

Step 1. Start with arbitrary s_1^0, s_2^0 from support of the posterior distribution, i.e. $(0, \infty)$.

Step 2. At stage n, generate proposal s_1^*, s_2^* from $q(s_1^*, s_2^* | s_1, s_2)$. The arbitrary proposal distribution is defined as $q(s_1^*, s_2^* | s_1, s_2) = \frac{1}{s_1 s_2} \exp\{-\frac{s_1^*}{s_1} - \frac{s_2^*}{s_2}\}$, the product of two exponentials with means s_1 and s_2.

Step 3. Take $s_1^{n+1} = s_1^*, s_2^{n+1} = s_2^*$ with probability $\alpha = \min\{\frac{q(s_1, s_2 | s_1^*, s_2^*)\pi_J(s_1^*, s_2^*) f(y | s_1^*, s_2^*)}{q(s_1^*, s_2^* | s_1, s_2)\pi_J(s_1, s_2) f(y | s_1, s_2)}, 1\}$. Otherwise, increase n and return to *Step 2*. This random acceptance is done by generating $u \sim$ Uniform (0, 1) and accepting the proposal s_1^*, s_2^* if $u \leq \alpha$.

We burn in 1000 out of 2000 simulations (usually 100-500 is enough) to make sure that there is no influence of the initial values for s_1 and s_2, so only 1000 variates have been used to approximate the posteriors, from which we select one in every ten and make records of 100 pairs of (s_1^*, s_2^*). The convergence is justified by the result of Gelman and Rubin's convergence diagnostic in the "coda" package of R language.

4.3.3 MC Method for integration of KL divergence

We evaluate $E_{Y|\theta}[D(\varphi, \widehat{\varphi}) - D(\varphi, \tilde{\varphi})]$ for fixed θ by the following algorithm.

step 1: Generate $Y^{(l)}$ for $l = 1, 2, \ldots, L$ using the model (28) for fixed θ.

step 2: For each $Y^{(l)}$, compute the REML estimator $\hat{\theta} = (\hat{s_1}, \hat{s_2})$ using (35) and (36), and the corresponding REML predictive density function is given by $\hat{\varphi}(z; Y) = \varphi(z; Y, \hat{\theta})$.

step 3: Approximate $\tilde{\varphi}(z; Y^{(l)})$ by $\frac{1}{n}\sum_i \varphi(z; Y^{(l)}, \theta^i)$, where θ^i is generated as described in 4.3.2, with Jeffreys and improper priors, respectively.

step 4: The difference between $D(\varphi, \hat{\varphi})$ and $D(\varphi, \tilde{\varphi})$ is approximated by quadrature integration method.

step 5: To calculate the expected KL divergence for fixed θ, we approximate it by $\frac{1}{L}\sum_l (D(\varphi, \widehat{\varphi}|Y^{(l)}, \theta) - D(\varphi, \tilde{\varphi}|Y^{(l)}, \theta))$, where the summation is taken over $Y^{(l)}$

We set L as 100 here, which is also justified by the convergence diagnostic in "coda" package of R programming.

4.3.4 Simulation Results

In the simulation studies, we set $s_2 = 1$, $\beta = 0$, $m = 2, 5, 10$, $n = 10, 20, 50, 100$, and $s_1 = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8$, and carried out the computation for both the Jeffreys prior and the improper prior with $\alpha = 0.75$. The results are summarized in Figure 2.

The first row of Figure 2 describes simulation results for $m = 2$. The left panel shows the results under Jeffreys prior, and the right one under the improper prior with $\alpha = 0.75$. The left panel indicates that under the Jeffreys prior and when s_1 is less than 0.5, the REML plug-in density performs better than the Bayesian predictive density in terms of KL divergence, while the Bayes predictive density performs better than the REML competitor otherwise. The right panel indicates that, under the improper prior the Bayesian predictive density always performs better than the REML estimative density. Both results are consistent with our theoretical findings.

The second row of Figure 2 gives simulation results for $m = 5$. The left panel indicates that when $m = 5$, the Bayesian predictive density under Jeffreys prior performs better than REML plug-in estimative density when $\frac{s_1}{s_2}$ is greater than some value around 0.2, and

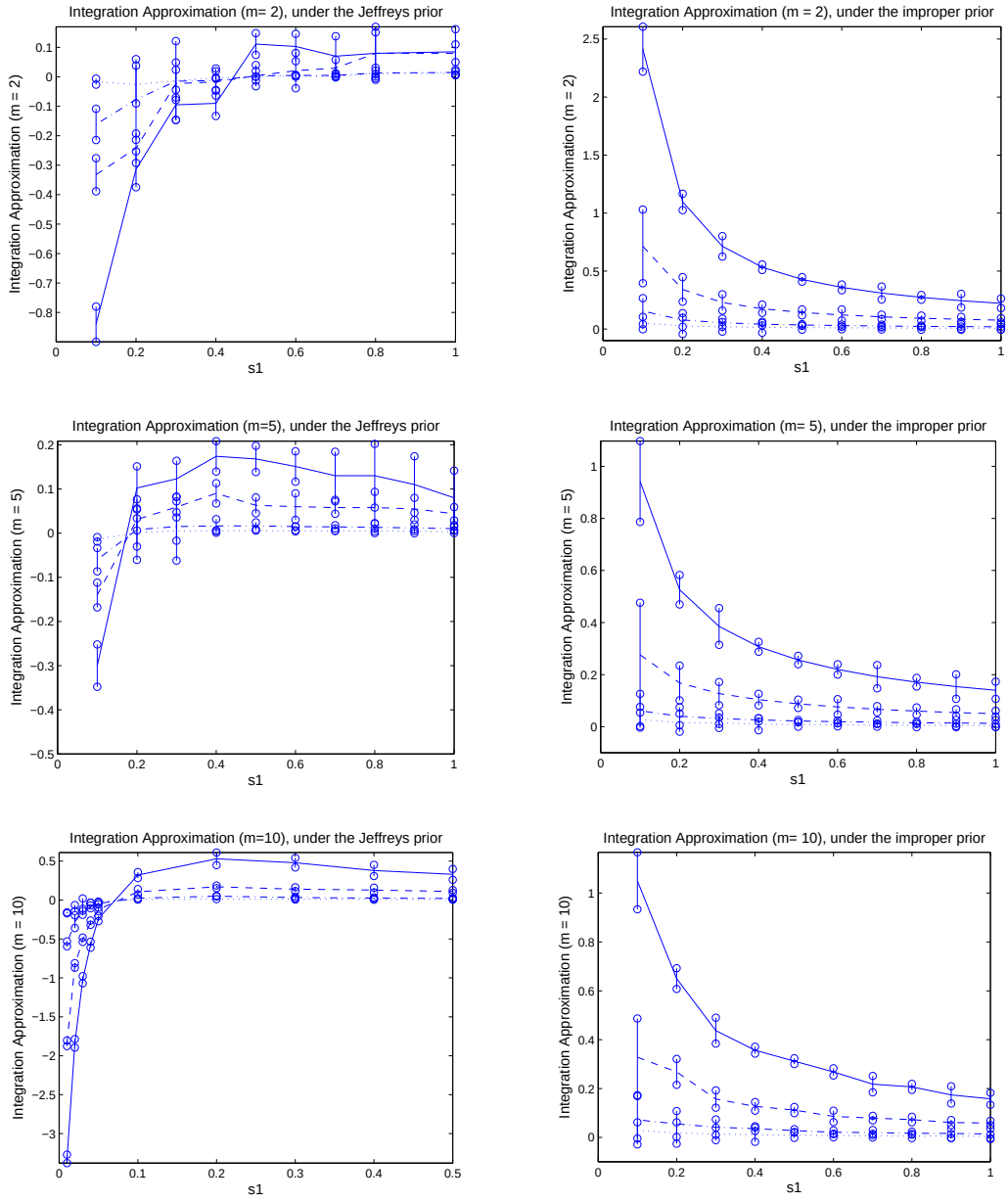

Figure 2: Simulation results of expected difference of KL divergence against s_1 with $s_2 = 1$ for n =10 (solid line), 20 (broken line), 50 (broken/dotted line), and 100 (dotted line). The plots on the left are under the Jeffreys prior, and those on the right are under the improper prior, with $\alpha = 0.75$. The three rows from top to bottom correspond to $m = 2, 5$, and 10 respectively.

the REML competitor performs better otherwise, which is consistent with the asymptotic results in Figure 1. The right panel displays simulation results under the improper prior with $\alpha = 0.75$, which indicates that the Bayesian predictive densities always performs better than the REML estimative density, which are also consistent with the theoretical results.

The third row of Figure 2 gives simulation results for $m = 10$. We obtain similar conclusions as for $m = 2$ and 5, except that in the left panel, the change point is around 0.1 this time.

5 Discussion

In this paper we used the asymptotic expansion of the KL divergence as the main tool to compare different predictive distributions, and derived some explicit results for one-way random effects models. In particular, we find a class of improper priors which lead to predictive distributions that are asymptotically superior to the REML based estimative distributions. It is reasonable to believe that similar results hold for more general mixed effects models, including the spatial linear models commonly used in geostatistics, though we expect the proof to be more difficult. The asymptotic expressions we derived for KL divergence is quite general, and can be used for other purpose, such as spatial sampling design in the context of spatial linear model.

GARCIA-DONATO and SUN (2007) discussed objective priors for hypothesis testing for one-way random effects models, and derived the divergence based (DB) prior and the intrinsic prior. Their work is related with ours, while their emphasis is on the use of these priors to develop consistent objective Bayesian factors, which is different from our purpose. It is interesting to check whether their priors are also second order KL REML dominant priors, and whether some of their priors can dominant other priors in the sense of second order KL divergence.

References

Aitchison, J. (1975), "Goodness of prediction fit," *Biometrika*, 62, 547.

Bleistein, N. and Handelsman, R. A. (1986), *Asymptotic Expansions of Integrals*, Courier Dover.

GARCIA-DONATO, G. and SUN, D. (2007), "Objective priors for hypothesis testing in one-way random effects models," *The Canadian Journal of Statistics*, 35, 303 – 320.

Hartigan, J. A. (1998), "The maximum likelihood prior," *The Annals of Statistics*, 26, 2083–2103.

Hobert, J. P. and Casella, G. (1996), "The effect of improper priors on Gibbs Sampling in Hierarchical Linear Mixed Models," *Jounal of the American Statistical Association*, 91, 1461 – 1473.

Komaki, F. (1996), "On asymptotic properties of predictive distributions," *Biometrika*, 83, 299–313.

K.V., M. and R.J., M. (1984), "Maximum likelihood estimation of models for residual covariance in spatial regression," *Biometrika*, 71, 135–146.

Murray, G. D. (1977), "A note on the estimation of probability density functions," *Biometrika*, 64, 150–2.

Ren, C., Sun, D., and Dey, D. K. (2006), "Bayesian and frequentist estimation and prediction for exponential distributions," *Jounal of Statistical Planning and Inference*, 136, 2873 – 2897.

Smith, R. L. and Zhu, Z. (2004), "Asymptotic Theory for Kriging with Estimated Parameters and Its Application to Network Design," .

Stein, M. L. (1999), *Interpolation of Spatial Data: Some Theory for Kriging*, Springer.