Predictive Inference in Linear Mixed Models

Gourab Mukherjee
Department of Data Sciences and Operations
University of Southern California
and
Keisuke Yano
The Institute of Statistical Mathematics

July 24, 2025

Abstract

We consider predictive inference in Linear Mixed Models (LMMs). Specifically, we study the problem of estimating the predictive density under Kullback-Leibler (KL) loss in LMMs with a large number of units. We consider flexible classes of empirical Bayes (EB) predictive densities and develop a novel risk estimation based methodology for selecting hyper-parameters of EB predictive density estimates. Our risk estimation based hyper-parameter selection methodology uses the decision-theoretic identity in Lemma 2 of George et al. [2006] that connects predictive KL risk for density estimation to Stein's unbiased estimate of the quadratic risk in point estimation. Direct construction of unbiased KL risk estimates is not possible in LMMs. We leverage information in the covariates and exchangeability of the individual effects to construct asymptotically efficient estimates of the KL risks for wide classes of predictive density estimators. We derive the rates of convergences of the proposed KL risk estimates (KLRE) and show that EB predictors calibrated by minimizing KLREs are asymptotically optimal in LMMs.

Keywords: Empirical Bayes; Kullback-Leibler loss; Predictive density estimation; Oracle optimality; Shrinkage; Risk estimation; Predictive inference; Linear Mixed Models.

1 Introduction

We consider the problem of estimating the predictive density [Aitchison and Dunsmore, 1975, Geisser, 1993] in a linear mixed model [Jiang and Nguyen, 2007, McCulloch and Searle, 2004, Pinheiro and Bates, 2000]. Linear mixed models (LMMs) are widely used across various empirical fields of study for their capability to handle data that do not meet the assumptions of traditional linear regression models [Demidenko, 2013, Verbeke and Molenberghs, 2012]. LMMs allow fixed effects that are common to all units as well as random effects which are specific to individual units, and thus, provide a flexible modeling framework for complex, highly heterogeneous big data analysis [Faraway, 2016].

Consider the following LMM. For each unit $i = 1, \ldots, n$, we observe K_i replicates X_{ik}. The expected value of the response X_{ik} is linearly related to covariates $\boldsymbol{a}_{ik}$ and u_{ik} by parameters $\boldsymbol{\beta}$ and γ_i respectively. The model is:

$$X_{ik} = \boldsymbol{a}_{ik}'\boldsymbol{\beta} + \gamma_i u_{ik} + \sigma_i\, \epsilon_{ik} \text{ for } k = 1, \ldots, K_i \text{ and, } i = 1, \ldots, n. \tag{1}$$

The parameter $\boldsymbol{\beta}$ is global and invariant across replicates as well as units whereas the parameters γ_i changes over units. We use bold symbols to denote vectors. While $\boldsymbol{\beta}$ can be multivariate, γ_is are scalars and (1) has n unit-specific slope coefficients. The noise terms ϵ_{ik}s are independently distributed from standard normal distribution. Based on observing $\{X_{ik} : k = 1, \ldots, K_i;\, i = 1, \ldots, n\}$, consider predicting future observations Y_{ik} from units $i = 1, \ldots, n$ based on the following model:

$$Y_{ik} = \tilde{\boldsymbol{a}}_{ik}'\boldsymbol{\beta} + \gamma_i v_{ik} + \tilde{\sigma}_i\, \tilde{\epsilon}_{ik} \text{ for } k = 1, \ldots, \tilde{K}_i \text{ and, } i = 1, \ldots, n. \tag{2}$$

The covariates in (1) are allowed to change in (2). However, the parameters are unchanged. The noise terms $\tilde{\epsilon}_{ik}$ have standard normal distributions and are independent not only across (i, k) pairs but also of the noise terms in (1). The variances $\sigma_i^2, \tilde{\sigma}_i^2$ are known but can vary across units. The variances of the future observations in (2) are allowed to be different from the variances of the past observations in (1).

In (1)-(2), we consider estimating the predictive density of $\boldsymbol{Y} = \{Y_{ik} : k = 1, \ldots, \tilde{K}_i;\, i = 1, \ldots, n\}$ based on observing $\boldsymbol{X} = \{X_{ik} : k = 1, \ldots, K_i;\, i = 1, \ldots, n\}$. Let $\boldsymbol{\mu} = (\boldsymbol{\beta}, \gamma_i : i =$

$1, \ldots, n$). The covariates and variances of (1) are stored in A, U, Σ with proper indexing and those of (2) are stored in $\tilde{A}$, V, $\tilde{\Sigma}$. Let $C = (A, U, \Sigma)$ and $\tilde{C} = (\tilde{A}, U, \tilde{\Sigma})$. Denote the true density of $\boldsymbol{X}$ and $\boldsymbol{Y}$ conditioned on $\boldsymbol{\mu}$ by $\boldsymbol{p}_{\boldsymbol{\mu}}(\boldsymbol{X}; C)$ and $\boldsymbol{p}_{\boldsymbol{\mu}}(\boldsymbol{Y}; \tilde{C})$ respectively. Note that, both $\boldsymbol{p}_{\boldsymbol{\mu}}(\boldsymbol{X}; C)$ and $\boldsymbol{p}_{\boldsymbol{\mu}}(\boldsymbol{Y}; \tilde{C})$ are Gaussian densities. Our objective is to choose a probability distribution that will effectively predict the behavior of future samples $\boldsymbol{Y}$. We construct an estimate $\hat{\boldsymbol{p}}(\boldsymbol{Y}; \boldsymbol{X}, C, \tilde{C})$ of the future conditional density of $\boldsymbol{Y}$ based on $\boldsymbol{X}$ which is referred to as the predictive density [Seymour, 1971].

Estimating the predictive density is a very important problem in statistical prediction analysis [see Liang, 2002, Mukherjee, 2013, Xu, 2005 and the references therein]. Unlike point estimation methods, predictive densities assign probabilities to all potential future outcomes and thus, offer superior risk assessment and decision-making capabilities [Geisser, 1993]. Predictive densities find application across various domains such as weather forecasting [Taylor and Buizza, 2004], finance [Tay and Wallis, 2000], information theory [Barron et al., 1998, Liang and Barron, 2005, Yuan and Clarke, 1999] as well as for model diagnostics and validation [Gelman et al., 2013, 2014, Pardoe, 2001, Yano and Komaki, 2017a].

To evaluate performance of the predictive density estimate (prde) $\hat{\boldsymbol{p}}$ we use the familiar Kullback–Leibler (KL) distance [Cover, 1999] as the loss function:

$$L(\boldsymbol{\mu}, \hat{\boldsymbol{p}}(\,\cdot\,; \boldsymbol{x}, C, \tilde{C})) = \left(\sum_{i=1}^{n} \tilde{K}_i \right)^{-1} \int \boldsymbol{p}_{\boldsymbol{\mu}}(\boldsymbol{Y}; \tilde{C}) \, \log \left(\frac{\hat{\boldsymbol{p}}(\boldsymbol{Y}; \boldsymbol{x}, C, \tilde{C})}{\boldsymbol{p}_{\boldsymbol{\mu}}(\boldsymbol{Y}; \tilde{C})} \right) d\boldsymbol{Y} \ .$$

The corresponding KL risk follows by averaging over the distribution of the past observations:

$$R(\boldsymbol{\mu}, \hat{\boldsymbol{p}}\,;\, C, \tilde{C}) = \int \boldsymbol{p}_{\boldsymbol{\mu}}(\boldsymbol{X}; C) \, L(\boldsymbol{\mu}, \hat{\boldsymbol{p}}(\,\cdot\,; \boldsymbol{X}, C, \tilde{C})) \, d\boldsymbol{X} \ . \tag{3}$$

Aslan [2006], Hartigan [1998], Komaki [2001] characterized the asymptotic KL risk in fixed dimensions through the use of information geometry. In particular, Hartigan [1998] showed that the KL risk of MLE-based plug-in predictive density asymptotically equals the risk of any regular Bayes prde in large samples. However, when dimension increases with sample size, this plug-in and Bayes congruence collapses and plug-in methods can be highly sub-optimal [George et al., 2012, Mukherjee and Johnstone, 2015]. For predictive density

estimation in multivariate Gaussian models, the canonical minimax prde, which is the Bayes prde under uniform prior, is not admissible for dimensions greater than 2 [George et al., 2006, Ghosh et al., 2008, Komaki, 2001]. In multivariate models, the role of shrinkage priors is critical for constructing efficient prdes [Brown et al., 2008, Gangopadhyay and Mukherjee, 2021, Ghosh et al., 2011, Komaki, 2004, Maruyama and Strawderman, 2012, Matsuda and Komaki, 2015, Matsuda and Strawderman, 2022, Mukherjee and Johnstone, 2022, Yano et al., 2021]. The role of shrinkage in predictive density estimation and its connection with point-estimation under quadratic loss have been established in [Fourdrinier et al., 2011, Kubokawa et al., 2013, Mukherjee and Johnstone, 2015, Xu and Liang, 2010].

We consider prdes in (1)-(2) where there are a large number of units, i.e., in regimes where $n \to \infty$. Thus, the resultant model is high-dimensional. However, this high-dimensional parametric space has a low-dimensional structure in LMMs as we assume that $\boldsymbol{\gamma} = \{\gamma_i : 1 \leq i \leq n\}$ are independent and identically distributed (i.i.d.) according to a higher level prior g. Hierarchical modeling using the second level prior provides a framework to combine information across units and pooling information from similar resources can provide significant gains in inference.

The conditional Bayes risk of prior g for any fixed $\boldsymbol{\beta}$ is given by:

$$R((\boldsymbol{\beta}, g), \hat{\boldsymbol{p}}\,;\, C, \tilde{C}) = \int R(\boldsymbol{\mu}, \hat{\boldsymbol{p}}\,;\, C, \tilde{C})\, g(\gamma_1) \ldots g(\gamma_n)\, d\boldsymbol{\gamma}\ . \tag{4}$$

For any fixed $\boldsymbol{\beta}$ this risk is minimized by the conditional Bayes prde $\hat{\boldsymbol{p}}^g$:

$$\hat{\boldsymbol{p}}^g[\boldsymbol{\beta}, C, \tilde{C}](\boldsymbol{Y}; \boldsymbol{X}) = \int \boldsymbol{p}_{(\boldsymbol{\beta},\boldsymbol{\gamma})}(\boldsymbol{Y}; \tilde{C}) g(\boldsymbol{\gamma}|\boldsymbol{X}; C, \boldsymbol{\beta}) d\boldsymbol{\gamma}, \tag{5}$$

where, $g(\boldsymbol{\gamma}|\boldsymbol{X}; C, \boldsymbol{\beta})$ is the posterior distribution of $\boldsymbol{\gamma}$ for any fixed $\boldsymbol{\beta}$ based on the observations and is given by:

$$(m^g(\boldsymbol{X}; C, \boldsymbol{\beta}))^{-1} \prod_{i=1}^{n} g(\gamma_i) \boldsymbol{p}_{(\boldsymbol{\beta},\boldsymbol{\gamma})}(\boldsymbol{X}; C)\,, \text{ where, } m^g(\boldsymbol{X}; C, \boldsymbol{\beta}) = \int \prod_{i=1}^{n} g(\gamma_i) \boldsymbol{p}_{(\boldsymbol{\beta},\boldsymbol{\gamma})}(\boldsymbol{X}; C) d\boldsymbol{\gamma}$$

is the marginal distribution based on the prior g and the observed data.

As g is not known, we can not directly use the conditional Bayes prde $\hat{\boldsymbol{p}}^g$. We undertake an empirical Bayes (EB) approach [Efron, 2012] to construct an efficient prde. We consider

a family of distributions $\mathcal{G}$ for the prior g. Let the distributions g_h in some class $\mathcal{G}$ be indexed by a hyper-parameter $h \in \mathcal{H}$. The class of all Bayes prdes for this class $\mathcal{G}$ of priors is

$$\mathcal{B}_{\mathcal{G}} = \big\{\hat{\boldsymbol{p}}^h[\boldsymbol{\beta}, C, \tilde{C}](\boldsymbol{Y}; \boldsymbol{X}) : h \in \mathcal{H}\big\}.$$

For ease of presentation, we switched the sub-script of the prdes from explicitly referencing to the prior g_h to only its index h in the class. Correspondingly the frequentist risk (3) of such an estimator is denoted by $R(\boldsymbol{\mu}, h\,;\, C, \tilde{C})$.

For popular families $\mathcal{B}_{\mathcal{G}}$ is a wide class of prdes and our goal here is to select the optimal member from the class that minimizes the KL risk. We construct estimates $\hat{R}(h\,;\, C, \tilde{C})$ of the risk that are uniformly efficient for all $h \in \mathcal{H}$ and propose to select $\hat{h} = \arg\min_h \hat{R}(h\,;\, C, \tilde{C})$. We show that such a choice is efficient as the resultant estimator is asymptotically optimal.

Empirical Bayes (EB) predictive density estimation in Gaussian sequence models (which are much simpler versions of (1)-(2) with $\boldsymbol{\beta} = 0, K_i = \tilde{K}_i = 1$ and $u_{ik} = v_{ik} = 1$) has been studied in George et al. [2021], Xu and Zhou [2011]. The most natural approach in prde is to use a non-informative prior for g in (5). The resultant conditional Bayes prde $\hat{\boldsymbol{p}}^{\mathtt{U}}$ is minimax that has constant risk for any $\boldsymbol{\gamma}$ [Liang and Barron, 2004]. However, shrinkage prdes can have much lower risk than $\hat{\boldsymbol{p}}^{\mathtt{U}}$ over large subsets of the parametric space and asymptotically ($n \to \infty$) no worse risk over the entire space. EB provides a disciplined framework for constructing shrinkage prdes by calibrating Bayes prde from a wide family of priors. Risk estimation [Fourdrinier et al., 2018] prescribes calibration of EB prdes by minimizing unbiased risk estimates. However, unlike in point estimation, constructing unbiased estimates for the KL risk of prdes is not straight forward. If $\mathcal{G}$ is a Gaussian family of priors then unbiased estimates of the KL risk can be derived [George et al., 2021]. However, for non-Gaussian families of priors constructing unbiased estimates of the KL risk is not straightforward. The main difficulty is that the connections between the unbiased risk estimates for the KL risk and quadratic risk of point-estimation (see (23) in George et al., 2012) is through an integral equation that involves unobserved quantities. So, while one

of Stein's lemmas [Stein, 1981] readily yields unbiased risk estimates of point estimators in Gaussian location models we can not use them for KL risk of prdes. Here, we leverage the information in the covariate associated with the slope coefficient $\boldsymbol{\gamma}$ to construct an estimate of the predictive KL risk that is asymptotically unbiased for exchangeable $\boldsymbol{\gamma}$. We derive large sample characteristics of these risk estimates and demonstrate their applicability for constructing EB prdes in large LMMs.

The paper is organized as follows. In Sec. 2, we present shrinkage prdes and derive their KL risk. In Sec. 3, we present the proposed risk estimation methodology and its decision theoretic properties. We provide a procedure to estimate the KL risk of shrinkage prdes. In Sec 3.1 we quantify the sampling fluctuations of these risk estimates in LMMs with large number of units. In Sec. 3.3, we provide the asymptotic properties of shrinkage prdes from popular prior families that are calibrated using the proposed risk estimation method. We end with a discussion in Sec. 4. All the proofs of the results in this paper are provided in the Appendix.

2 Shrunken Predictive Density Estimators

In this section, we present shrinkage classes of prdes and derive their KL risk in LMMs. The risk of shrunken prdes has been widely studied in recent works on predictive inference [Bhagwat and Marchand, 2023, George et al., 2019, Ghosh et al., 2019, L'Moudden and Marchand, 2019, Marchand and Sadeghkhani, 2018, Yano and Komaki, 2017b]. To rigorously study their risk in LMMs, we next assume $\boldsymbol{\beta}$ known and use the principles of sufficiency [Lehmann and Casella, 2006] to reduce (1) to a cross-sectional model:

$$\dot{X}_i = \left(\sum_{k=1}^{K_i} u_{ik}^2 \right)^{-1} \left(\sum_{k=1}^{K_i} u_{ik}(X_{ik} - \boldsymbol{a}_{ik}'\boldsymbol{\beta}) \right) \text{ for } i = 1, \ldots, n.$$

Note that $\dot{X}_i$ follows a normal distribution with mean γ_i and variance $\sigma_i^2/(\sum_k u_{ik}^2)$ and also $\{\dot{X}_i : 1 \leq i \leq n\}$ are independent among themselves. Denote: $u_i = (\sum_{k=1}^{K_i} u_{ik}^2)^{1/2}$. Thus,

we have

$$\dot{X}_i = \gamma_i + u_i^{-1}\sigma_i\varepsilon_i \text{ for } i = 1, \ldots, n. \tag{6}$$

Also, for the conditional prediction model we predict $\dot{Y}_{ik} = Y_{ik} - \tilde{\boldsymbol{a}}'_{ik}\boldsymbol{\beta}$ where,

$$\dot{Y}_{ik} = v_{ik}\gamma_i + \tilde{\sigma}_i\tilde{\varepsilon}_{ik} \text{ for } i = 1, \ldots, n \text{ and } k = 1, \ldots, \tilde{K}_i. \tag{7}$$

The noise term ε and $\tilde{\varepsilon}$ are i.i.d. from standard normal distributions. $\boldsymbol{\gamma}$ is the only parameter in conditional prediction model (6)-(7). When we consider the conditional model, we use dots on X and Y to mark that they are centered variables using information of A, $\tilde{A}$ and $\boldsymbol{\beta}$.

The conditional (fixed $\boldsymbol{\beta}$) Bayes prde based on $\dot{\boldsymbol{X}} = \{\dot{X}_i : 1 \leq i \leq n\}$ from (6) decouples as a product of co-ordinatewise Bayes prdes $\hat{p}_g[(u_i, \sigma_i), (v_{ik}, \tilde{\sigma}_i)](Y_{ik}; X_i)$, which is given by:

$$\left(\int \dot{p}_\gamma(X_i; u_i, \sigma_i) g(\gamma)\, d\gamma \right)^{-1} \int \dot{p}_\gamma(Y_{ik}; v_{ik}, \tilde{\sigma}_i)\, \dot{p}_\gamma(X_i; u_i, \sigma_i)\, g(\gamma)\, d\boldsymbol{\gamma}$$

where, $p_\gamma(Y_{ik}; v_{ik}, \tilde{\sigma}_i)$ and $\dot{p}_\gamma(X_i; u_i, \sigma_i)$ are the true densities based on (7) and (6) respectively. Noting, both of them are normal densities we further simplify the expressions and summarize the result in the following proposition. We denote normal densities with mean m and standard deviation s by $\phi(\cdot|m, s)$.

Proposition 1 *If $\gamma_1, \ldots, \gamma_n$ i.i.d. from g then for any fixed β the prde of $\dot{\boldsymbol{Y}}$ in* (7) *based on observing $\dot{\boldsymbol{X}}$ from* (6)*, that minimizes* (4) *is given by:*

$$\hat{\boldsymbol{p}}^g(\dot{\boldsymbol{Y}}; \dot{\boldsymbol{X}}) = \prod_{i=1}^{n} \prod_{k=1}^{\tilde{K}_i} \frac{\int \phi(\dot{Y}_{ik}; v_{ik}\gamma, \tilde{\sigma}_i)\phi(v_{ik}\dot{X}_i; v_{ik}\gamma, u_i^{-1}v_{ik}\sigma_i)g(\gamma)\, d\gamma}{\int \phi(v_{ik}\dot{X}_i; v_{ik}\gamma, u_i^{-1}v_{ik}\sigma_i)g(\gamma)d\gamma} \ .$$

As the conditional Bayes prde is a product rule, for conducting risk analysis the fundamental calculations follows from coordinatewise analysis. We next present the corresponding univariate predictive model and risk properties of aforementioned prdes in such models.

2.1 Properties of Univariate Predictive KL Risk

Consider the following univariate predictive problem where we observe $\ddot{X}$ and would like to predict $\ddot{y}$ based on the following model:

$$\ddot{x} = v\gamma + u^{-1}v\sigma\varepsilon \text{ and } \ddot{y} = v\gamma + \tilde{\sigma}\tilde{\varepsilon} \tag{8}$$

Note that compared to (6), the above model has a scale change in the past observations whereas the model for the future observations is just the univariate version of (7). This is done so that the past observation is adjusted to have the same expected value as that of the future. We use double dots to represent this model and to distinguish it from (6).

Define $r = \tilde{\sigma}^2/\sigma^2$. When g is uniform, the Bayes prde in (8) is given by:

$$\hat{p}^{\mathtt{U}}(\ddot{y}; \ddot{x}, u, v) = \phi(\ddot{y}; \ddot{x}, (r + v^2u^{-2})\sigma^2) \tag{9}$$

and the frequentist risk of $\hat{p}^{\mathtt{U}}$ as in (3) but in the setting of (8) is given by:

$$r(\gamma, \hat{p}^{\mathtt{U}}; u, v) = 2^{-1}\log(1 + v^2u^{-2}r^{-1}). \tag{10}$$

For the univariate case, we keep the dependence of σ and r in the risk function implicit. Next, consider the following variable which is a linear combination of $\ddot{X}$ and $\ddot{Y}$:

$$\ddot{W}(u, v, r) = r(r + v^2u^{-2})^{-1}\ddot{X} + v^2u^{-2}(r + v^2u^{-2})^{-1}\ddot{Y}.$$

$\ddot{W}$ is the the uniformly minimum variance unbiased estimator of γ if both $\ddot{X}$ and $\ddot{Y}$ were observed. As $\ddot{Y}$ is not observed, $\ddot{W}$ is not really an estimator. It connects Bayes prdes to $\hat{p}^{\mathtt{U}}$. As such it follows directly from Lemma 2 of George et al. [2006] that the Bayes prde in (8) for any prior g is given by:

$$\hat{p}^{g}(\ddot{y}|\ddot{x}, u, v) = q^{g}(\ddot{w}; u, v)/m^{g}(\ddot{x}; u, v) * \hat{p}^{\mathtt{U}}[u, v](\ddot{y}; \ddot{x}) \ , \tag{11}$$

where, q and m are the marginal distributions of $\ddot{x}$ and $\ddot{w}$ respectively. Note that, $[\ddot{X}|\gamma, u, v] \stackrel{d}{=} N(v\gamma, u^{-2}v^2\sigma^2)$ and $[\ddot{W}|\gamma, u, v, r] \stackrel{d}{=} N(v\gamma, \sigma^2/(r^{-1} + u^2v^{-2}))$, thus,

$$m^{g}(\ddot{x}; u, v, \sigma) = \int \phi(\ddot{x}; v\gamma, u^{-1}v\sigma)g(\gamma)d\gamma \text{ and,} \tag{12}$$

$$q^{g}(\ddot{w}; u, v, \sigma, r) = \int \phi(\ddot{w}; v\gamma, \sigma/(r^{-1} + u^2v^{-2})^{1/2})g(\gamma)d\gamma \tag{13}$$

The risk of any Bayes prde can be written in terms of the expected values of the log-marginal likelihood of $\ddot{X}$ and the predictive log-likelihood of $\ddot{W}$.

Proposition 2 *The risk of the Bayes prde in* (8) *based on prior g is given by:*

$$\ddot{R}(\gamma, \hat{p}^g; u, v) = \frac{1}{2}\log(1 + r^{-1}u^{-1}v) - \mathbb{E}_\gamma \log q^g(\ddot{W}; u, v, \sigma, r) + \mathbb{E}_\gamma \log m^g(\ddot{X}; u, v, \sigma) \ .$$

2.2 Risk of Conditional Bayes prdes

Define $r_i = \tilde{\sigma}_i^2/\sigma_i^2$. Then, from (10) and Proposition 1 we obtain the risk of conditional Bayes prde in (7) based on uniform prior on $\boldsymbol{\gamma}$.

Proposition 3 *For any fixed $\boldsymbol{\beta}$, the conditional risk of the uniform prior based Bayes prde in* (7) *is:*

$$\dot{R}(\boldsymbol{\gamma}, \hat{\boldsymbol{p}}^U) = \left(2\sum_{i=1}^n \tilde{K}_i\right)^{-1} \sum_{i=1}^n \sum_{k=1}^{\tilde{K}_i} \log\left(1 + \frac{v_{ik}^2}{u_i^2} \cdot \frac{\sigma_i^2}{\tilde{\sigma}_i^2}\right) .$$

Similarly, extending proposition 2 for univariate predictive risk to the risk of product prdes in (6)-(7) we obtain the conditional risk of any Bayes prde. We report the difference of the risks of Bayes prdes from an arbitrary prior g and that of the uniform prior based prde.

Proposition 4 *For any fixed $\boldsymbol{\beta}$, the conditional risk of the Bayes prde based on prior g for predicting $\dot{\boldsymbol{Y}}$ from* (7) *based on observing $\dot{\boldsymbol{X}}$ from* (6) *is:*

$$2\big[\, \dot{R}(\boldsymbol{\gamma}, \hat{\boldsymbol{p}}^g) - \dot{R}(\boldsymbol{\gamma}, \hat{\boldsymbol{p}}^U) \,\big] = \dot{R}_1(\boldsymbol{\gamma}, g) - \dot{R}_2(\boldsymbol{\gamma}, g), \ \textit{where,}$$

$$\dot{R}_1(\boldsymbol{\gamma}, g) = \left(\sum_{i=1}^n \tilde{K}_i\right)^{-1} \sum_{i=1}^n \sum_{k=1}^{\tilde{K}_i} \mathbb{E}_{\gamma_i}\big\{\log m^g(v_{ik}\dot{X}_i; u_i, v_{ik}, \sigma_i)\big\} \tag{14}$$

$$\dot{R}_2(\boldsymbol{\gamma}, g) = \left(\sum_{i=1}^n \tilde{K}_i\right)^{-1} \sum_{i=1}^n \sum_{k=1}^{\tilde{K}_i} \mathbb{E}_{\gamma_i}\big\{\log q^g(\dot{W}_{ik}; u_i, v_{ik}, \sigma_i, r_i)\big\}, \tag{15}$$

and, $\dot{W}_{ik} = r_i(r_i + v_{ik}^2 u_i^{-2})^{-1} v_{ik}\dot{X}_i + v_{ik}^2 u_i^{-2}(r_i + v_{ik}^2 u_i^{-2})^{-1}\dot{Y}_{ik}$ *for* $k = 1, \ldots, \tilde{K}_i$ *and* $i = 1, \ldots, n$.

Ideally, we would like to minimize the excess risk of Bayes prdes $\hat{\boldsymbol{p}}^g$ with respect to $\hat{\boldsymbol{p}}^{\mathtt{U}}$ over the class $\mathcal{G}$ of priors. However, as $\boldsymbol{\gamma}$ is not known, direct evaluations of $\dot{R}_1$ and $\dot{R}_2$ are not possible. We propose to estimate $\dot{R}_1(\boldsymbol{\gamma}, g)$ and $\dot{R}_1(\boldsymbol{\gamma}, g)$ respectively by $\hat{R}_1^g$ and $\hat{R}_2^g$.

Before proceeding further, we impose the following assumption on (2):

Assumption A1: $\quad 1 \leq \inf_i \tilde{K}_i \leq \sup_i \tilde{K}_i \leq \kappa < \infty$.

The assumption means we are only predicting a few replicates for each unit and thus, no unit dominates the other units in terms of prediction tasks. It is a benign assumption that protects us against the imbalance among the units and facilitates the presentation of rigorous proof.

Consider the following estimator for $\dot{R}_1(\boldsymbol{\gamma}, g)$:

$$\hat{R}_1^g = \Big(\sum_{i=1}^n \tilde{K}_i \Big)^{-1} \sum_{i=1}^n \sum_{k=1}^{\tilde{K}_i} \log m^g(v_{ik}\dot{X}_i; u_i, v_{ik}, \sigma_i). \tag{16}$$

As $\dot{X}_i$ are independent and by assumption 1, $\dot{R}_1(\boldsymbol{\gamma}, g)$ can be well estimated by $\hat{R}_1^g$. As such, for well-behaved prior classes $\mathcal{G}$, we will show that $\hat{R}_1^g$ convergences to the true $\dot{R}_1$ at a near-parametric rate as $n \to \infty$, i.e., $|\dot{R}_1(\boldsymbol{\gamma}, g) - \hat{R}_1^g| = O_p(n^{-1/2} \log\log n)$. A formal proof is presented in the following section.

To provide a reasonable estimate of $\dot{R}_2$ is difficult. We can not use the approach used in constructing $\hat{R}_1$ as unlike $\dot{X}_i$ we do not observe $\dot{W}_{ik}$. However, if $\mathcal{G}$ is a simple family of priors such as $\mathcal{G} = \{\phi(\,\cdot\,|0, \tau^{1/2}) : \tau > 0\}$, then unbiased estimates of $\dot{R}_2$ can be constructed. In this case the risk function reduces to terms involving second moments of $\dot{W}$ which can be easily calculated leading to quadratic functions of $\boldsymbol{\gamma}$. The difference $\dot{R}_1(\boldsymbol{\gamma}, g) - \dot{R}_2(\boldsymbol{\gamma}, g)$ equals:

$$-\Big(\sum_{i=1}^n 2\tilde{K}_i \Big)^{-1} \sum_{i=1}^n \sum_{k=1}^{\tilde{K}_i} \Big\{ \log \frac{(\tau + \sigma_i^2 u_i^{-2})(r_i^{-1} v_{ik}^2 + u_i^2)}{\tau(r_i^{-1} v_{ik}^2 + u_i^2) + \sigma_i^2} + \frac{r_i^{-1} v_{ik}^2 \sigma_i^2 (\gamma_i^2 - \tau)}{(\tau u_i^2 + \sigma_i^2)\{\tau(r_i^{-1} v_{ik}^2 + u_i^2) + \sigma_i^2\}} \Big\} \tag{17}$$

and it can be estimated by replacing γ_i^2 above by $\dot{X}_i^2 - u_i^{-2}\sigma_i^2$. By theorems 4.1 and 4.2 of George et al. [2021] we know that this estimate convergences at $\sqrt{n}$ rate. For most popular

families of priors we do not have a natural estimate of $\dot{R}_2$. In the following section, we develop a novel method for estimating $\dot{R}_2$.

3 Proposed Methodology

3.1 KL Risk Estimation

Note that, though $\dot{W}_{ik}$ has the same mean as $v_{ik}\dot{X}_i$ it has much lower variance. To construct an estimate of $\dot{R}_2$, we plan to:
(a) first construct estimates of $\dot{W}_{ik}$ based on $\{\dot{X}_j : j = 1, \ldots, n\}$ such that those estimates each have the same variance as $\dot{W}_{ik}$. This is possible as u_is, σ_is varies over i in (6).
For each $i = 1, \ldots, n$ and $k = 1, \ldots, \tilde{K}_i$ define the set:

$$S_{ik} = \{j \in \{1, \ldots, n\} : \sigma_i^2 \sigma_j^{-2} u_j^2 - u_i^2 - r_i^{-1} v_{ik}^2 \geq 0\} \ .$$

For all $j \in S_{ik}$ define,

$$d_{ikj} = u_j^{-2} v_{ik}^2 \sigma_j^2 \{\sigma_i^2 \sigma_j^{-2} (v_{ik}^2 + r_i u_i^2)^{-1} r_i u_j^2 - 1\}.$$

By construction note that $d_{ikj} \geq 0$. Next, for each $i = 1, \ldots, n$ and $k = 1, \ldots, \tilde{K}_i$ and $j \in S_{ik}$, we construct a new variable by adding scaled i.i.d. standard normal noise Z_{ikj}:

$$\dot{T}_{ikj} = v_{ik}\dot{X}_j + d_{ikj}^{1/2} Z_{ijk} \ .$$

Note that, by construction we have: $\mathrm{Var}(\dot{T}_{ikj})$=$\mathrm{Var}(\dot{W}_{ik})$ and $\mathbb{E}(\dot{T}_{ikj} - \dot{W}_{ik}) = v_{ik}(\gamma_j - \gamma_i)$.
(b) next, we replace the $\dot{W}_{ik}$ in $\dot{R}_2$ by $\dot{T}_{ijk}$ in S_{ik}. Let $|S_{ik}|$ denote the cardinality of the set S_{ik}. For the time being, consider S_{ik}s are non-empty for all (i, j). Then, we replace $\mathbb{E}_{\gamma_i}\big\{ \log q^g(\dot{W}_{ik}; u_i, v_{ik}, \sigma_i, r_i)\big\}$ in (15) by

$$\rho_{ik}(\boldsymbol{\gamma}, g) = |S_{ik}|^{-1} \sum_{j \in S_{ik}} \mathbb{E}_{\gamma_j}\big\{ \log q^g(\dot{T}_{ikj}; u_i, v_{ik}, \sigma_i, r_i)\big\}$$

and consider,

$$\dot{R}_3(\boldsymbol{\gamma}, g) = \left(\sum_{i=1}^{n} \tilde{K}_i \right)^{-1} \sum_{i=1}^{n} \sum_{k=1}^{\tilde{K}_i} \rho_{ik}(\boldsymbol{\gamma}, g)$$

as a surrogate for $\dot{R}_2(\boldsymbol{\gamma}, g)$. Next, we construct an estimate of $\dot{R}_3(\boldsymbol{\gamma}, g)$.

Note, that $\dot{T}_{ikj}$ is a random quantity as it was constructed by adding scaled white noise. We denote its dependence on white noise by using $\dot{T}_{ikj}(Z)$. Consider enumerating $\mathbb{E}_Z\big\{ \log q^g(\dot{T}_{ikj}(Z); u_i, v_{ik}, \sigma_i, r_i)\big\}$ where the expectation of the predictive log-likelihood is taken over white noise and so, the resultant quantity is not random (once the data $\dot{\boldsymbol{X}}$ is fixed). We propose the following estimate of $\rho_{ik}(\boldsymbol{\gamma}, g)$:

$$\hat{\rho}^g_{ik} = |S_{ik}|^{-1} \sum_{j \in S_{ik}} \mathbb{E}_Z\big\{ \log q^g(\dot{T}_{ikj}(Z); u_i, v_{ik}, \sigma_i, r_i)\big\} \ .$$

Finally, we consider the following estimate of $\dot{R}_2(\boldsymbol{\gamma}, g)$:

$$\hat{R}^g_2 = \Big(\sum_{i=1}^n \tilde{K}_i \Big)^{-1} \sum_{i=1}^n \sum_{k=1}^{\tilde{K}_i} \hat{\rho}^g_{ik} \ .$$

If $\boldsymbol{\gamma}$ are i.i.d. from some distribution even outside of $\mathcal{G}$ then $|\dot{R}_3(\boldsymbol{\gamma}, g) - \dot{R}_2(\boldsymbol{\gamma}, g)| \to 0$ in P. This asymptotic equivalence holds under model mispecification. Also, in the similar lines to the discussion below (16), it follows that $\hat{R}^g_2$ converges to $\dot{R}_3(\boldsymbol{\gamma}, g)$ at near parametric rate. Thus, $\hat{R}^g_2$ is an efficient estimator of $\dot{R}_2(\boldsymbol{\gamma}, g)$. We formally present this result in the Sec 3.3.

3.2 Calibrated EB prdes

In the previous subsection, we assumed that all S_{ik}s are non-empty. That will never be the case. Consider all (i, k) such that $|S_{i,k}| \geq 1$. Put all those (i, k) in S and all others in $\bar{S}$. For (i, k) in $\bar{S}$, we will not do any shrinkage as it will be difficult to get good risk estimates. In this case, we will just use the predictive density estimator based on the uniform prior. Consider the following prde $\hat{\boldsymbol{p}}^g_c(\dot{\boldsymbol{Y}}; \dot{\boldsymbol{X}})$ that fuses $\hat{p}^g$ and $\hat{p}^{\mathtt{U}}$ based on S:

$$\prod_{i=1}^n \prod_{k=1}^{\tilde{K}_i} \Big(\hat{p}^g[(u_i, \sigma_i), (v_{ik}, \tilde{\sigma}_i)](Y_{ik}; X_i) \Big)^{1\{(i,k) \in S\}} \Big(\hat{p}^{\mathtt{U}}[(u_i, \sigma_i), (v_{ik}, \tilde{\sigma}_i)](Y_{ik}; X_i) \Big)^{1\{(i,k) \in \bar{S}\}} \ .$$

The risk of $\dot{R}(\boldsymbol{\gamma}, \hat{\boldsymbol{p}}^g_c)$ of $\hat{\boldsymbol{p}}^g_c(\dot{\boldsymbol{Y}}; \dot{\boldsymbol{X}})$ directly follows from Proposition 4 with the sum in the two terms $\dot{R}_1$ and $\dot{R}_2$ in right side being replaced by the sum over only (i, k) pairs in S.

We consider the following KL risk estimator (KLRE):

$$\hat{R}_c^g(\dot{\boldsymbol{X}}) = \dot{R}(\boldsymbol{\gamma}, \hat{\boldsymbol{p}}^{\mathtt{U}}) + \left(\sum_{i=1}^{n} \tilde{K}_i\right)^{-1} \sum_{(i,k)\in S} \left\{ \log m^g(v_{ik}\dot{X}_i; u_i, v_{ik}, \sigma_i) - \hat{\rho}_{ik}^g \right\} , \qquad (18)$$

where, $\dot{R}(\boldsymbol{\gamma}, \hat{\boldsymbol{p}}^{\mathtt{U}})$ follows from Proposition 3 and is independent of any parameters.

3.3 Asymptotic Properties

To facilitate a rigorous proof on the asymptotic efficiency of the proposed KLRE we make the following assumptions:

(A2) Consider (a)$0 < \inf_i |u_i| \leq \sup_i |u_i| < \infty$, $0 < \inf_{i,k} v_{ik} \leq \sup_{i,k} v_{i,k} < \infty$ (b) $0 < \inf_i \sigma_i \leq \sup_i \sigma_i < \infty$, $\inf r_i > 0$. These are benign assumptions on the LMMs which impose that we have non-trivial information for each units and are not in a non-degenerate predictive set-up.

(A3) Define, the dependency index: $d_n = \sum_{i=1}^{n} \sum_{k=1}^{\tilde{K}_i} |S_{ik}|^{-1} 1\{|S_{ik}| > 0\}$, which reflects the total usage (by weights) of the $\dot{X}_i$s in $\hat{R}_c^g$. Assume:

$$\lim_{n\to\infty} \left(\sum_{i=1}^{n} \tilde{K}_i\right)^{-1} d_n^2 = 0 \ .$$

The above assumption depends on $\boldsymbol{u}$, $\boldsymbol{v}$, $\boldsymbol{\sigma}$, $\tilde{\boldsymbol{\sigma}}$ and is verifiable. Also, note that if $u_1^2/\sigma_1^2 < u_2^2/\sigma_2^2 \ldots < u_n^2/\sigma_n^2$ then $|S_{ik}| \leq (i-1)$ for all $K = 1, \ldots, \tilde{K}_i$. As such, if additionally $\sup_i v_i^2/r_i \to 0$ then we have $|S_{ik}| = i - 1$. In this case, by assumption A1, $d_n \leq \kappa \sum_{i=2}^{n} |S_{i1}|^{-1} \leq \kappa \log n$. Thus, we expect the assumption to be valid in most LMMs.

(A4) For any fixed $g \in \mathcal{G}$, the fourth moment of g is bounded.

Under these assumptions, we now state our main result on the asymptotic behavior of our proposed KLRE method.

Theorem 1 *Under assumptions A1-A4, for any $g \in \mathcal{G}$ and for any i.i.d. sequence of $\boldsymbol{\gamma}$ with a finite fourth moment of the unit-specific effects, the estimate $\hat{R}_c^g$ of the KL risk of the prde $\hat{\boldsymbol{p}}_c^g$ in (7) satisfies:*

$$c_n^{-1} \left(\dot{R}(\boldsymbol{\gamma}, \hat{\boldsymbol{p}}_c^g) - \hat{R}_c^g\right) = O_p(1) \ \textit{as} \ n \to \infty,$$

where, $c_n = n^{-1/2} (\log \log n) d_n$.

Note that, as $n \to \infty$ if d_n grows at a polynomial rate in $\log n$ (which would be a feature of the design), the KLRE estimator converges to the true risk at a parametric rate barring poly-log terms. Next, consider three popular families of priors:

1. Spike and Slab priors. We consider the spike-and-slab prior class with an atom of mass $1 - \eta$ at the origin and a laplace distribution with scale a. The class $\mathcal{G}$ is indexed by hyper-parameters $\mathcal{H} = \{(\eta, a) : 0 < \eta < 1 \text{ and } a > 0\}$. The members of this class are:

$$g[\eta, \alpha](\tau) = (1 - \eta)\delta_0 + (\eta/2)\alpha \exp(-a|\tau|).$$

This class $\mathcal{G}_1$ has been extensively studied in the Lasso problem [Ročková and George, 2018].

2. Scaled Mixture of Gaussians: Consider $g(\tau) = \sum_{l=1}^{L} \pi_l \phi(\cdot\,; 0, \nu_l)$, where $0 \leq \nu_1 \ldots < \nu_L$ is a pre-specified grid of component variances and $\pi_1, \ldots, \pi_L$ are the unknown mixture proportions. The hyper-parameters in this family $\mathcal{G}_2$ are the $L - 1$ mixture proportion parameters.

3. Discrete Prior: Consider $g(\tau) = \sum_{l=1}^{L} \pi_l \delta_{\tau_l}$ where τ_l are on a uniform fixed grid with the weights $\pi_l \geq 0$ and $\sum_l \pi_l = 1$. The hyper-parameters in this family $\mathcal{G}_3$ are the $L - 1$ weight parameters.

Our proposed prde is $\hat{\boldsymbol{p}}_c[\hat{h}]$ where $\hat{h}_n = \arg\min_{h \in H_n} \hat{R}_c[h]$ where H_n is a dense grid in $\mathcal{H}$ that depends only on the class of estimators and n. Note, we now index the family of prdes by hyper-parameter h instead of g. To understand the risk properties of our proposed prde, we introduce the oracle risk hyper-parameters as those which minimize the true risk function:

$$h_* = \arg\min_{h \in \mathcal{H}} \dot{R}(\boldsymbol{\gamma}, \hat{\boldsymbol{p}}_c^h) \ .$$

The oracle prde $\hat{\boldsymbol{p}}_c[h_*](\dot{\boldsymbol{Y}}; \dot{\boldsymbol{X}})$ is not really an estimator since h_* depends on the unknown parameter values. The oracle is not obtainable in practice but provide the theoretical benchmark that one can ever hope to reach. Indeed, no prdes in $\mathcal{G}$ can have smaller risk than the oracle risk prde.

Using Theorem 1, we show the risk of our proposed prde for the above mentioned three classes converges to the risk of the oracle prde.

Theorem 2 *Under assumptions A1-A3, prdes in the classes $\mathcal{G}_1$, $\mathcal{G}_2$ and $\mathcal{G}_3$ that are calibrated by RKLE satisfy the following:*

$$\dot{c}_n^{-1}\big(\dot{R}(\boldsymbol{\gamma}, \hat{\boldsymbol{p}}_c[\hat{h}_n]) - \dot{R}(\boldsymbol{\gamma}, \hat{\boldsymbol{p}}_c[h_*])\big) = O_p(1) \text{ as } n \to \infty,$$

where, $\dot{c}_n = c_n(\log n)^L$.

3.4 EB prdes tuned by KLRE

In the previous subsection, we prescribe choosing the hyper-parameter $\hat{h}$ based on the KLRE of the conditional Bayes estimate when $\boldsymbol{\beta}$ is fixed. As $\boldsymbol{\beta}$ is a low-dimensional parameter, following the EB perspectives in Jiang and Zhang [2010], we can extend our approach to simultaneously estimate $(\hat{\boldsymbol{\beta}}, \hat{h})$ using KLRE. For that purpose, consider rewriting (18) as

$$\hat{R}_c^h(\boldsymbol{X}, \boldsymbol{\beta}) = \dot{R}(\boldsymbol{\gamma}, \hat{\boldsymbol{p}}^{\mathtt{U}}) + \left(\sum_{i=1}^n \tilde{K}_i\right)^{-1} \sum_{(i,k)\in S} \left\{ \log m^h(v_{ik}(f(X_i, \boldsymbol{\beta}); u_i, v_{ik}, \sigma_i) - \hat{\rho}_{ik}^h(\boldsymbol{\beta}) \right\},$$

where, we replace $\{\dot{X}_i : 1 \le i \le n\}$ in m^h and $\hat{p}_{ik}^h$ in (18) by $f(X_i, \boldsymbol{\beta})$ where the functional form f is provided in the display above (6). We simultaneously estimate $(\hat{\boldsymbol{\beta}}, \hat{h})$ as $\arg\min_{h\in\mathcal{H}, \boldsymbol{\beta}} \hat{R}_c^h(\boldsymbol{X}, \boldsymbol{\beta})$ and prescribe using the prde $\hat{\boldsymbol{p}}^{(\hat{h}, \hat{\boldsymbol{\beta}})}(\boldsymbol{Y}; \boldsymbol{X})$ based on the form of the prde given in proposition 1.

If the variances are not known and $\sigma_i = \tilde{\sigma}_i = \sigma$ for $i = 1, \ldots, n$, where σ is unknown, we plug-in estimates of σ in prdes. Applying our proposed methodology on prdes based on a consistent point estimate of σ will lead to good predictive performance. In this case, as we have large n and replicates for several cases, obtaining an consistent estimator for σ is not difficult. Consider the set $\mathcal{I} = \{i : 1 \le i \le n \text{ and } K_i \ge 2\}$. Define the contrasts:

$$f(i, \boldsymbol{\beta}) = \{u_{i2}(X_{i1} - \boldsymbol{a}_{i1}'\boldsymbol{\beta}) - u_{i1}(X_{i2} - \boldsymbol{a}_{i2}'\boldsymbol{\beta})\}(u_{i1}^2 + u_{i2}^2)^{-1/2} \quad \text{for } i \in \mathcal{I}.$$

Then, the estimate $\hat{\sigma}^2 = |\mathcal{I}|^{-1} \sum_{i\in\mathcal{I}} f^2(i, \hat{\boldsymbol{\beta}})$, where $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i\in\mathcal{I}} f^2(i, \boldsymbol{\beta})$, is a consistent estimator for σ as $|\mathcal{I}| \to \infty$.

4 Discussion and Future Work

We consider predictive density estimation under Kullback-Leibler loss in linear mixed models. Specifically, we adopt an empirical Bayes approach in which prdes are optimally shrunken using a hyper-parameter selection method based on risk estimation. In forthcoming work, we provide a detailed study, including verifiable conditions on the covariates under which Assumption A3—governing the rates in Theorem 1—holds. We also include extensive simulation studies illustrating the finite-sample performance of the proposed method across a variety of covariate distributions.

It is worth noting that the exchangeability of individual effects is crucial for the shrinkage theory developed in this paper. Extending this theory to non-exchangeable settings would be an interesting direction for future research. Ghosh and Kubokawa [2018], L'Moudden and Marchand [2019], Maruyama et al. [2019], Matsuda and Strawderman [2021], Suzuki and Komaki [2010] examined prdes under general divergence loss, albeit within sequence models without covariates. In future work, it would be valuable to develop the proposed method for broader divergence losses.

In the absence of covariates, Kubokawa et al. [2017], Li and Gal [2017], Sadeghkhani [2022], Sadeghkhani and Ahmed [2020] explored shrunken prdes beyond normal models. In future, it will be useful to extend the proposed approach with covariates to generalized linear mixed effects models.

Acknowlegdement

We thank the editors and two referees for their constructive comments that have improved the quality of this paper. This work was supported by Japan Society for the Promotion of Science (JSPS) [Grant Nos. 21H05205, 21K12067, 23K11024, 24H00247], and the MEXT Project for Seismology toward Research Innovation with Data of Earthquake (STAR-E) [Grant No. JPJ010217].

5 Appendix

Proof of (9)**.** As g is a uniform prior on γ, it is also uniform on $v\gamma$. Writing $v\gamma$ as $\ddot{\gamma}$ we have, the Bayes prde based on the uniform prior as:

$$\int \phi(\ddot{y}\,|\ddot{\gamma}, r^{1/2}\sigma)\,\phi(\ddot{x}\,|\ddot{\gamma}, u^{-1}v\sigma)\{m(\ddot{x}; u, v)\}^{-1}\,d\ddot{\gamma}$$

where, $m(\ddot{x}; u, v) = \int \phi(\ddot{x}|\ddot{\gamma}, u^{-1}v\sigma)\,d\ddot{\gamma} = 1$. Define, $\mu = v\gamma$, $s = v/u$ and $\kappa = (r^{-1}+s^{-2})^{1/2}$. Next, note that $\phi(\ddot{y}\,|\ddot{\gamma}, \tilde{\sigma})\,\phi(\ddot{x}\,|\ddot{\gamma}, u^{-1}v\sigma)$ is proportional to

$$\exp\{-(\ddot{y}-\mu)^2/(2r\sigma^2) - (\ddot{y}-\mu)^2/(2\sigma^2 s^2)\}.$$

This decomposes as:

$$\exp\left\{-\frac{1}{2\sigma^2}\left(\mu\kappa - \frac{1}{\kappa}\left(\frac{\ddot{y}}{r}+\frac{\ddot{x}}{s^2}\right)\right)\cdot\exp\left(-\frac{1}{2\sigma^2}\left(\frac{\ddot{y}^2}{r}+\frac{\ddot{x}^2}{s^2}-\frac{1}{\kappa^2}\left(\frac{\ddot{y}}{r}+\frac{\ddot{x}}{s^2}\right)^2\right)\right.\right.$$

The first term is a normal density function with respect to μ and is integrable. The constants are independent of $\ddot{x}$ and $\ddot{y}$. For the second term it can be further simplified to:

$$\exp\left(-\frac{1}{2\sigma^2}\cdot\frac{(\ddot{y}-\ddot{x})^2}{r+s^2}\right),$$

and the result follows.

Proof of (10)**.** By definition, the risk of $\hat{p}^{\mathtt{U}}$ is given by

$$2^{-1}\log\left(\frac{2(r+v^2u^{-2})\sigma^2}{r\sigma^2}\right) + \left\{\frac{\mathbb{E}(\ddot{Y}-\ddot{X})^2}{2\sigma^2(r+v^2u^{-2})} - \frac{1}{2}\right\}.$$

Decomposing $\mathbb{E}(\ddot{Y}-\ddot{X})^2 = \mathbb{E}(\ddot{Y}-v\gamma)^2 + \mathbb{E}(\ddot{X}-v\gamma)^2 = \sigma^2 r + \sigma^2 u^{-2}v^2$, the result follows.

Proof of proposition 1. First note that conditioned on $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$, the true density of $\dot{\boldsymbol{Y}}$ is a product density across (i,k) pairs and the true density of $\dot{\boldsymbol{X}}$ is also a product density across $i = 1, \ldots, n$. As we consider product priors on $\boldsymbol{\tau}$, for any fixed $\boldsymbol{\beta}$, the corresponding Bayes predictive density estimates are co-ordinatewise Bayes rule. The proof then follows directly from the expression of the product of co-ordinatewise Bayes prdes that is presented in the display just before the proposition. We consider a change of variable and consider the density of $v_{ik}\dot{X}_i$ instead of $\dot{X}_i$ for the (i,k) coordinate. The scale transformation does

not add any terms in the density as the terms in the Jacobian appear both in the numerator and denominator and so, get cancelled off.

Proof of proposition 2. The proof follows from (16) in Lemma 2 of George et al. [2006].

Proof of proposition 3. By proposition 1, as the Bayes prde from the uniform prior is a product rule (conditioned on β) its risk decomposes into sum of univariate risks:

$$\dot{R}(\boldsymbol{\gamma}, \hat{\boldsymbol{p}}^{\mathtt{U}}) = \Big(\sum_{i=1}^{n} \tilde{K}_i \Big)^{-1} \sum_{i=1}^{n} \sum_{k=1}^{\tilde{K}_i} r(\gamma_i, \hat{p}^{\mathtt{U}}; u_i, v_{ik}),$$

and then, the result follows by applying (10).

Proof of proposition 4. By proposition 1, conditional on $\boldsymbol{\beta}$ the Bayes prde $\hat{\boldsymbol{p}}^{g}$ is a product rule and so, its risk decouples into sum of univariate predictive risks which are derived based on the proposition 2.

Proof of (17) Direct evaluations of Gaussian convolutions yield

$$m^{g}(v_{ik}\dot{X}_i; u_i, v_{ik}, \sigma_i) = (2\pi(v_{ik}^2\tau + u_i^{-2}v_{ik}^2\sigma_i^2))^{-1/2} \exp\{-(v_{ik}\dot{X}_i)^2(2(v_{ik}^2\tau + u_i^{-2}v_{ik}^2\sigma_i^2))^{-1}\} \text{ and,}$$

$$q^{g}(\dot{W}_{ik}; u_i, v_{ik}, \sigma_i, r_i) = \left(2\pi \left\{ v_{ik}^2\tau + \frac{\sigma_i^2}{(r_i^{-1} + u_i^2/v_{ik}^2)} \right\} \right)^{-1/2} \exp\left\{ - \frac{(\dot{W}_{ik})^2}{2\{v_{ik}^2\tau + \sigma_i^2/(r_i^{-1} + u_i^2/v_{ik}^2)\}} \right\}.$$

Thus, we have $2\,\mathbb{E}[\log m^{g}(v_{ik}\dot{X}_i; u_i, v_{ik}, \sigma_i) - \log q^{g}(\dot{W}_{ik}; u_i, v_{ik}, \sigma_i, r_i)]$ equals

$$\begin{aligned}
&- \log \frac{v_{ik}^2\tau + u_i^{-2}v_{ik}^2\sigma_i^2}{v_{ik}^2\tau + \{\sigma_i^2/(r_i^{-1} + u_i^2/v_{ik}^2)\}} - \frac{\mathbb{E}[(v_{ik}\dot{X}_i)^2]}{v_{ik}^2\tau + u_i^{-2}v_{ik}^2\sigma_i^2} + \frac{\mathbb{E}[(\dot{W}_{ik})^2]}{v_{ik}^2\tau + \{\sigma_i^2/(r_i^{-1} + u_i^2/v_{ik}^2)\}} \\
&= - \log \frac{v_{ik}^2\tau + u_i^{-2}v_{ik}^2\sigma_i^2}{v_{ik}^2\tau + \{\sigma_i^2/(r_i^{-1} + u_i^2/v_{ik}^2)\}} - \frac{\gamma_i^2 v_{ik}^2 + u_i^{-2}v_{ik}^2\sigma_i^2}{v_{ik}^2\tau + u_i^{-2}v_{ik}^2\sigma_i^2} + \frac{\gamma_i^2 v_{ik}^2 + \sigma_i^2/(r_i^{-1} + u_i^2/v_{ik}^2)}{v_{ik}^2\tau + \sigma_i^2/(r_i^{-1} + u_i^2/v_{ik}^2)} \\
&= - \log \frac{\tau + u_i^{-2}\sigma_i^2}{\tau + \{\sigma_i^2/(v_{ik}^2 r_i^{-1} + u_i^2)\}} - \frac{\gamma_i^2 + u_i^{-2}\sigma_i^2}{\tau + u_i^2\sigma_i^2} + \frac{\gamma_i^2 + \sigma_i^2/(v_{ik}^2 r_i^{-1} + u_i^2)}{\tau + \sigma_i^2/(v_{ik}^2 r_i^{-1} + u_i^2)} \\
&= - \log \frac{(\tau + \sigma_i^2 u_i^{-2})(r_i^{-1}v_{ik}^2 + u_i^2)}{\tau(r_i^{-1}v_{ik}^2 + u_i^2) + \sigma_i^2} - \frac{r_i^{-1}v_{ik}^2\sigma_i^2(\gamma_i^2 - \tau)}{(\tau u_i^2 + \sigma_i^2)\{\tau(r_i^{-1}v_{ik}^2 + u_i^2) + \sigma_i^2\}}.
\end{aligned}$$

Thus, the difference between $\dot{R}_1(\boldsymbol{\gamma}, g)$ and $\dot{R}_2(\boldsymbol{\gamma}, g)$ is

$$-\Big(\sum_{i=1}^{n} 2\tilde{K}_i \Big)^{-1} \sum_{i=1}^{n} \sum_{k=1}^{\tilde{K}_i} \Big\{ \log \frac{(\tau + \sigma_i^2 u_i^{-2})(r_i^{-1}v_{ik}^2 + u_i^2)}{\tau(r_i^{-1}v_{ik}^2 + u_i^2) + \sigma_i^2} + \frac{r_i^{-1}v_{ik}^2\sigma_i^2(\gamma_i^2 - \tau)}{(\tau u_i^2 + \sigma_i^2)\{\tau(r_i^{-1}v_{ik}^2 + u_i^2) + \sigma_i^2\}} \Big\},$$

which completes the proof.

Proof of Theorem 1. The proof of the theorem involves the following three steps. For any fixed $\boldsymbol{\gamma}$ we need to prove:

Step 1. $\dot{R}_1(\boldsymbol{\gamma}, g) - \hat{R}_1^g = O_p((\log\log n)/\sqrt{n})$.

Step 2. $\dot{R}_3(\boldsymbol{\gamma}, g) - \hat{R}_2^g = O_p((\log\log n)(d_n^2/\sqrt{n})$.

Step 3. $\dot{R}_2(\boldsymbol{\gamma}, g) - \dot{R}_3(\boldsymbol{\gamma}, g) = O_p((\log\log n)d_n^2/\sqrt{n})$.

Proof of Step 1. We begin with bounding

$$M_i^2 := \mathrm{Var}\left[\left(\sum_{i=1}\tilde{K}_i\right)^{-1}\sum_{k=1}^{\tilde{K}_i}\{\log m^g(v_{ik}\dot{X}_i; u_i, v_{ik}, \sigma_i)\}\right].$$

By applying Jensen's inequality to the function $y \to y^2$, we have

$$\begin{aligned} M_i^2 &\leq \left(\sum_{i=1}^n \tilde{K}_i\right)^{-2}\mathbb{E}\left[\left\{\sum_{k=1}^{\tilde{K}_i}\log m^g(v_{ik}\dot{X}_i; u_i, v_{ik}, \sigma_i)\right\}^2\right] \\ &\leq \left(\sum_{i=1}^n \tilde{K}_i\right)^{-2}\tilde{K}_i\sum_{k=1}^{\tilde{K}_i}\mathbb{E}\left[\left(\log m^g(v_{ik}\dot{X}_i; u_i, v_{ik}, \sigma_i)\right)^2\right]. \end{aligned}$$

Then, consider bounding $\mathbb{E}[(\log m^g(v_{ik}\dot{X}_i; u_i, v_{ik}, \sigma_i))^2$. Observe

$$m^g(v_{ik}\dot{X}_i; u_i, v_{ik}, \sigma_i) = \int \phi(v_{ik}\dot{X}_i; u_i\gamma, u_i^{-1}v_{ik}\sigma_i)g(\gamma)d\gamma \leq 1/\{\sqrt{2\pi}(u_i^{-1}v_{ik}\sigma_i)\},$$

and so, we can upper bound $\mathbb{E}[(\log m^g(v_{ik}\dot{X}_i; u_i, v_{ik}, \sigma_i))^2$ by

$$\begin{aligned} &\mathbb{E}\left[\left(\log\frac{m^g(v_{ik}\dot{X}_i; u_i, v_{ik}, \sigma_i)}{(\sqrt{2\pi}(u_i^{-1}v_{ik}\sigma_i))^{-1}} + \log\frac{1}{\sqrt{2\pi}(u_i^{-1}v_{ik}\sigma_i)}\right)^2\right] \\ &\leq 2\mathbb{E}\left[\left(\log\frac{m^g(v_{ik}\dot{X}_i; u_i, v_{ik}, \sigma_i)}{(\sqrt{2\pi}(u_i^{-1}v_{ik}\sigma_i))^{-1}}\right)^2\right] + 2(\log\{\sqrt{2\pi}(u_i^{-1}v_{ik}\sigma_i)\})^2. \end{aligned}$$

By Jensen's inequality the above is again bounded by

$$2\mathbb{E}\left\{\mathbb{E}_{\gamma\sim g}\left(\log\frac{\phi(v_{ik}\dot{X}_i; u_i\gamma, u_i^{-1}v_{ik}\sigma_i)}{(\sqrt{2\pi}(u_i^{-1}v_{ik}\sigma_i)^{-1})}\right)^2\right\} + 2(\log\{\sqrt{2\pi}(u_i^{-1}v_{ik}\sigma_i)\})^2.$$

Next, by using the inequalities $(a+b)^2 \le 2a^2+2b^2$ and $(a+b)^4 \le 53(a^4+b^4)$, we finally bound $\mathbb{E}[(\log m^g(v_{ik}\dot{X}_i; u_i, v_{ik}, \sigma_i))^2$ by

$$
\begin{aligned}
&4\mathbb{E}\left\{\mathbb{E}_{\gamma\sim g}\left[\frac{(v_{ik}\dot{X}_i - u_i\gamma)^2}{2(u_i^{-1}v_{ik}\sigma_i)^2}\right]^2\right\} + 4(\log\{\sqrt{2\pi}(u_i^{-1}v_{ik}\sigma_i)\})^2 \\
&\le c\left(\frac{u_i^4}{\sigma_i^4}\mathbb{E}_{\dot{X}_i}\left[\dot{X}_i^4\right] + \frac{u_i^4}{v_{ik}^4\sigma_i^4}\mathbb{E}_{\gamma\sim g}\left[\gamma^4\right]\right) + (\log 2\pi)^2 + 4\{\log(u_i^{-1}v_{ik}\sigma_i)\}^2,
\end{aligned}
$$

where, c is an absolute constants (specifically, say, 500). Substituting $\mathbb{E}[\dot{X}_i^4] = \mathbb{E}_{\gamma\sim g^*}\mathbb{E}_{\dot{X}_i|\gamma}[\dot{X}_i^4]$ where the true distribution of γ is denoted by g^*, we further simplify the upper bound on $\mathbb{E}[(\log m^g(v_{ik}\dot{X}_i; u_i, v_{ik}, \sigma_i))^2$ as

$$
c'\left(\frac{u_i^4}{\sigma_i^4}\mathbb{E}_{\gamma\sim g^*}[\gamma^4] + 3 + \frac{u_i^4}{v_{ik}^4\sigma_i^4}\mathbb{E}_{\gamma\sim g}\left[\gamma^4\right]\right) + (\log 2\pi)^2 + 4\{\log(u_i^{-1}v_{ik}\sigma_i)\}^2,
$$

where c' is an absolute constant. Thus, we get

$$
\begin{aligned}
M_i^2 &\le \left(\sum_{i=1}^n \tilde{K}_i\right)^{-2}\left(\tilde{K}_i^2 c'' + \tilde{K}_i^2\frac{u_i^4}{\sigma_i^4}\mathbb{E}_{\gamma\sim g^*}[\gamma^4] + \tilde{K}_i\frac{u_i^4}{\sigma_i^4}\mathbb{E}_{\gamma\sim g}[\gamma^4]\sum_{k=1}^{\tilde{K}_i} v_{ik}^{-4}\right) \\
&= \left(\sum_{i=1}^n \tilde{K}_i\right)^{-2}\tilde{K}_i^2\left(c'' + \max_i(u_i^{-1}\sigma_i)^{-4}\mathbb{E}_{\gamma\sim g^*}[\gamma^4] + \max_{i,k}(u_i^{-1}v_{ik}\sigma_i)^{-4}\mathbb{E}_{\gamma\sim g}[\gamma^4]\right)
\end{aligned}
$$

where c'' is an absolute constant plus $4\max_{i,k}(\log u_i^{-1}v_{ik}\sigma_i)^2$. So, we finally get

$$
\sum_{i=1}^n M_i^2 \le n^{-1}c''' \quad \text{with a constant } c''' \text{ independent from } n,
$$

which completes Step 1 by the Chebyshev inequality.

Proof of Step 2. Hereafter, without any loss of generality assume $\tilde{K}_i = 1$, $\sigma_i = 1$, $r_i = r$ for all $i = 1, \ldots, n$. We denote S_{ik} as S_i as $\tilde{K}_i = 1$. Also, without loss of generality we can assume: $0 < u_1^2 < u_2^2 < \ldots < u_n^2$. This implies:

(a) $|S_i| \le i-1$ for $i = 1, \ldots, n$;

(b) $|S_i|$ is monotonically non-decreasing in i;

(c) if $j \in S_i$ then $k \in S_i$ for all $k < j$.

Thus, we will using $\hat{p}^{\mathtt{U}}$ for $i = 1$ and for all other i such that $|S_i| = 0$. We use $\hat{p}^g$ for

$i \geq 2$. For ease of presentation, we skip using the suffix c for the prdes as well as their risks henceforth. Also, we do not explicitly denote the dependence of r in the risk calculations.

Define $s_i = |S_i|$. Note that:

$$\dot{R}_3(\boldsymbol{\gamma}, g) = n^{-1} \sum_{i=2}^{n} \rho_i(\boldsymbol{\gamma}, g) = n^{-1} \sum_{i=2}^{n} s_i^{-1} \sum_{j=1}^{i-1} \mathbb{E}_{\gamma_j} \big\{ \log q^g(\dot{T}_{ij}; u_i, v_i) \big\} \cdot 1\{j \in S_i\},$$
$$= n^{-1} \sum_{j=1}^{n-1} \sum_{i=j+1}^{n} s_i^{-1} \mathbb{E}_{\gamma_j} \big\{ \log q^g(\dot{T}_{ij}; u_i, v_i) \big\} \cdot 1\{j \in S_i\}.$$

Similarly, we have:

$$\hat{R}_2^g = n^{-1} \sum_{j=1}^{n-1} \sum_{i=j+1}^{n} s_i^{-1} \mathbb{E}_Z \big\{ \log q^g(\dot{T}_{ij}(Z); u_i, v_i) \big\} \cdot 1\{j \in S_i\} \tag{19}$$

It is easy to see that $\hat{R}_2^g$ is an unbiased estimate of $\dot{R}_3(\boldsymbol{\gamma}, g)$. We next want to calculate its variance. For that purpose, we represent the terms in the right side of (19) as functions of $\dot{X}_j$ for $j = 1, \ldots, n$ and rewrite $\hat{R}_2^g$ as:

$$n -^{-1} \sum_{j=1}^{n-1} \mathcal{A}_j \text{ where, } \mathcal{A}_j = \sum_{i=j+1}^{n} s_i^{-1} f(v_i \dot{X}_j; u_i, v_i) \cdot 1\{j \in S_i\}.$$

Note that, $\mathcal{A}_j$ are independent of each other and so,

$$\mathrm{Var}(\hat{R}_2^g) = n^{-2} \sum_{j=1}^{n-1} \mathrm{Var}(\mathcal{A}_j).$$

Next, we bound $\mathrm{Var}(\mathcal{A}_j) \leq \{\sum_{i=j+1}^{n} s_i^{-1} \cdot 1\{j \in S_i\}\}^2 * \{\max_{i \geq j+1} \mathrm{Var}(f(v_i \dot{X}_j; u_i, v_i))\}$. The first term is bounded by $n d_n^2$. The second term is bounded via the similar calculus as in Step 1 and so, $\mathrm{Var}(\hat{R}_2^g) = O_p(d_n^2/n)$.

Proof of Step 3. By definition,

$$\dot{R}_2(\boldsymbol{\gamma}, g) = \frac{1}{n} \sum_{i=2}^{n} f(\gamma_i, u_i, v_i),$$

where,

$$f(\gamma, u, v) = \int \left\{ \log \int \phi\left(\frac{\gamma - a}{\sigma(v^2 r^{-1} + u^2)^{-1/2}} + z \right) g(a)\, da \right\} \phi(z)\, dz.$$

We assume that $\boldsymbol{\gamma}$ are i.i.d. from $\tilde{g}$ (this is different from g), $\boldsymbol{u}, \boldsymbol{v}$ are i.i.d. from p_1 and p_2 respectively. Also, $\boldsymbol{\gamma}$, $\boldsymbol{u}$, $\boldsymbol{v}$ are independent among themselves.

Next, consider:

$$\dot{R}_3(\boldsymbol{\gamma}, g) = \frac{1}{n-1} \sum_{i=2}^{n} \sum_{j=1}^{i-1} s_i^{-1} f(\gamma_j, u_i, v_i) 1\{j \in S_i\}.$$

As $\boldsymbol{\gamma}$, $\boldsymbol{u}$ and $\boldsymbol{v}$ are independent, by Tower property of conditional expectation, we have:

$$\mathbb{E}\{\dot{R}_3(\boldsymbol{\gamma}, g)\} = \mathbb{E}\{\dot{R}_2(\boldsymbol{\gamma}, g)\} = \int f(\gamma, u, v) \tilde{g}(\gamma) p_1(u) p_2(v) d\gamma du dv.$$

Next, note that $\mathrm{Var}(\dot{R}_2) = n^{-1} \mathrm{Var}(f(\gamma, u, v))$. Also,

$$\mathrm{Var}(\dot{R}_3) = \mathrm{Var}(E(\dot{R}_3 | \boldsymbol{u}, \boldsymbol{v})) + \mathbb{E}(\mathrm{Var}(\dot{R}_3 | \boldsymbol{u}, \boldsymbol{v})).$$

The first term in the right side above is $\mathrm{Var}(n^{-1} \sum_{i=2}^{n} \mathbb{E}\{f(\gamma | u_i, v_i)\})$ which is $O_p(n^{-1})$. The second term is $O_p(n^{-1} d_n^2)$ in the similar lines to the proof of step 2. This completes the proof.

Proof of Theorem 2. Noting that all priors in the above classes satisfy Assumption A4 of Theorem 1, the proof of the asymptotic control on the risk of the prescribed PRDEs follows similarly to the proof of Theorem 3.3 in Xie et al. [2012].

References

J. Aitchison and I. R. Dunsmore. *Statistical prediction analysis*. Cambridge University Press, 1975.

M. Aslan. Asymptotically minimax Bayes predictive densities. *Ann. Statist.*, 34(6):2921–2938, 2006. ISSN 0090-5364.

A. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, 44(6):2743–2760, 1998.

P. Bhagwat and É. Marchand. Bayesian inference and prediction for mean-mixtures of normal distributions. *Electronic Journal of Statistics*, 17(2):1893–1922, 2023.

L. D. Brown, E. I. George, and X. Xu. Admissible predictive density estimation. *Ann. Statist.*, 36(3):1156–1170, 2008. ISSN 0090-5364.

T. M. Cover. *Elements of information theory*. John Wiley & Sons, 1999.

E. Demidenko. *Mixed models: theory and applications with R*. John Wiley & Sons, 2013.

B. Efron. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press, 2012.

J. J. Faraway. *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. CRC press, 2016.

D. Fourdrinier, É. Marchand, A. Righi, and W. E. Strawderman. On improved predictive density estimation with parametric constraints. *Electronic Journal of Statistics*, 5:172–191, 2011. ISSN 1935-7524.

D. Fourdrinier, W. E. Strawderman, and M. T. Wells. *Shrinkage estimation*. Springer, 2018.

U. Gangopadhyay and G. Mukherjee. On discrete priors and sparse minimax optimal predictive densities. *Electronic Journal of Statistics*, 2021.

S. Geisser. *Predictive inference*, volume 55 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, New York, 1993. ISBN 0-412-03471-9.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

A. Gelman, J. Hwang, and A. Vehtari. Understanding predictive information criteria for Bayesian models. *Statistics and computing*, 24(6):997–1016, 2014.

E. George, E. Marchand, G. Mukherjee, and D. Paul. New and evolving roles of shrinkage in large-scale prediction and inference. In *BIRS Workshop Report*, pages 1–14, 2019.

E. George, G. Mukherjee, and K. Yano. Optimal shrinkage estimation of predictive densities under α-divergences. *Bayesian Analysis*, 16(4):1139–1155, 2021.

E. I. George, F. Liang, and X. Xu. Improved minimax predictive densities under Kullback–Leibler loss. *The Annals of Statistics*, pages 78–91, 2006.

E. I. George, F. Liang, and X. Xu. From minimax shrinkage estimation to minimax shrinkage prediction. *Statistical Science*, 27(1):82–94, 2012. ISSN 0883-4237.

M. Ghosh and T. Kubokawa. Hierarchical Bayes versus empirical Bayes density predictors under general divergence loss. *Biometrika*, 2018.

M. Ghosh, V. Mergel, and G. S. Datta. Estimation, prediction and the Stein phenomenon under divergence loss. *Journal of Multivariate Analysis*, 99(9):1941–1961, 2008. ISSN 0047-259X.

M. Ghosh, V. Mergel, and R. Liu. A general divergence criterion for prior selection. *Annals of the Institute of Statistical Mathematics*, 63(1):43–58, 2011.

M. Ghosh, T. Kubokawa, and G. S. Datta. Density prediction and the Stein phenomenon. *Sankhya A*, pages 1–23, 2019.

J. A. Hartigan. The maximum likelihood prior. *The Annals of Statistics*, 26(6):2083–2103, 1998. ISSN 0090-5364.

J. Jiang and T. Nguyen. *Linear and generalized linear mixed models and their applications*, volume 1. Springer, 2007.

W. Jiang and C.-H. Zhang. Empirical bayes in-season prediction of baseball batting averages. In *Borrowing Strength: Theory Powering Applications–A Festschrift for Lawrence D. Brown*, volume 6, pages 263–274. Institute of Mathematical Statistics, 2010.

F. Komaki. A shrinkage predictive distribution for multivariate normal observables. *Biometrika*, 88(3):859–864, 2001. ISSN 0006-3444.

F. Komaki. Simultaneous prediction of independent Poisson observables. *The Annals of Statistics*, 32(4):1744–1769, 2004. ISSN 0090-5364.

T. Kubokawa, É. Marchand, W. E. Strawderman, and J.-P. Turcotte. Minimaxity in predictive density estimation with parametric constraints. *Journal of Multivariate Analysis*, 116:382–397, 2013.

T. Kubokawa, É. Marchand, and W. E. Strawderman. On predictive density estimation for location families under integrated absolute error loss. *Bernoulli*, 23(4B):3197–3212, 2017.

E. L. Lehmann and G. Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

Y. Li and Y. Gal. Dropout inference in Bayesian neural networks with alpha-divergences. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2052–2061. JMLR. org, 2017.

F. Liang. *Exact minimax procedures for predictive density estimation and data compression*. ProQuest LLC, Ann Arbor, MI, 2002. ISBN 978-0493-60397-1. Thesis (Ph.D.)–Yale University.

F. Liang and A. Barron. Exact minimax strategies for predictive density estimation, data compression, and model selection. *IEEE Transactions on Information Theory*, 50(11): 2708–2726, 2004. ISSN 0018-9448.

F. Liang and A. Barron. *Exact Minimax Predictive Density Estimation and MDL*, chapter 7, pages 177–194. Advances in Minimum Description Length: Theory and Applications (P. Grunwald, I. Myung and M. Pitt eds). MIT Press, 2005.

A. L'Moudden and É. Marchand. On predictive density estimation under α-divergence loss. *Mathematical Methods of Statistics*, 28:127–143, 2019.

É. Marchand and A. Sadeghkhani. On predictive density estimation with additional information. *Electronic Journal of Statistics*, 12:4209–4238, 2018.

Y. Maruyama and W. E. Strawderman. Bayesian predictive densities for linear regression models under α-divergence loss: Some results and open problems. In *Contemporary developments in Bayesian analysis and statistical decision theory: a Festschrift for William E. Strawderman*, volume 8, pages 42–57. Institute of Mathematical Statistics, 2012.

Y. Maruyama, T. Matsuda, and T. Ohnishi. Harmonic Bayesian prediction under α-divergence. *IEEE Transactions on Information Theory*, 2019.

T. Matsuda and F. Komaki. Singular value shrinkage priors for bayesian prediction. *Biometrika*, 102:843–854, 2015.

T. Matsuda and W. Strawderman. Predictive density estimation under the Wasserstein loss. *Journal of Statistical Planning and Inference*, 210:53–63, 2021.

T. Matsuda and W. Strawderman. Estimation under matrix quadratic loss and matrix superharmonicity. *Biometrika*, 109:503–519, 2022.

C. E. McCulloch and S. R. Searle. *Generalized, linear, and mixed models*. John Wiley & Sons, 2004.

G. Mukherjee. *Sparsity and Shrinkage in Predictive Density Estimation*. PhD thesis, Stanford University, 2013.

G. Mukherjee and I. M. Johnstone. Exact minimax estimation of the predictive density in sparse Gaussian models. *The Annals of Statistics*, 43:937–961, 2015.

G. Mukherjee and I. M. Johnstone. On minimax optimality of sparse bayes predictive density estimates. *Annals of Statistics*, 2022.

I. Pardoe. A bayesian sampling approach to regression model checking. *Journal of Computational and Graphical Statistics*, 10(4):617–627, 2001.

J. C. Pinheiro and D. M. Bates. Linear mixed-effects models: basic concepts and examples. *Mixed-effects models in S and S-Plus*, pages 3–56, 2000.

V. Ročková and E. I. George. The spike-and-slab lasso. *Journal of the American Statistical Association*, 113(521):431–444, 2018.

A. Sadeghkhani. On improving the posterior predictive distribution of the difference between two independent poisson distribution. *Sankhya B*, 84(2):765–777, 2022.

A. Sadeghkhani and S. E. Ahmed. On predictive distribution of k-inflated poisson models with and without additional information. *Revista Colombiana de Estadística*, 43(2): 173–182, 2020.

G. Seymour. The inferential use of predictive distributions. *Foundations of Statistical Inference*, pages 456–469, 1971.

C. M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, pages 1135–1151, 1981.

T. Suzuki and F. Komaki. On prior selection and covariate shift of β-Bayesian prediction under α-divergence risk. *Communications in Statistics: Theory and Methods*, 39:1655–1673, 2010.

A. S. Tay and K. F. Wallis. Density forecasting: a survey. *Journal of forecasting*, 19(4): 235–254, 2000.

J. W. Taylor and R. Buizza. A comparison of temperature density forecasts from GARCH and atmospheric models. *Journal of Forecasting*, 23(5):337–355, 2004.

G. Verbeke and G. Molenberghs. *Linear mixed models in practice: a SAS-oriented approach*, volume 126. Springer Science & Business Media, 2012.

X. Xie, S. Kou, and L. D. Brown. Sure estimates for a heteroscedastic hierarchical model. *Journal of the American Statistical Association*, 107(500):1465–1479, 2012.

X. Xu. *Estimation of high dimensional predictive densities*. PhD thesis, University of Pennsylvania, 2005.

X. Xu and F. Liang. Asymptotic minimax risk of predictive density estimation for non-parametric regression. *Bernoulli*, 16(2):543–560, 2010. ISSN 1350-7265.

X. Xu and D. Zhou. Empirical Bayes predictive densities for high-dimensional normal models. *Journal of Multivariate Analysis*, 102(10):1417–1428, 2011.

K. Yano and F. Komaki. Information criteria for prediction when the distributions of current and future observations differ. *Statistica Sinica*, 27:1205–1223, 2017a.

K. Yano and F. Komaki. Asymptotically minimax prediction in infinite sequence models. *Electronic Journal of Statistics*, 11(2):3165–3195, 2017b.

K. Yano, R. Kaneko, and F. Komaki. Minimax predictive density for sparse count data. *Bernoulli*, 27:1212–1238, 2021.

A. Yuan and B. Clarke. An information criterion for likelihood selection. *IEEE Transactions on Information Theory*, 45(2):562–571, 1999.