Bayesian Analysis (2021) **16**, Number 4, pp. 1139–1155

Optimal Shrinkage Estimation of Predictive Densities Under α-Divergences*

Edward George†, Gourab Mukherjee‡, and Keisuke Yano§

Abstract. We consider the problem of estimating the predictive density in a heteroskedastic Gaussian model under general divergence loss. Based on a conjugate hierarchical set-up, we consider generic classes of shrinkage predictive densities that are governed by location and scale hyper-parameters. For any α-divergence loss, we propose a risk-estimation based methodology for tuning these shrinkage hyper-parameters. Our proposed predictive density estimators enjoy optimal asymptotic risk properties that are in concordance with the optimal shrinkage calibration point estimation results established by Xie, Kou, and Brown (2012) for heteroskedastic hierarchical models. These α-divergence risk optimality properties of our proposed predictors are not shared by empirical Bayes predictive density estimators that are calibrated by traditional methods such as maximum likelihood and method of moments. We conduct several numerical studies to compare the non-asymptotic performance of our proposed predictive density estimators with other competing methods and obtain encouraging results.

MSC2020 subject classifications: Primary 62L20; secondary 60F15, 60G42.

Keywords: predictive density, α-divergences, predictive inference, optimal shrinkage, risk estimation, empirical Bayes.

1 Introduction

Predictive density estimation (prde) is one of the fundamental problems in statistical prediction analysis (see chapters 2, 7 and 10 of Aitchison and Dunsmore, 1975 and chapters 2, 3 and 9 of Geisser, 1993). Predictive density estimates assign probabilities to all possible future outcomes and can be used for better risk assessment and decision making than traditional point estimation methods (Mukherjee, 2013; Liang, 2002; Xu, 2005). Predictive densities have been widely used in a host of statistical applications in weather forecasting (Taylor and Buizza, 2004), finance (Tay and Wallis, 2000), information theory (Liang and Barron, 2005; Yuan and Clarke, 1999; Barron et al., 1998) as well as for model diagnostics and validation (Gelman et al., 2013; Pardoe, 2001; Gelman et al., 2014).

In this paper, we consider multivariate predictive density estimation under general divergence loss in a heteroskedastic Gaussian model. For point estimation, since the seminal work of James and Stein (1961) there has been substantial research toward

*The research was partly supported by NSF grants DMS-1811866, DMS-1916245 and JST grant KAKENHI 19K20222.

†Department of Statistics, University of Pennsylvania, edgeorge@wharton.upenn.edu

‡Department of Data Sciences and Operations, University of Southern California, gourab@usc.edu

§The Institute of Statistical Mathematics, yano@ism.ac.jp

 https://doi.org/10.1214/21-BA1264

understanding the risk properties of shrinkage estimators for the homoscedastic hierarchical normal models (see Fourdrinier et al., 2018, Efron, 2012 and the references therein). The concept of shrinkage is important because it provides an elegant framework for combining information from related populations and often leads to substantial improvements in the performances of estimators used for simultaneous inference. Komaki (2001, 2004); George et al. (2006); Brown et al. (2008) demonstrated the critical role of shrinkage priors for constructing efficient predictive density estimates (prdes) under Kullback-Leibler (KL) loss. High-dimensional decision theoretic parallels between prde under Kullback-Leibler loss and point-estimation under quadratic loss have been established in Fourdrinier et al. (2011); Xu and Liang (2010); George et al. (2012); Kubokawa et al. (2013); Mukherjee and Johnstone (2017, 2015); Yano and Komaki (2017); Ghosh et al. (2019). Ghosh et al. (2008), Suzuki and Komaki (2010), Maruyama and Strawderman (2010), L'Moudden and Marchand (2018) and Ghosh and Kubokawa (2018) extended those parallels for prde under general α-divergences. Using KL loss as a divergence measure between the true and estimated predictive density leads to convenient tractable analysis. However, predictive densities calibrated by KL loss are often non-robust to outliers and may under-estimate the variance or ignore important local attributes of the true density. To circumvent these issues, it is becoming increasingly popular in complex prediction approaches (Wang et al., 2018; Hernández-Lobato et al., 2016; Cichocki and Amari, 2010; Zhang et al., 2017) to use the class of α-divergences (Amari, 2009; Liese and Vajda, 2006; Barron et al., 1992) that covers a wide spectrum of divergence measures with contrasting attributes.

For predictive density estimation in multivariate Gaussian models, Ghosh et al. (2008) showed that the canonical minimax prde, which is the Bayes prde under uniform prior, is not admissible under general divergence loss for dimensions greater than 2. For dominating the canonical minimax prde, Ghosh et al. (2008) used prdes that are not necessarily Bayes, whereas Maruyama et al. (2019) established the domination results for the Bayes prde under the harmonic prior of George et al. (2006). Ghosh and Kubokawa (2018) established that the hierarchical Bayes prde has lower frequentist risk than that of the empirical Bayes prde in a regression set-up. While Ghosh et al. (2008); Ghosh and Kubokawa (2018) showcased enhanced predictive efficiency of the Bayes prde from non-informative priors over plug-in prdes, L'Moudden and Marchand (2018) proposed improving plug-in prdes directly. However, all of these results are based on the homoskedastic model. They also do not provide any prescription for selecting a particular prde among the host of feasible and admissible prdes.

Here, we study the prde in a heteroskedastic set-up where our target density is no longer spherically symmetric, and consider the problem of finding optimal shrinkage directions. We provide a data driven program for determining the optimal directions (location) and magnitude (scale) of shrinkage such that the resultant prde has minimal frequentist risk among a wide class of shrinkage estimators. Our proposed prde not only possesses the plug-in-dominance properties of the Bayes prde as in Ghosh et al. (2008), but also obtains the minimal risk among a wide class of shrinkage rules. These α-predictive risk optimality properties parallel those established by Xie, Kou, and Brown (2012) for point estimation in heteroskedastic hierarchical models.

Recent point estimation results of Xie, Kou, and Brown (2012); Xie et al. (2016); Tan (2015); Weinstein et al. (2018) have brought to light new shrinkage phenomena in heteroskedastic models. A hierarchical set-up specifying a second-level structure to motivate the shrinkage is considered, and the corresponding hyper-parameters are subsequently estimated. Whereas, the common practice is to choose the conjugate hierarchical structure and estimate the hyper-parameters through empirical Bayes maximum likelihood estimator (EBMLE) or empirical Bayes method of moments (EBMOM), we instead consider tuning the hyper-parameters by minimizing efficient risk estimates as in Xie, Kou, and Brown (2012).

A significant finding reported in Xie, Kou, and Brown (2012); Xie et al. (2016); Tan (2015); Weinstein et al. (2018) is that, under heteroskedasticity, EBMLE or EBMOM provide sub-optimal predictive performance and are far outperformed by algorithms tuned using risk estimation-based approaches. We establish asymptotic optimality of our proposed predictive methods akin to the point estimation results in Xie, Kou, and Brown (2012). These asymptotic properties are not shared by EBMLE or EBMOM based prdes. We establish asymptotic convergence rates of our risk estimates as dimension increases. Dimension independent non-asymptotic characterizations of the predictive risk of our proposed estimators are also provided using maximal inequalities for martingales. We compare these comprehensive results for general α-divergence with those of Xu and Zhou (2011) who studied empirical Bayes prde in spherically symmetric homoskedastic Gaussian model under KL loss. Our general α-divergence results well reconcile with the KL results in the existing literature. Through numerical studies, we demonstrate the benefits of using α-divergence based risk calibrated prdes over EBMLE or EBMOM based prdes. The direction of shrinkage and the shape of the optimally shrunken prdes greatly varies as α changes.

2 Predictive Set-Up

Predictive Sequence Model Consider observing a vector $\boldsymbol{X} = \{X_1, \ldots, X_n\}$ where X_i are independent among each other and X_i follows $N(\theta_i, \sigma_i^2)$, $i = 1, \ldots, n$. Based on observing $\boldsymbol{X}$ we would like to predict the unknown density of future observations $\boldsymbol{Y} = \{Y_i : 1 \leq i \leq n\}$, where Y_i independently follow $N(\theta_i, \nu_i^2)$. Here, σ_i, ν_i are known and thus, $r_i = \nu_i^2/\sigma_i^2$ is the known ratio of the future-to-past variances. The observed past $\boldsymbol{X}$ and unobserved future $\boldsymbol{Y}$ are only related through the unknown location parameter $\boldsymbol{\theta} = \{\theta_i : 1 \leq i \leq n\}$. This is the heteroskedastic version of the Gaussian predictive model studied in Komaki (2001); George et al. (2006); Brown et al. (2008); Xu and Zhou (2011). Let $\hat{p}(\boldsymbol{y}|\boldsymbol{x})$ be any prde for the true density $p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\nu}) = \prod_{i=1}^n \phi(y_i - \theta_i; \nu_i)$ of $\boldsymbol{Y}$. Note that, here we denote normal probability density function (pdf) with variance v by $\phi(\cdot\,; v)$ and thus, $\phi(y_i - \theta_i; \nu_i) = \nu_i^{-1/2}\phi(\nu^{-1/2}(y_i - \theta_i))$ where $\phi(\cdot)$ denotes standard normal pdf.

Loss Function Consider α-divergence as the measure of discrepancy between the prde and the true density. For any fixed $\alpha \in \mathbb{R}$, the loss defined as

$$L_{n,\alpha}(\boldsymbol{\theta}, \hat{p}(\,\cdot\,;\boldsymbol{x})) := \int p(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\nu})\, \ell_\alpha \left(\frac{\hat{p}(\boldsymbol{y}|\boldsymbol{x})}{p(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\nu})} \right) d\boldsymbol{y},$$

where the *f-divergence* function ℓ_α with domain in $[0, \infty)$ is:

$$\ell_\alpha(z) := \begin{cases} \{4/(1-\alpha^2)\}\{1 - z^{(1+\alpha)/2}\}, & \alpha \neq \pm 1, \\ -\log z, & \alpha = -1, \\ z \log z, & \alpha = 1. \end{cases}$$

Integrating over the density of the observed past, we have the predictive α-risk as:

$$R_{n,\alpha}(\boldsymbol{\theta}, \hat{p}) := \int p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\sigma}) L_{n,\alpha}(\boldsymbol{\theta}, \hat{p}(\,\cdot\,;\boldsymbol{x}))\, d\boldsymbol{x}.$$

When $\alpha = 3$, we have Pearson χ^2-divergence where as $\alpha = -3$ corresponds to Neyman χ^2-divergence; $\alpha = -1$ is KL loss and $\alpha = 1$ yields reverse KL loss; $\alpha = 0$ corresponds to Bhattacharyya-Hellinger (BH) loss Bhattacharyya (1943). Note that, the predictive risk might not be well-defined for all $\alpha \in \mathbb{R}$ on which we reflect later. For $\alpha_0 \in \{-1, 1\}$, $R_{n,\alpha_0}(\boldsymbol{\theta}, \hat{p}) = \lim_{\alpha \to \alpha_0} R_{n,\alpha}(\boldsymbol{\theta}, \hat{p})$, so that the KL and reverse KL predictive risk expressions will follow from the general α-predictive risk. Note that, the α predictive risk is the posterior predictive relative entropy regret criterion introduced in Sweeting et al. (2006) and differs from the reference prior inducing criterion studied in Clarke and Yuan (2010); Ghosh et al. (2011).

Hierarchical Set-Up and Shrinkage Prdes Next, we assume a hierarchical higher level exchangeable structure on the unknown location parameters. Let $\{\theta_i : 1 \leq i \leq n\}$ be independent and identically distributed (i.i.d.) from a $N(\eta, \tau)$ prior where $\eta \in \mathbb{R}$ and $\tau \geq 0$ are unknown hyper-parameters. This exchangeable hierarchical set-up, well-used in the literature (Xie, Kou, and Brown, 2012; Efron and Morris, 1973; Robbins, 1964; Zhang, 2003), allows partial pooling of information from quantities of interest for different yet related groups of populations. Define the integrated Bayes risk with respect to any prior π_n on $\boldsymbol{\theta}$ as $B_n(\pi, \hat{p}) = \int R_{n,\alpha}(\boldsymbol{\theta}, \hat{p}) \pi_n(\boldsymbol{\theta})\, d\boldsymbol{\theta}$. Let $\alpha_U = 1 + 2\min\{r_i : 1 \leq i \leq n\}$, $b = (1-\alpha)/2$ and $\bar{b} = 1 - b$. The following result (proved in George et al. (2021, Section 1.2)) shows that for all $\alpha \leq \alpha_U$ and for any product normal priors, i.e., $\pi_n(\boldsymbol{\theta}) = \prod \phi(\theta_i - \eta; \tau)$, the integrated Bayes risk has a well-defined minima and the resultant Bayes predictive density estimator is also a product of normal densities.

Lemma 2.1. *Consider $\theta_1, \ldots, \theta_n \overset{i.i.d.}{\sim} N(\eta, \tau)$ where $\eta \in \mathbb{R}$ and $\tau \geq 0$. Then, the Bayes predictive density estimate with respect to α-divergence loss for any fixed $\alpha \leq \alpha_U$ is*

$$\hat{p}[\eta, \tau](\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^{n} \phi\big(y_i - (\omega_i x_i + \bar{\omega}_i \eta);\ (r_i + b\omega_i)\sigma_i^2\big), \tag{2.1}$$

where $\omega_i = \tau/(\tau + \sigma_i^2)$ and $\bar{\omega}_i = 1 - \omega_i$.

Henceforth, we consider only α-divergences with $\alpha \leq \alpha_U$. Note that as $\min_i r_i \geq 0$, BH, KL, reverse KL, Neyman χ^2 divergences are always covered in our results. If $\min_i r_i \geq 1$, then the Pearson χ^2 divergence is also covered. Based on the above result, we consider the following flexible class $\mathcal{S} = \{\hat{p}[\eta, \tau] : \eta \in [\hat{q}_1, \hat{q}_2], 0 \leq \tau\}$ of shrinkage prdes, where $\hat{p}[\eta, \tau](\boldsymbol{y}|\boldsymbol{x}) = \prod_i \phi\big(y_i - (\omega_i x_i + \bar{\omega}_i \eta_i);\ (r_i + b\omega_i)\sigma_i^2\big)$, with $\omega_i := \omega_i[\tau] = \tau/(\tau + \sigma_i^2)$, $\bar{\omega}_i := \bar{\omega}_i[\tau] = 1 - \omega_i[\tau]$, and $\hat{q}_1, \hat{q}_2$ are $q/2$ and $(1 - q/2)$th quantiles of $X_1, \ldots, X_n$ for any prefixed $q \in (0, 1)$. The class is indexed by the location and scale hyper-parameters, η and τ. For any sensible shrinkage predictors, it is enough to confine the location hyper-parameter η within the $100q\%$ range of the observed data. The scale hyper-parameter τ varies over the non-negative axis. In the following section, we provide a methodology for choosing the hyper-parameters so that the resultant prde has optimal risk properties among all prdes in the class $\mathcal{S}$.

Extension to Non-diagonal Predictive Set-Ups The results and methodology developed here can encompass non-diagonal predictive set-ups where $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \boldsymbol{\Sigma}_p)$ and $\boldsymbol{Y} \sim N(\boldsymbol{\theta}, \boldsymbol{\Sigma}_f)$ with $\boldsymbol{\Sigma}_p$ and $\boldsymbol{\Sigma}_f$ being known positive definite matrices. For a non-diagonal prior $\boldsymbol{\theta} \sim N(\boldsymbol{\eta}, \Lambda)$, Lemma 2.1 can be extended as follows with α_U being $1 + 2\lambda_{\min}(\boldsymbol{\Sigma}_f \boldsymbol{\Sigma}_p^{-1})$, which is the generalization of the scalar case. The proof of the lemma is presented in George et al. (2021, Section 1.2).

Lemma 2.2. *Let $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \boldsymbol{\Sigma}_p)$, $\boldsymbol{Y} \sim N(\boldsymbol{\theta}, \boldsymbol{\Sigma}_f)$, and $\boldsymbol{\theta} \sim N(\boldsymbol{\eta}, \Lambda)$ with known positive definite matrices $\boldsymbol{\Sigma}_p$, $\boldsymbol{\Sigma}_f$, and Λ. Then, the Bayes predictive density estimate with respect to α-divergence loss for $\alpha < \alpha_U$ is*

$$\hat{p}[\boldsymbol{\eta}, \Lambda](\boldsymbol{y} \mid \boldsymbol{x}) = \boldsymbol{\phi}\big(\boldsymbol{y} - (\Omega\, \boldsymbol{x} + \overline{\Omega}\, \boldsymbol{\eta});\ \boldsymbol{\Sigma_f} + b\,(\boldsymbol{\Sigma}_p^{-1} + \Lambda^{-1})^{-1}\big), \tag{2.2}$$

where $\Omega := \Lambda(\Lambda + \boldsymbol{\Sigma}_p)^{-1}$ and $\overline{\Omega} := I - \Omega$.

For tractable shrinkage classes, we need to impose lower dimensional structures on $\boldsymbol{\eta}$ and Λ in (2.2). Perhaps, the most popular choice is $\boldsymbol{\eta} = \eta \mathbf{1}$ and $\Lambda = \tau I$, which extends the class $\mathcal{S}$ based on (2.1) to the non-diagonal set-up. In the following sections, we describe our method and its associated results for the class $\mathcal{S}$. However, note that the methodology can be extended to other shrinkage classes based on (2.2).

Hereon, we describe our method first for the diagonal predictive set-up as it produces comparatively simpler expressions that can be intuitively studied and compared to the point estimation results of Xie, Kou, and Brown (2012) and the predictive KL results in Xu and Zhou (2011). The general results for non-diagonal set-ups along with their complete proofs are provided in George et al. (2021, Section 1).

3 Risk Estimation and Hyper-parameter Calibration

Denote by $R_{n,\alpha}(\boldsymbol{\theta}; \eta, \tau)$ the risk $R_{n,\alpha}(\boldsymbol{\theta}, \hat{p}[\eta, \tau])$ of any arbitrary member $\hat{p}[\eta, \tau]$ in $\mathcal{S}$. The following result proved in George et al. (2021, Section 1.3) shows that this multivariate predictive risk decouples as functions of the corresponding coordinate-wise risks and subsequently can be explicitly written through closed form expressions as functions of

$\eta, \tau, \boldsymbol{\theta}$ and α. Letting $\tau \to \infty$, we get the second display in Theorem 2.4 of Ghosh et al. (2008).

Theorem 3.1. *For $\alpha \le \alpha_U$ and $\alpha \neq \pm 1$, the risk of any prde $\hat{p}[\eta, \tau] \in \mathcal{S}$ can be expressed as*

$$\log\big(1 - c_\alpha\, R_{n,\alpha}(\boldsymbol{\theta}; \eta, \tau)\big) = \sum_{i=1}^{n} H_i(\theta_i, \eta, \tau; \alpha), \text{ where, } c_\alpha = (1-\alpha^2)/4, \text{ and,}$$
$$H_i(\theta_i, \eta, \tau; \alpha) = f\big((\theta_i - \eta)^2, \omega_i[\tau], r_i, \sigma_i, (1-\alpha)/2\big) \text{ with } \omega_i[\tau] = \tau/(\tau + \sigma_i^2),$$
$$f(t, w, r, \sigma, b) = \frac{\bar{b}}{2}\log\left(\frac{r}{r + wb}\right) + \frac{1}{2}\log\left(\frac{r + wb}{r + wb^2 + w^2 b\bar{b}}\right) - \frac{b\bar{b}\bar{w}^2 t}{2\sigma^2(r + wb^2 + w^2 b\bar{b})}$$

and $\bar{w} = 1 - w$, $\bar{b} = 1 - b$.

The risk for the KL and reverse KL losses can be derived from the above expression by noting that for $\alpha = \alpha_0 \in \{-1, 1\}$,

$$R_{n,\alpha_0} = \lim_{\alpha \to \alpha_0} R_{n,\alpha} = \lim_{\alpha \to \alpha_0} \frac{1 - \exp\{\sum_{i=1}^n H_i(\theta_i, \eta, \tau; \alpha)\}}{c_\alpha} = 2\alpha_0 \sum_{i=1}^{n} \frac{\partial}{\partial \alpha} H_i(\theta_i, \eta, \tau; \alpha_0),$$

where, the last equality follows from the fact that $H_i = 0$ when $\alpha \in \{-1, 1\}$, and L'Hôpital's rule. Thus,

$$R_{n,-1} = \sum_{i=1}^{n} f_b((\theta_i - \eta)^2, w_i[\tau], r_i, \sigma_i; 1) \text{ and } R_{n,1} = -\sum_{i=1}^{n} f_b((\theta_i - \eta)^2, w_i[\tau], r_i, \sigma_i; 0),$$

where, $f_b(t, w, r, \sigma; b) = \frac{\partial}{\partial b} f(t, w, r, \sigma, b)$. Finally, it yields

$$R_{n,-1}(\boldsymbol{\theta}; \eta, \tau) = \sum_{i=1}^{n} \frac{\log(1 + r_i^{-1}\omega_i)}{2} + \frac{\bar{\omega}_i^2(\theta_i - \eta)^2 - \omega_i\bar{\omega}_i\sigma_i^2}{2\sigma_i^2(r_i + \omega_i)},$$
$$R_{n,1}(\boldsymbol{\theta}; \eta, \tau) = \sum_{i=1}^{n} \frac{\bar{\omega}_i^2(\theta_i - \eta)^2 + \omega_i^2\sigma_i^2}{2\,\sigma_i^2 r_i}.$$

Note that $R_{n,-1}$ matches the KL risk expression in equation (11) of Xu and Zhou (2011). We next estimate the predictive risk $R_{n,\alpha}(\boldsymbol{\theta}; \eta, \tau)$. Noting that $(X_i - \eta)^2 - \sigma_i^2$ is an unbiased estimator for $(\theta_i - \eta)^2$, an unbiased estimate of $H_i(\theta_i, \eta, \tau; \alpha)$ is given by $\hat{H}_i(\eta, \tau; \alpha)$, where

$$\hat{H}_i(\eta, \tau; \alpha) = f\left((X_i - \eta)^2 - \sigma_i^2,\ \frac{\tau}{\tau + \sigma_i^2},\ r_i,\ \sigma_i,\ \frac{1-\alpha}{2}\right).$$

Consider their average $\hat{\mathcal{H}}_n(\tau,\eta;\alpha) = n^{-1}\sum_{i=1}^n \hat{H}_i(\tau,\eta;\alpha)$. For any fixed $\alpha \le \alpha_U$, select the hyper-parameters that minimize the *average risk* estimate, i.e.,

$$(\hat{\eta}_{n,\alpha}, \hat{\tau}_{n,\alpha}) = \underset{0\le\tau,\,\eta\in[\hat{q}_1,\hat{q}_2]}{\arg\min} \quad \operatorname{sign}(\alpha^2-1)\ \hat{\mathcal{H}}_n(\eta,\tau;\alpha) \tag{3.1}$$

to obtain our proposed prde $\hat{p}[\hat{\eta}_{n,\alpha},\hat{\tau}_{n,\alpha}](\boldsymbol{y}|\boldsymbol{x})$. Thus, when $|\alpha| < 1$, we maximize $\hat{\mathcal{H}}_n(\eta,\tau;\alpha)$ and we minimize it when $|\alpha| > 1$. However, note that in both the cases, this corresponds to minimizing risk estimates of the actual risk. For $\alpha = \pm 1$, we directly minimize the unbiased estimates of $R_{n,1}$ and $R_{n,-1}$. Taking the limit of the hyper-parameters $(\hat{\eta}_{n,\alpha},\hat{\tau}_{n,\alpha})$ as $\alpha \to 1_-$ or $\alpha \to -1_+$ also yields similar results.

Figure 1 shows the BH risk of prdes for $n = 5$ and 10 at $\boldsymbol{\theta} = t\mathbf{1}$ where t varies from 0 to ∞. The risks of the best prde in $\mathcal{S}$ (which is characterized later in (4.1)) are plotted in blue. The risk of our proposed method is calculated by Monte-Carlo integration and is plotted in red. In dotted black lines, we have the risk of the best invariant prde $\hat{p}_{\mathsf{U}}$ which is the Bayes prde from the uniform prior. We see that compared to $\hat{p}_{\mathsf{U}}$ significant gains in risk can be obtained by estimators in $\mathcal{S}$ when $|t|$ is near the origin and the gains decrease when $|t|$ is large. Also, the risk of the proposed method is reasonably close to the minimum attainable risk in $\mathcal{S}$. Next, we rigorously document these risk properties of prdes tuned by the proposed procedure.

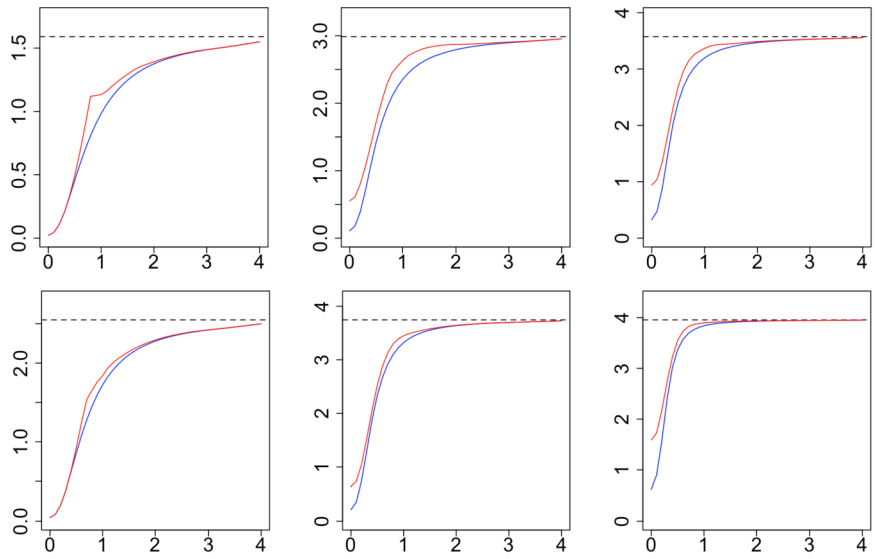

Figure 1: Plot of BH risks of the prdes in a homoskedastic normal model at $\boldsymbol{\theta} = t\mathbf{1}$ as t varies in the abscissa. Here, $\sigma_i = 1$ and $r_i = r$ for all $i = 1,\ldots,n$. From left to right, $r = 1, 0.25, 0.1$ respectively; $n = 5$ in the top and $n = 10$ in the bottom plots. The risk of the best-invariant predictive density (dotted black), the risk of the oracle estimator based on hyper-parameters in (4.1) (in blue) and the risk of our proposed method (in red) are presented here.

4 Theory Results

We first establish a non-asymptotic concentration bound on the deviation of $\hat{\mathcal{H}}_n$ from the true log-risk. Consider the expected absolute deviation:

$$D_{n,\alpha}(\boldsymbol{\theta},\tau,\eta) = \mathbb{E}\left[\sup_{\tau\geq 0}\left|n^{-1}\log\left\{1-c_\alpha\, R_{n,\alpha}(\boldsymbol{\theta};\eta,\tau)\right\} - \hat{\mathcal{H}}_n(\eta,\tau;\alpha)\right|\right].$$

We establish an upper bound on $D_{n,\alpha}$ that depends on the L_2 norm of the signal strength:

$$g_n(\boldsymbol{\theta},\eta) := \max\left\{1, \left\{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\theta_i-\eta}{\sigma_i}\right)^2\right\}^{1/2}\right\}.$$

Theorem 4.1. *For any $\alpha \leq \alpha_U$, any fixed $\eta\in\mathbb{R}$ and $\boldsymbol{\theta}\in\mathbb{R}^n$, for all $n\geq 1$,*

$$D_{n,\alpha}(\boldsymbol{\theta},\tau,\eta) \leq \kappa_0\, r_*^{-1}\, \max\{1,|c_\alpha|\}\, g_n(\boldsymbol{\theta},\eta)\, n^{-1/2}\ ,$$

where, $\kappa_0 = 12$ is an absolute constant and $r_* = \inf_i r_i$.

When $\alpha\to\pm 1$, both the $D_{n,\alpha}(\boldsymbol{\theta},\tau,\eta)$ and the $c_\alpha = (1-\alpha^2)/4$ on the RHS tends to 0. Applying L'Hôpital's rule yields the analogous bound $\mathbb{E}\sup_{\tau\geq 0}\left|n^{-1}R_{n,\alpha}(\boldsymbol{\theta};\eta,\tau)\right\}-\hat{\mathcal{H}}_n(\eta,\tau;\alpha)| \leq \kappa_0\, r_*^{-1} g_n(\theta,\eta)\, n^{-1/2}$ for $\alpha=\pm 1$. The proof of the theorem is presented in George et al. (2021, Section 1.3).

Our next result which uses Theorem 4.1 shows that our proposed risk estimate approximates the average of the logarithm of the true multivariate risk uniformly well for all prdes in the shrinkage class $\mathcal{S}$. Thus, calibrating the hyper-parameters by minimizing the risk estimates is a sensible choice. To facilitate shorter mathematical proofs we assume the following asymptotic conditions:

[A1] $\overline{\lim}_{n\to\infty} n^{-1}\sum_{i=1}^{n}\sigma_i^2 < \infty$, $\underline{\lim}_i \sigma_i > 0$ and $0 < \underline{\lim}_i r_i \leq \overline{\lim}_i r_i < \infty$,
[A2] $\overline{\lim}_{n\to\infty} n^{-1}\sum_{i=1}^{n}\theta_i^2 < \infty$.

Though these conditions may be further relaxed as noted after Theorem 3.1 of Xie, Kou, and Brown (2012), we do not seek the full generality as the conditions are not restrictive and the proofs presented under these assumptions contain all the essential statistical perspectives.

Theorem 4.2. *Under Assumptions A1–A2, for any $\alpha\leq\alpha_U$ and $a_n = o(n^{1/2})$,*

$$a_n\left(\sup_{\eta\in[\hat{q}_1,\hat{q}_2],\tau\geq 0}\left|n^{-1}\log\left\{1-c_\alpha\, R_{n,\alpha}(\boldsymbol{\theta};\eta,\tau)\right\} - \hat{\mathcal{H}}_n(\eta,\tau;\alpha)\right|\right) \to 0 \text{ in } L_1 \quad \text{as } n\to\infty\ .$$

The above result (proved in George et al. (2021, Section 1.5)) shows that the risk estimates $\hat{\mathcal{H}}_n$ have near-parametric $\sqrt{n}$-rates of convergence barring some poly-log terms. Thus, we expect the risk estimates to be reasonably precise even for moderate dimensions n. This attribute is reflected in the simulation studies in Section 5.

To study the risk properties of our proposed prde $\hat{p}[\hat{\eta}_{n,\alpha}, \hat{\tau}_{n,\alpha}]$, we next introduce the oracle risk (OR) hyper-parameters as those which minimize the true risk function:

$$(\eta^{\mathsf{or}}_{n,\alpha}, \tau^{\mathsf{or}}_{n,\alpha}) = \underset{0\leq\tau,\, \eta\in[\hat{q}_1,\hat{q}_2]}{\arg\min} R_{n,\alpha}(\boldsymbol{\theta}; \eta, \tau). \tag{4.1}$$

These oracle choices are not really estimators since they depend on the unknown $\boldsymbol{\theta}$ values. They are not obtainable in practice but provide the theoretical benchmark that one can ever hope to reach. Indeed, no prdes in $\mathcal{S}$ can have smaller risk than the oracle risk prde $\hat{p}[\eta^{\mathsf{or}}_{n,\alpha}, \tau^{\mathsf{or}}_{n,\alpha}]$.

Consider the average logarithmic ratio of the risk deviations

$$\rho_{n,\alpha}(\boldsymbol{\theta}) = -\frac{\gamma_\alpha}{n} \log\left\{ \frac{1 - c_\alpha R_{n,\alpha}(\boldsymbol{\theta}; \eta^{\mathsf{or}}_{n,\alpha}, \tau^{\mathsf{or}}_{n,\alpha})}{1 - c_\alpha \mathbb{E}\{L_{n,\alpha}(\boldsymbol{\theta}; \hat{\eta}_{n,\alpha}, \hat{\tau}_{n,\alpha})\}} \right\} \text{ where } \gamma_\alpha = \text{sign}(\alpha^2 - 1).$$

By construction, $\rho_{n,\alpha}(\boldsymbol{\theta}) \geq 0$. For any fixed $0 < a_0 < a_1 < \infty$ and $\epsilon > 0$, consider the neighborhood $\Theta[\varepsilon, a_0, a_1] := \{(\eta, \tau) : |\tau - \tau^{\mathsf{or}}_{n,\alpha}| \leq \varepsilon, |\eta - \eta^{\mathsf{or}}_{n,\alpha}| \leq \varepsilon, \tau \geq 0, \eta \in [a_0, a_1]\}$ around oracle hyper-parameters. We impose the following regularity condition on the sensitiveness (non-flatness) on the true risk functions around the oracle hyper-parameters.

[A3] For any $\varepsilon > 0$ and $-\infty < a_0 < a_1 < \infty$,

$$\lim_{n\to\infty} n^{-1/2} \left[\inf_{(\eta,\tau)\notin\Theta[\varepsilon,\, a_0,\, a_1]} \log\left(\frac{1 - c_\alpha R_{n,\alpha}(\boldsymbol{\theta}; \eta, \tau)}{1 - c_\alpha R_{n,\alpha}(\boldsymbol{\theta}; \eta^{\mathsf{or}}_{n,\alpha}, \tau^{\mathsf{or}}_{n,\alpha})} \right)^{\gamma_\alpha} \right] = \infty.$$

The following result shows that our estimated hyper-parameters are close to the oracle hyper-parameters and our proposed prde has risk close to the oracle risk. This property is not shared by the EBMLE or EBMOM. The estimating equations for calibrating the hyper-parameters based on the EBMLE and EBMOM (discussed in the following section), differ from our proposed methodology so that the estimated hyper-parameters can be highly different (see cases IV–V in Section 5). In such cases the EBMLE and EBMOM tuned prdes have much higher risk than the oracle unless the risk function is completely flat. As our proposed estimator $\hat{p}[\hat{\eta}_{n,\alpha}, \hat{\tau}_{n,\alpha}]$ is asymptotically nearly as good as the oracle prde, its asymptotic risk is no larger than that of any other prdes in the general class $\mathcal{S}$.

Theorem 4.3. *For any $\alpha \leq \alpha_U$, under assumptions A1-A2, the logarithmic ratio $\rho_{n,\alpha}(\boldsymbol{\theta})$ converges asymptotically satisfying $\limsup_{n\to\infty} n^{1/2}\rho_{n,\alpha}(\boldsymbol{\theta}) < \infty$. Additionally, with assumption A3, we have:*

$$(\hat{\eta}_{n,\alpha} - \eta^{\mathsf{or}}_{n,\alpha}, \hat{\tau}_{n,\alpha} - \tau^{\mathsf{or}}_{n,\alpha}) \to 0 \text{ in probability as } n \to \infty \ .$$

The above result is proved in George et al. (2021, Section 1.6). It shows that as n increases, the average logarithmic ratio of risk deviations from the oracle (ALRORD) converges to 0 for any prde calibrated by the proposed method. As such, as $n \to \infty$ the ALRORD of any prde calibrated by the proposed method is always bound above by $O_p(n^{-1/2})$. Assumption A3 implies that the ALRORD for an arbitrary $\hat{p}[\eta, \tau]$ based on

hyper-parameter (η, τ) cannot be bounded below $O_p(n^{-1/2})$ unless the hyper-parameter (η, τ) is within any prefixed ϵ-neighborhood of the oracle hyper-parameter values. This ensures that as $n \to \infty$, the hyper-parameter estimates in (3.1) converge to the oracle values in (4.1).

5 Simulation Experiments

We conduct six simulation experiments to compare the performance of our estimation methodology with competing methods for calibrating estimators in $\mathcal{S}$. We consider the EBMLE tuned prde which uses hyper-parameters

$$(\hat{\eta}_{\mathsf{ML}}, \hat{\tau}_{\mathsf{ML}}) = \underset{\eta \in \mathbb{R}, \tau \geq 0}{\arg\min} \sum_{i=1}^{n} \log(\tau + \sigma_i^2) + \frac{(X_i - \eta)^2}{\tau + \sigma_i^2},$$

the EBMOM tuned prde whose hyper-parameters are solutions to the following equations

$$\hat{\eta}_{\mathsf{MM}} = \frac{\sum_{i=1}^{n} (\sigma_i^2 + \tau)^{-1} X_i}{\sum_{i=1}^{n} (\sigma_i^2 + \tau)^{-1}} \text{ and } \hat{\tau}_{\mathsf{MM}} = \frac{1}{n-1} \left(\sum_{i=1}^{n} (X_i - \hat{\eta}_{\mathsf{MM}})^2 - \frac{n-1}{n} \sum_{i=1}^{n} \sigma_i^2 \right)_+,$$

the extended James-Stein (Xie, Kou, and Brown, 2012) based prde in $\mathcal{S}$ with hyper-parameters

$$\hat{\eta}_{\mathsf{JS}} = \frac{\sum_{i=1}^{n} \sigma_i^{-2} X_i}{\sum_{i=1}^{n} \sigma_i^{-2}} \text{ and } \hat{\tau}_{\mathsf{JS}} = \left(\frac{1}{n} \sum_{i=1}^{n} \sigma_i^2 \right) \left(\frac{\sum_{i=1}^{n} \sigma_i^{-2} (X_i - \hat{\eta}_{\mathsf{JS}})^2}{(n-3)} - 1 \right)_+,$$

as well as the completely non-informative (NI) prde which has $\tau \to \infty$ and the oracle estimator of (4.1). Additionally, we also consider the Bayes prde $\hat{p}_{\mathsf{C}}$ from the heavy tailed Cauchy prior.

The six simulation regimes are inspired by experiments in section 4 of Xie, Kou, and Brown (2012). In Cases I to V, we have Gaussian noise with mean 0 and variances σ_i^2 being i.i.d. from Uniform$[0.5, 1.5]$ distribution. Thus, the mean noise variance is 1. In Case I, we generate $\boldsymbol{\theta} \sim N(0, 2I_n)$ and the r_i as uniformly distributed between 0.1 and 1. The average signal-to-noise-ratio (snr) is 2 here. Note, that this case is in perfect congruence with the hierarchical normal set-up of Section 2. The remaining cases are different from the normal models and conjugate priors set-ups of Lemmas 2.1 and 2.2. In Case II, θ_i are i.i.d. from a Uniform prior on $[-\sqrt{3}, \sqrt{3}]$ and thus has variability 1. Here, we consider stratification in $\boldsymbol{r}$. The r_i are generated from a mixture of three uniform distributions with significantly different supports. In Case III, we introduce dependence between the means and the future variances by setting $\boldsymbol{\theta} = \boldsymbol{r}$ while $\boldsymbol{r}$ is i.i.d. form Uniform$[0.1, 2]$. Case IV is similar to case 3, except r_i's are no longer bounded as before but are generated from an inverse-Chisquare with 5 degrees of freedom. In Case V, $\{(\theta_i, r_i) : i = 1, \ldots, n\}$ are independent among themselves, r_i takes two possible values with $P(r_i = 10) = 0.7$ and $P(r_i = 1) = 0.3$, and the true mean values are generated conditionally on the r_i values with $[\theta_i | r_i = 10]$ and $[\theta_i | r_i = 1]$ both being normal distributions with mean 0 and standard deviations 0.1 and 1, respectively. In Case VI, we consider $\boldsymbol{\theta}, \boldsymbol{r}$ as in case 3. However, here the noise is no longer Gaussian but is from a uniform distribution with mean 0 and variance 1. The snr is 1.4 in this set-up.

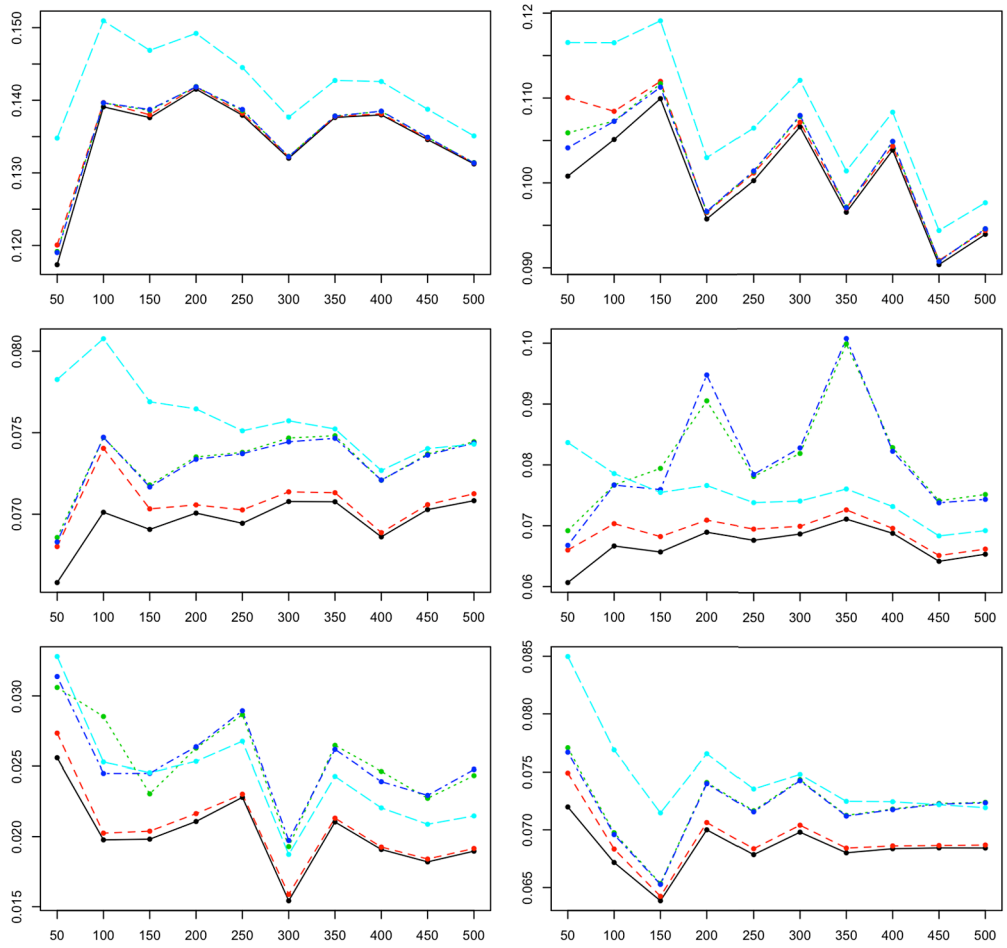

Figure 2: Adjusted BH risk of the prdes based on the oracle estimator of (4.1) (in black), EBMLE (in blue), EBMOM (in green), Cauchy prior based Bayes prde (in cyan) and our proposed method (in red) are plotted as n varies along the x-axis.

We report the Bhattacharyya-Hellinger predictive risk for the six cases in Table 1 and Figure 2. In each of the six cases, we generate $\{\theta_i, \sigma_i, \nu_i : 1 \leq i \leq n\}$ values once from the corresponding model and then calculate the adjusted BH risk for different $\hat{\eta}$ and $\hat{\tau}$ estimates across 100 replicates. We calculate the BH risk adjusted for dimension by $\text{ABH}_n(\boldsymbol{\theta}; \eta, \tau) = 1 - \{1 - c_0\, R_{n,0}(\boldsymbol{\theta}; \eta, \tau)\}^{1/n}$. For each case and n, the reported adjusted BH predictive risk in Figure 2 is the average of the ABH value across the 100 replicates.

In Table 1, $n = 1000$ and so, it reflects the asymptotic risks of these estimators. Figure 2 compares the risk profiles as sample size varies (NI and JS were avoided in the display for their significantly higher error rates reduced clarity in some of the plots).

	Proposed	EBMOM	EBMLE	JS	NI	Cauchy
Case I	0.05	0.17	0.19	0.04	24.63	1.63
Case II	0.23	0.71	0.65	0.16	48.30	2.32
Case III	0.21	4.69	4.81	6.70	69.70	2.96
Case IV	0.51	26.82	23.07	26.43	119.47	3.28
Case V	0.55	22.40	23.00	19.00	105.97	9.07
Case VI	0.19	5.70	5.67	5.77	72.54	2.94

Table 1: As $n \to \infty$, the limiting BH predictive risk (adjusted for dimensions) of different shrinkage prdes is reported in % of excess risk over that of the oracle estimator in (4.1).

From the table, we see our proposed estimation method is asymptotically close to the oracle in all the cases where as the figure displays that suitable accuracy can be attained for moderate n across all the concerned scenarios. The EBMLE and the EBMOM have risks similar to ours in cases I and II but has significantly worse performance in the other cases when there is dependence between $\boldsymbol{\theta}$ and $\boldsymbol{r}$. The Cauchy prior based Bayes prde $\hat{p}_{\mathsf{C}}$ performs considerably better than the EBMLE and EBMOM in Cases III, IV and V, when n is large (see Table 1), though its risk is still significantly higher than that of the proposed method. In moderate dimensions, the relative risk of $\hat{p}_{\mathsf{C}}$ is substantially higher than the EBMLE, the EBMOM and the proposed methods. As such for $n \leq 200$, the risk curve for $\hat{p}_{\mathsf{C}}$ is higher than the EBMLE or EBMOM in all cases except IV (see Figure 2). The JS based prde has erratic asymptotic risk behavior as it often fails to adapt to the heterogeneity in the data and the non-informative prior based prde has poor performance across all regimes. We present two numerical examples in the supplementary materials (George et al., 2021) where suboptimal performance of the JS based prde is witnessed.

6 Discussion and Future Work

We developed a risk estimation based methodology for tuning linearly shrunken prdes in Gaussian models under general α-divergence losses. The proposed risk estimation based method will be particularly useful under heterogeneity. If the set-up is homoskedastic ($\sigma_i = \sigma$ and $\nu_i = \nu$ for all i), then in high dimensions the proposed method will produce hyper-parameter estimates similar to the EBMOM and EBMLE. An interesting topic for future work will be to introspect the roles of prde under α-divergences when covariances are unknown as studied in Kato (2009) for the KL loss. Also, in applications it is important to choose an α-divergence loss that is tailored to the specific prediction task at hand (Kempthorne et al., 1988; Rosasco et al., 2004). Following Nguyen et al. (2009); Gül and Zoubir (2016) it will be important to study the roles different α predictive risks in hypothesis testing and classification problems. Another important direction would be understanding the roles of adaptive calibration under α-risk in the presence of latent structures in the mean parameters such as the sparsity restrictions studied in Mukherjee (2013); Yano et al. (2021) for the KL loss. Finally, extending the risk estimation based methodology developed here to non-normal models will be useful.

Supplementary Material

Supplementary materials for Optimal Shrinkage Estimation of Predictive Densities under α-divergences (DOI: 10.1214/21-BA1264SUPP; .pdf). The supplement contains the proofs of all the results stated in this paper as well as two numerical examples regarding prdes.

References

Aitchison, J. and Dunsmore, I. R. (1975). *Statistical prediction analysis*. Cambridge University Press. MR0408097. 1139

Amari, S.-I. (2009). "α-divergence is unique, belonging to both f-divergence and Bregman divergence classes." *IEEE Transactions on Information Theory*, 55(11): 4925–4931. MR2596950. doi: https://doi.org/10.1109/TIT.2009.2030485. 1140

Barron, A., Rissanen, J., and Yu, B. (1998). "The minimum description length principle in coding and modeling." *IEEE Transactions on Information Theory*, 44(6): 2743–2760. MR1658898. doi: https://doi.org/10.1109/18.720554. 1139

Barron, A. R., Gyorfi, L., and van der Meulen, E. C. (1992). "Distribution estimation consistent in total variation and in two types of information divergence." *IEEE Transactions on Information Theory*, 38(5): 1437–1454. MR1178189. doi: https://doi.org/10.1109/18.149496. 1140

Bhattacharyya, A. (1943). "On a measure of divergence between two statistical populations defined by their probability distributions." *Bulletin of the Calcutta Mathematical Society*, 99–109. MR0010358. 1142

Brown, L. D., George, E. I., and Xu, X. (2008). "Admissible predictive density estimation." *Ann. Statist.*, 36(3): 1156–1170. MR2418653. doi: https://doi.org/10.1214/07-AOS506. 1140, 1141

Cichocki, A. and Amari, S.-i. (2010). "Families of alpha-beta-and gamma-divergences: Flexible and robust measures of similarities." *Entropy*, 12(6): 1532–1568. MR2659408. doi: https://doi.org/10.3390/e12061532. 1140

Clarke, B. and Yuan, A. (2010). "Reference priors for empirical likelihoods." *Frontiers of Statistical Decision Making and Bayesian Analysis: In honor of James O. Berger*, 56–68. 1142

Efron, B. (2012). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press. MR2724758. doi: https://doi.org/10.1017/CBO9780511761362. 1140

Efron, B. and Morris, C. (1973). "Stein's estimation rule and its competitors—an empirical Bayes approach." *Journal of the American Statistical Association*, 68(341): 117–130. MR0388597. 1142

Fourdrinier, D., Marchand, É., Righi, A., and Strawderman, W. E. (2011). "On improved

predictive density estimation with parametric constraints." *Electron. J. Stat.*, 5: 172–191. MR2792550. doi: https://doi.org/10.1214/11-EJS603. 1140

Fourdrinier, D., Strawderman, W. E., and Wells, M. T. (2018). *Shrinkage estimation*. Springer. MR3887633. doi: https://doi.org/10.1007/978-3-030-02185-6. 1140

Geisser, S. (1993). *Predictive inference*, volume 55 of *Monographs on Statistics and Applied Probability*. New York: Chapman and Hall. An introduction. MR1252174. doi: https://doi.org/10.1007/978-1-4899-4467-2. 1139

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*. Chapman and Hall/CRC. MR3235677. 1139

Gelman, A., Hwang, J., and Vehtari, A. (2014). "Understanding predictive information criteria for Bayesian models." *Statistics and computing*, 24(6): 997–1016. MR3253850. doi: https://doi.org/10.1007/s11222-013-9416-2. 1139

George, E. I., Liang, F., and Xu, X. (2006). "Improved minimax predictive densities under Kullback-Leibler loss." *Ann. Statist.*, 34(1): 78–91. MR2275235. doi: https://doi.org/10.1214/009053606000000155. 1140, 1141

George, E. I., Liang, F., and Xu, X. (2012). "From minimax shrinkage estimation to minimax shrinkage prediction." *Statist. Sci.*, 27(1): 82–94. MR2953497. doi: https://doi.org/10.1214/11-STS383. 1140

George, E. I., Mukherjee, G., and Yano, K. (2021). "Supplementary Material of "Optimal Shrinkage Estimation of Predictive Densities Under α-Divergences"." *Bayesian Analysis*. doi: https://doi.org/10.1214/21-BA1264SUPP. 1142, 1143, 1146, 1147, 1150

Ghosh, M. and Kubokawa, T. (2018). "Hierarchical Bayes versus empirical Bayes density predictors under general divergence loss." *Biometrika*. MR3949318. doi: https://doi.org/10.1093/biomet/asy073. 1140

Ghosh, M., Kubokawa, T., and Datta, G. S. (2019). "Density prediction and the Stein phenomenon." *Sankhya A*, 1–23. MR4136238. doi: https://doi.org/10.1007/s13171-019-00186-z. 1140

Ghosh, M., Mergel, V., and Datta, G. S. (2008). "Estimation, prediction and the Stein phenomenon under divergence loss." *J. Multivariate Anal.*, 99(9): 1941–1961. MR2466545. doi: https://doi.org/10.1016/j.jmva.2008.02.002. 1140, 1144

Ghosh, M., Mergel, V., and Liu, R. (2011). "A general divergence criterion for prior selection." *Annals of the Institute of Statistical Mathematics*, 63(1): 43–58. MR2748933. doi: https://doi.org/10.1007/s10463-009-0226-4. 1142

Gül, G. and Zoubir, A. M. (2016). "Robust hypothesis testing with *alpha*-divergence." *IEEE Transactions on Signal Processing*, 64(18): 4737–4750. MR3538377. doi: https://doi.org/10.1109/TSP.2016.2569405. 1150

Hernández-Lobato, J. M., Li, Y., Rowland, M., Hernández-Lobato, D., Bui, T., and Turner, R. (2016). "Black-box α-divergence minimization." In *Proceedings of The*

33rd International Conference on Machine Learning. International Machine Learning Society. 1140

James, W. and Stein, C. M. (1961). "Estimation with quadratic loss." In *Proceedings of the 4th Berkeley Symposium on Probability and Statistics*, 367–379. MR0133191. 1139

Kato, K. (2009). "Improved prediction for a multivariate normal distribution with unknown mean and variance." *Ann. Inst. Statist. Math.*, 61(3): 531–542. MR2529965. doi: https://doi.org/10.1007/s10463-007-0163-z. 1150

Kempthorne, P. J. et al. (1988). "Controlling risks under different loss functions: The compromise decision problem." *Annals of statistics*, 16(4): 1594–1608. MR0964940. doi: https://doi.org/10.1214/aos/1176351055. 1150

Komaki, F. (2001). "A shrinkage predictive distribution for multivariate normal observables." *Biometrika*, 88(3): 859–864. MR1859415. doi: https://doi.org/10.1093/biomet/88.3.859. 1140, 1141

Komaki, F. (2004). "Simultaneous prediction of independent Poisson observables." *Ann. Statist.*, 32(4): 1744–1769. MR2089141. doi: https://doi.org/10.1214/009053604000000445. 1140

Kubokawa, T., Marchand, É., Strawderman, W. E., and Turcotte, J.-P. (2013). "Minimaxity in predictive density estimation with parametric constraints." *Journal of Multivariate Analysis*, 116: 382–397. MR3049911. doi: https://doi.org/10.1016/j.jmva.2013.01.001. 1140

Liang, F. (2002). *Exact minimax procedures for predictive density estimation and data compression*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)–Yale University. MR2703233. 1139

Liang, F. and Barron, A. (2005). *Exact Minimax Predictive Density Estimation and MDL*, chapter 7, 177–194. Advances in Minimum Description Length: Theory and Applications (P. Grunwald, I. Myung and M. Pitt eds). MIT Press. 1139

Liese, F. and Vajda, I. (2006). "On divergences and informations in statistics and information theory." *IEEE Transactions on Information Theory*, 52(10): 4394–4412. MR2300826. doi: https://doi.org/10.1109/TIT.2006.881731. 1140

L'Moudden, A. and Marchand, É. (2018). "On Predictive Density Estimation under α-divergence Loss." *arXiv preprint arXiv:1806.02600*. MR3989319. doi: https://doi.org/10.3103/S1066530719020030. 1140

Maruyama, Y., Matsuda, T., and Ohnishi, T. (2019). "Harmonic Bayesian prediction under α-divergence." *IEEE Transactions of Information Theory*. MR4009238. doi: https://doi.org/10.1109/TIT.2019.2915245. 1140

Maruyama, Y. and Strawderman, W. (2010). "Bayesian predictive densities for linear regression models under α-divergence loss: some results and open problems." Manuscript available at: arXiv:1002.3786v1. MR3202501. doi: https://doi.org/10.1214/11-IMSCOLL803. 1140

Mukherjee, G. (2013). "Sparsity and Shrinkage in Predictive Density Estimation." Ph.D. thesis, Stanford University. MR4187552. 1139, 1150

Mukherjee, G. and Johnstone, I. M. (2015). "Exact minimax estimation of the predictive density in sparse Gaussian models." *Annals of Statistics*. MR3346693. doi: https://doi.org/10.1214/14-AOS1251. 1140

Mukherjee, G. and Johnstone, I. M. (2017). "On Minimax Optimality of Sparse Bayes Predictive Density Estimates." *arXiv preprint arXiv:1707.04380*. 1140

Nguyen, X., Wainwright, M. J., Jordan, M. I., et al. (2009). "On surrogate loss functions and f-divergences." *The Annals of Statistics*, 37(2): 876–904. MR2502654. doi: https://doi.org/10.1214/08-AOS595. 1150

Pardoe, I. (2001). "A Bayesian sampling approach to regression model checking." *Journal of Computational and Graphical Statistics*, 10(4): 617–627. MR1938970. doi: https://doi.org/10.1198/106186001317243359. 1139

Robbins, H. (1964). "The empirical Bayes approach to statistical decision problems." *The Annals of Mathematical Statistics*, 35(1): 1–20. MR0163407. doi: https://doi.org/10.1214/aoms/1177703729. 1142

Rosasco, L., Vito, E. D., Caponnetto, A., Piana, M., and Verri, A. (2004). "Are loss functions all the same?" *Neural Computation*, 16(5): 1063–1076. 1150

Suzuki, T. and Komaki, F. (2010). "On Prior Selection and Covariate Shift of β-Bayesian prediction under α-divergence risk." *Comm. Statist. Theory and Methods*, 39: 1655–1673. 1140

Sweeting, T. J., Datta, G. S., and Ghosh, M. (2006). "Nonsubjective priors via predictive relative entropy regret." *The Annals of Statistics*, 441–468. MR2275249. doi: https://doi.org/10.1214/009053605000000804. 1142

Tan, Z. (2015). "Improved minimax estimation of a multivariate normal mean under heteroscedasticity." *Bernoulli*, 21(1): 574–603. MR3322331. doi: https://doi.org/10.3150/13-BEJ580. 1141

Tay, A. S. and Wallis, K. F. (2000). "Density forecasting: a survey." *Journal of Forecasting*, 19(4): 235–254. 1139

Taylor, J. W. and Buizza, R. (2004). "A comparison of temperature density forecasts from GARCH and atmospheric models." *Journal of Forecasting*, 23(5). 1139

Wang, D., Liu, H., and Liu, Q. (2018). "Variational inference with tail-adaptive f-divergence." In *Advances in Neural Information Processing Systems*, 5737–5747. 1140

Weinstein, A., Ma, Z., Brown, L. D., and Zhang, C.-H. (2018). "Group-linear empirical Bayes estimates for a heteroscedastic normal mean." *Journal of the American Statistical Association*, 113(522): 698–710. MR3832220. doi: https://doi.org/10.1080/01621459.2017.1280406. 1141

Xie, X., Kou, S., and Brown, L. (2016). "Optimal shrinkage estimation of mean param-

eters in family of distributions with quadratic variance." *Annals of Statistics*, 44(2): 564. MR3476610. doi: https://doi.org/10.1214/15-AOS1377. 1141

Xie, X., Kou, S., and Brown, L. D. (2012). "SURE estimates for a heteroscedastic hierarchical model." *Journal of the American Statistical Association*, 107(500). MR3036408. doi: https://doi.org/10.1080/01621459.2012.728154. 1139, 1140, 1141, 1142, 1143, 1146, 1148

Xu, X. (2005). "Estimation of high dimensional predictive densities." Ph.D. thesis, University of Pennsylvania. MR2707474. 1139

Xu, X. and Liang, F. (2010). "Asymptotic minimax risk of predictive density estimation for non-parametric regression." *Bernoulli*, 16(2): 543–560. MR2668914. doi: https://doi.org/10.3150/09-BEJ222. 1140

Xu, X. and Zhou, D. (2011). "Empirical Bayes predictive densities for high-dimensional normal models." *J. Multivariate Analysis*, 102(10): 1417–1428. MR2819959. doi: https://doi.org/10.1016/j.jmva.2011.05.008. 1141, 1143, 1144

Yano, K., Kaneko, R., and Komaki, F. (2021). "Minimax predictive density for sparse count data." Forthcoming in *Bernoulli*. 1150

Yano, K. and Komaki, F. (2017). "Asymptotically minimax prediction in infinite sequence models." *Electronic Journal of Statistics*, 11(2): 3165–3195. MR3697133. doi: https://doi.org/10.1214/17-EJS1312. 1140

Yuan, A. and Clarke, B. (1999). "An information criterion for likelihood selection." *IEEE Transactions on Information Theory*, 45(2): 562–571. MR1677018. doi: https://doi.org/10.1109/18.749003. 1139

Zhang, C.-H. (2003). "Compound decision theory and empirical Bayes methods: invited paper." *Ann. Statist.*, 31(2): 379–390. MR1983534. doi: https://doi.org/10.1214/aos/1051027872. 1142

Zhang, F., Shi, Y., Ng, H. K. T., and Wang, R. (2017). "Information geometry of generalized Bayesian prediction using α-divergences as loss functions." *IEEE Transactions on Information Theory*, 64(3): 1812–1824. MR3766316. doi: https://doi.org/10.1109/TIT.2017.2774820. 1140