Electronic Journal of Statistics
Vol. 15 (2021) 1636–1660
ISSN: 1935-7524
https://doi.org/10.1214/21-EJS1818

On discrete priors and sparse minimax optimal predictive densities*

Ujan Gangopadhyay and Gourab Mukherjee

University of Southern California
e-mail: ujan.gangopadhyay@usc.edu; gourab@usc.edu

Abstract: We consider the problem of predictive density estimation under Kullback-Leibler loss in a high-dimensional Gaussian model with exact sparsity constraints on the location parameters. For non-asymptotic sparsity levels, the least favorable prior is discrete. Here, we study the first order asymptotic minimax risk of Bayes predictive density estimates where the proportion of non-zero coordinates converges to zero as dimension increases. Motivated by an optimal thresholding rule in Mukherjee and Johnstone (2015), we propose a discrete prior and show that its Bayes predictive density estimate is minimax optimal. This produces a nonsubjective discrete prior distribution that minimizes the maximum posterior predictive relative entropy regret. We discuss the decision theoretic implications and the structural differences between our proposed prior and its closest predecessor – the geometrically decaying discrete prior of Johnstone (1994a) that produced minimax optimal point estimators under quadratic loss. Through numerical experiments, we present non-asymptotic worst-case risk of our proposed estimator across different sparsity levels.

MSC2020 subject classifications: Primary 62L20; secondary 60F15, 60G42.
Keywords and phrases: Minimax risk, sparsity, discrete priors, predictive density estimation, predictive inference, information loss.

Received September 2020.

1. Introduction and main results

A fundamental problem in statistical prediction analysis is to choose a probability distribution based on observed data that will be good in predicting the behavior of future samples (Aitchison and Dunsmore, 1975; Geisser, 1993; Aitchison, 1975). The future probability density conditioned on the observed past is referred to as the predictive density and estimating it plays an important role in a number of statistical applications (Liang, 2002; Mukherjee, 2013). Consider the problem of predictive density estimation in a n-dimensional Gaussian location model where the observed past vector $X \sim N_n(\theta, v_x I)$ and the future vector $Y \sim N_n(\theta, v_y I)$. The variances v_x and v_y are known. The future and past vectors are related only through the unknown location vector θ. Consider predictive density estimators (PRDE) $\hat{p}(y|x)$ and measure their performance in estimating the true future density $p(y|\theta, v_y) = N_n(\theta, v_y I)$ by the global divergence measure

*The research here was partially supported by NSF DMS-1811866.

of Kullback and Leibler (1951),

$$L(\theta, \hat{p}(\cdot|x)) = \int p(y|\theta, v_y) \log \left(\frac{p(y|\theta, v_y)}{\hat{p}(y|x)} \right) dy. \tag{1.1}$$

The KL risk integrates the above loss over the past distribution and is given by

$$\rho(\theta, \hat{p}) = \int L(\theta, \hat{p}(\cdot|x)) p(x|\theta, v_x)\, dx.$$

Sweeting et al. (2006) showed that $\rho(\theta, \hat{p})$ constitutes a posterior predictive relative entropy regret criterion that can be used for the construction of non-subjective prior distributions.

Given any prior π on θ, the Bayes PRDE $\hat{p}_\pi(y|x) = \int p(y|\theta, v_y)\pi(d\theta|x)$. The average integrated risk $B(\pi, \hat{p}) = \int \rho(\theta, \hat{p})\pi(d\theta)$, when well-defined, is minimized by $\hat{p}_\pi$ yielding the Bayes risk $B(\pi) = \inf_{\hat{p}} B(\pi, \hat{p})$.

Decision theoretic parallels between PRDE and point estimation (PE) of the multivariate normal mean under square loss are established in Komaki (2001); George et al. (2006); Brown et al. (2008); Ghosh et al. (2008); Kato (2009a); Maruyama and Ohnishi (2019). These risk analysis results hold for any dimension n. In higher dimensions, Fourdrinier et al. (2011); Xu and Liang (2010); Kubokawa et al. (2013) developed minimax optimal PRDE for constrained parameter spaces [see George et al. (2012), Ch. 1 of Mukherjee (2013) and George et al. (2019) for extensive reviews]. Sparse PRDE under exact ℓ_0 sparsity constraints on the location parameter is studied in Mukherjee and Johnstone (2017, 2015) where efficacy of different PRDEs were evaluated with respect to the minimax benchmark risk $R^*(\Theta) = \inf_{\hat{p}} \sup_{\theta \in \Theta} \rho(\theta, \hat{p})$. For an ℓ_0 constrained parameter space $\Theta_0[s_n] = \{\theta \in \mathbb{R}^n : \sum_{i=1}^n 1\{\theta_i \neq 0\} \leq s_n\}$ when $\eta_n = s_n/n \to 0$, the first order asymptotic minimax risk was evaluated as

$$R^*(\Theta_0[s_n]) = (1+r)^{-1} n\, \eta_n \log \eta_n^{-1} (1 + o(1)) \text{ as } n \to \infty, \tag{1.2}$$

where $r = v_y/v_x$. The minimax risk increases as r decreases. The difficulty of the density estimation problem increases as r decreases as we need to estimate the future observation density based on increasingly noisy past observations. The rate of convergence of the minimax risk with n does not depend on r, and so exact determination of the constants is needed to show the role of r in this prediction problem. Several predictive phenomena that contrast with point estimation results have been reported with the divergence becoming palpable as r decreases.

Here, we study the risk of Bayes predictive density estimators based on sparse discrete priors. In order to incorporate the knowledge on sparsity of the parameters, we consider priors with an atom of probability (spike) at the origin. Spike-and-slab priors based procedures have been shown to be very successful for sparse estimation (Johnstone and Silverman, 2004; Clyde and George, 2000; Rockova and George, 2018). Here, we consider slabs based on discrete priors. In regimes with non-asymptotic sparsity levels, i.e., $\eta_n \to \eta \in (0, 1)$,

the least favorable prior is unique and discrete (Berger and Bernardo, 1992; Zhang, 1994). Risk analysis of estimators based on discrete priors has a rich history in statistical decision theory (Johnstone, 2013; Marchand et al., 2004; Bickel, 1983; Kempthorne, 1987), particularly for studying the worst-case geometry of parametric spaces. For tractable analysis and detailed insights, minimax optimality based on discrete priors is studied in the asymptotic regime (Johnstone, 1994b; Bickel, 1983, 1981). Johnstone (1994a) (henceforth referred to as J94) established that for sparse point estimation a product prior based on discrete marginals containing equi-spaced support-points with geometrically decaying probability is asymptotically minimax optimal. Mukherjee and Johnstone (2017) (referred hereon as MJ17) showed that Bayes PRDE from the J94 prior is minimax sub-optimal under Kullback-Leibler loss. In this paper, we construct a discrete prior whose marginals have geometrically decaying tail probabilities akin to J94 but have different prior spacings so that the resultant Bayes PRDE is minimax optimal.

The discrete prior we study here is inspired by the risk diversification phenomenon introduced in Mukherjee and Johnstone (2015) (henceforth referred to as MJ15) for constructing minimax optimal PRDEs. MJ15 showed that in contrast to point estimation, for obtaining minimax optimality in sparse PRDE we need to incorporate the notion of diversification of the future risk. The optimal thresholding rule of MJ15 used two Bayes PRDEs: One of those is the Bayes PRDE from a symmetric product prior whose marginals have finitely many support points with atoms except at the origin having equal probability. Here, we conduct detailed worst-case risk analysis of PRDEs based on generic versions of such discrete priors. Unlike MJ15, our proposed prior has marginals with clusters of equi-probable atoms and the clusters have different probabilities. Compared to MJ15, our proposed clustered prior based Bayes PRDE has the advantage of avoiding the discontinuous thresholding operation in order to obtain sparse minimax optimality.

We first present our main result regarding minimax optimality of the Bayes PRDE from the proposed clustered discrete prior. Thereafter, we discuss the implications of the result along with detailed background and connections to the existing literature.

1.1. Main result: Minimax optimality

For any fixed positive r, consider the Bayes PRDE from a discrete product prior consisting of symmetric marginals π_{CL} (defined below). The marginal has equi-spaced clusters of atoms with geometrically decaying probability content in the clusters as they move away from the origin. For any $\eta \in (0,1)$ and $r \in (0,\infty)$ consider the univariate clustered discrete prior:

$$\pi_{\mathsf{CL}}[\eta, r; \gamma, \kappa] = (1-\eta)\delta_0 + \frac{1-\eta}{2}\sum_{i=1}^{\infty} \eta^i \big\{ C_i(\eta, r; \gamma, \kappa) + C_{-i}(\eta, r; \gamma, \kappa) \big\} \ , \quad (1.3)$$

which has an atom of probability $1-\eta$ at the origin and the remaining η probability shared across clusters. Each of the clusters C_i has κ atoms $\{\mu_{ij} : j = 1, \ldots, \kappa\}$ of equal probability which is the reason for referring such prior distributions as *clustered priors*. Let $v = (1 + r^{-1})^{-1}$, $\lambda_e := \lambda_e(\eta) = (-2v_x \log \eta)^{1/2}$ and $\lambda_f := \lambda_f(\eta, r) = v^{1/2}\lambda_e$. For any fixed $\gamma \geq 1$, the atoms in C_1 are aligned in between λ_f and λ_e in a geometric progression with common ratio γ, i.e., $\mu_{1j}(\eta, r, \gamma) = \gamma^{j-1}\lambda_f \wedge \lambda_e$ for $1 \leq j \leq \kappa$. Such geometric spacing was introduced in MJ15 (see Theorem 1C). For $i \geq 2$ the atoms are extended periodically to cluster C_i as $\mu_{ij} = (i-1)\mu_{1\kappa} + \mu_{1j}$ and by symmetry $\mu_{-ij} = -\mu_{ij}$ to the negative axis. Thus, the clusters themselves are equidistant at a separation of λ_f and while the atoms within each cluster has equal probability, the clusters themselves have geometrically decaying probabilities:

$$C_i(\eta, r; \gamma, \kappa) = \frac{1}{\kappa}\sum_{j=1}^{\kappa} \delta_{\mu_{ij}} \text{ and } P(C_i) = 2^{-1}(1-\eta)\eta^{|i|} \text{ for } i \in \mathbb{Z} \setminus \{0\}. \quad (1.4)$$

Our proposed cluster prior π_{C} has $\gamma = \gamma_r$ and $\kappa = K_r$ where,

$$\gamma_r = 1 + 4r \quad (1.5)$$

$$K_r = 1 + \left\lceil \log(1 + r^{-1})/(2\log\gamma_r) \right\rceil \cdot 1\{r < r_0\}\ . \quad (1.6)$$

Thus, $\pi_{\mathsf{C}}[\eta, r] := \pi_{\mathsf{CL}}[\eta, r; \gamma_r, K_r]$. Here, $r_0 = 0.5$. Note that, $K_r = 1$ iff $r \geq r_0$.

TABLE 1
The size K_r of each cluster in our proposed univariate cluster prior π_{C} as r varies.

r	0.0654	0.0759	0.0910	0.1150	0.1601	0.2826	0.4999	0.5 and above
K_r	8	7	6	5	4	3	2	1

When $K_r \geq 3$ and $i \geq 1$, all atoms except the K_rth one in any cluster C_i are aligned in a geometric progression starting from $\mu_{i1} = (i-1)\lambda_e + \lambda_f$, with common ratio $1 + 4r$ and $\mu_{iK_r} = i\lambda_e$. Table 1 shows the cluster size as r varies. Figure 1 shows the schematic diagram of the (truncated) prior with 6 clusters for two instances when $r = 0.38$ and $r = 0.14$ respectively. While the former has clusters of size 2, the latter has cluster size 4. Figure 1 illustrates a key aspect of the cluster prior: for $r < r_0$ the gap $\mu_{i,K_r} - \mu_{i,K_r-1}$ is allowed to vary widely with r while $\mu_{i+1,1} - \mu_{i,K_r}$ is fixed at λ_f for all i.

Now, consider the multivariate clustered prior $\pi_n^{\mathsf{C}}[\eta_n, r](d\theta) = \prod_{i=1}^n \pi_{\mathsf{C}}[\eta_n, r](d\theta_i)$ on $\mathbb{R}^n$. Then, the Bayes PRDE $\hat{p}_{\mathsf{C}}[\eta_n, r]$ based on $\pi_n^{\mathsf{C}}[\eta, r]$ is asymptotically minimax optimal.

Theorem 1.1. *Fix any $r \in (0, \infty)$. If $\eta_n = s_n/n \to 0$, then*

$$\lim_{n\to\infty} \left\{ \sup_{\theta \in \Theta_0[s_n]} \rho\big(\theta, \hat{p}_{\mathsf{C}}[\eta_n, r]\big) \right\} \Big/ R^*(\Theta_0[s_n]) = 1.$$

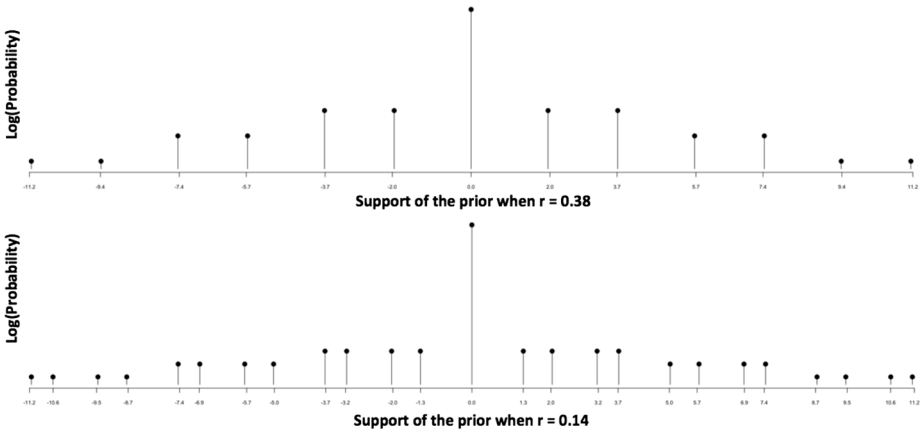

FIG 1. *Schematic for our proposed univariate cluster prior when r equals 0.38 (top) and 0.14 (bottom) respectively. The x-axis shows the spacings between and within the clusters and the y-axis the logarithm of the prior probabilities. Figure drawn to scale with* $\eta = 0.001$. *Only the six clusters are displayed with the rest being truncated.*

1.2. Geometrically decaying priors: Background and risk analysis

For understanding the decision theoretic implications of the above result, we briefly revisit the risk properties of sparse product priors based on symmetric marginals. It follows from J94 that for point estimation of the normal mean over $\Theta_0[s_n]$ under ℓ_2 loss, the posterior mean of the grid prior π_n^{EG} is minimax optimal as $\eta_n \to 0$. π_n^{EG} constitutes of i.i.d. copies of univariate grid prior $\pi_{\mathsf{EG}}[\eta_n]$ which is defined for any fixed r and $\eta \in (0,1)$ as

$$\pi_{\mathsf{EG}}[\eta] = (1-\eta)\delta_0 + \frac{1-\eta}{2}\sum_{i=1}^{\infty}\eta^i\left\{\delta_{i\lambda_e} + \delta_{-i\lambda_e}\right\} .$$

As $\eta_n \to 0$, the *geometric contamination* based discrete prior $\pi_{\mathsf{EG}}[\eta_n]$ attains the least Fisher information. In contrast to π_{C}, π_{EG} always has only one point in each cluster. However, they have identical probability decay rate (geometric contamination) as the clusters extend away from the origin. MJ17 showed that the PRDE based on π_n^{EG} is sub-optimal for PRDE estimation based on KL loss. The Bayes PRDE based on a product grid prior whose univariate marginals π_{PG} (subscripts PG and EG denote predictive and estimative grids) has reduced spacing between the atoms and reduced probability decay rate, was established to be minimax optimal in the predictive regime abet for $r \geq \tilde{r}_0 = (\sqrt{5}-1)/4$:

$$\pi_{\mathsf{PG}}[\eta, r] = (1-\eta)\delta_0 + \frac{\eta(1-\eta^v)}{2}\sum_{i=1}^{\infty}\eta^{(i-1)v}\left\{\delta_{i\lambda_f} + \delta_{-i\lambda_f}\right\} .$$

However, π_{PG} is sub-optimal for $r < \tilde{r}_0$. Note that, unlike the univariate grid priors $\pi_{\mathsf{EG}}, \pi_{\mathsf{PG}}$ where support points have geometric probability decay, π_{C} has

support points with identical probability within each clusters. The clusters in π_{C} however has the same decay rate as the support points in π_{EG}. The maximum gap between atoms in π_{C} equals the spacing in π_{PG}. Equiprobable atoms in the clusters was introduced in MJ15 to control predictive risk via the new notion of risk diversification. As such consider a truncated cluster prior with only two clusters:

$$\pi_{\mathsf{TC}}[\eta, r] = (1-\eta)\delta_0 + \eta/2\{C_1 + C_{-1}\}$$

where, $C_1 = C_1(\eta, r; \tilde{\gamma}_r, \tilde{K}_r)$ as in (1.4) with $\tilde{\gamma}_r = 1 + 2r$ and $\tilde{K}_r$ given by $K_r - 1$ with the formula in (1.6) used with $\tilde{\gamma}_r$ in place of γ_r. As the prior π_{TC} is bounded at λ_e, its corresponding Bayes PRDE $\hat{p}_{\mathsf{CT}}$ has unbounded risk. Thresholded product PRDE $\hat{p}_n^{\mathsf{T}}(y|x) = \prod_{i=1}^n \hat{p}_{\mathsf{T}}(y_i|x_i)$ with

$$\hat{p}_{\mathsf{T}}(y_i|x_i) = \hat{p}_{\mathsf{TC}}[\eta_n, r](y_i|x_i)1\{|x_i| \le \lambda_e(\eta_n)\} + \phi(y_i|x_i, v_x + v_y)1\{|x_i| > \lambda_e(\eta_n)\}$$

was shown in MJ15 to be minimax optimal for any $r \in (0, \infty)$. In $\hat{p}_{\mathsf{T}}$ the threshold is $\lambda_e(\eta_n)$; above the threshold the Bayes PRDE based on the uniform prior, which is Gaussian with variance $v_x + v_y$ was used where as below the threshold the Bayes PRDE from π_{CT} is used. Thresholding rules are not smooth functions of the data and it was conjectured in Sec. 6 of MJ15 that periodic clustered priors of the form of (1.3)-(1.4) can attain minimax optimality without the discontinuous thresholding operation. Here, we study the risk properties of such cluster priors and establish minimax optimality of the properly calibrated prior π_C. We found that the common ratio $\tilde{\gamma}_r$ used in MJ15 was not optimal and can be increased to γ_r. However, as a consequence of removing thresholding we needed one more atom than MJ15 in our proposed cluster prior π_{C} for small values of r. We show that the number of support points in π_{C} as used here is necessary by proving the following risk properties of Bayes PRDEs based on generic cluster priors of (1.3). First, we show that our prescribed choice of $r_0 = 0.5$ is sharp and can not be further lowered: Any cluster prior with cluster-size one and the probabilty decay rate of at least η_n is sub-optimal for all $r < r_0$. Consider the following univariate prior with singleton atoms in each cluster:

$$\pi_{\mathsf{SI}}[\eta; \nu, l] = (1-\eta)\delta_0 + \frac{1-\eta}{2}\sum_{i=1}^{\infty} \eta^i \{\delta_{\mu_i} + \delta_{-\mu_i}\} \quad \text{and } \mu_i = \nu + (i-1)l, \quad (1.7)$$

with $\nu \ge 0$ and $l \ge L(\eta) := (-2\log\eta)^{1/2}$. As ν, L vary, let SI be the class of Bayes PRDEs $\hat{p}_{\mathsf{SI}}[\eta_n; \nu, L]$ based on n i.i.d. copies of $\pi_{\mathsf{SI}}[\eta_n; \nu, L]$. Note, that this class includes Bayes PRDEs based on π_n^{EG} which correspond to $\nu = l = L(\eta_n)$ as well as $\hat{p}_C$ for all $r \ge r_0$ in which case $l = L(\eta_n)$ but $\nu = (1 + r^{-1})^{-1/2} L(\eta_n)$. The following result shows that the class SI is sub-optimal.

Lemma 1.2. *If $\eta_n = s_n/n \to 0$, then for any $r < r_0$,*

$$\liminf_{n\to\infty} \left\{ \inf_{\hat{p}\in\mathsf{SI}} \sup_{\theta\in\Theta_0[s_n]} \rho(\theta, \hat{p}) \right\} \Big/ R^*(\Theta_0[s_n]) > 1.$$

Second, we show that our prescribed cluster size K_r can not be further reduced for any $r < r_0$. The following result shows that any priors of the form (1.3)-(1.4) with $\gamma > \gamma_r$ and $K_\gamma = 1 + \lceil \log(1 + r^{-1})/(2\log\gamma) \rceil$ (which gets specified once γ is fixed by the structure of the atoms in (1.4)) will produce sub-optimal Bayes PRDEs. Also, dropping atoms from π_{C} will lead to sub-optimality. For any non-empty subset $S \subset \{1, \ldots, K_r\}$, instead of (1.4) consider priors $\pi_{\mathsf{S}}[\eta, r]$ with clusters

$$C_i(\eta, r; \gamma_r, K_r) = \Big(\sum_{j=1}^{K_r} I(j \in S) \Big)^{-1} \sum_{j=1}^{K_r} \delta_{\mu_{ij}} I(j \in S)$$

and $P(C_i) = 2^{-1}(1-\eta)\eta^{|i|}$ for $i \in \mathbb{Z} \setminus \{0\}$. Let $\hat{p}_{\mathsf{CL}}[\eta_n, r; S]$ denote the multivariate Bayes PRDE based on product of $\pi_{\mathsf{S}}[\eta_n, r]$. The following result shows that it is sub-optimal.

Lemma 1.3. *For any fixed $r < r_0$, as $\eta_n = s_n/n \to 0$, for any $\gamma > \gamma_r$ and $K_\gamma = 1 + \lceil \log(1 + r^{-1})/(2\log\gamma) \rceil$ we have:*

$$\liminf_{n\to\infty} \left\{ \sup_{\theta \in \Theta_0[s_n]} \rho\big(\theta, \hat{p}_{\mathsf{CL}}[\eta_n, r; \gamma, K_\gamma]\big) \right\} \Big/ R^*(\Theta_0[s_n]) \; > 1.$$

As $\eta_n = s_n/n \to 0$, there exists $r < r_0$ such that for any $S \subset \{1, \ldots, K_r\}$,

$$\liminf_{n\to\infty} \left\{ \sup_{\theta \in \Theta_0[s_n]} \rho\big(\theta, \hat{p}_{\mathsf{CL}}[\eta_n, r; S]\big) \right\} \Big/ R^*(\Theta_0[s_n]) \; > 1.$$

The proof of the above two lemmas are provided in section 3. We end this section by summarizing the key features of our results.

(a) We construct a prior π_n^{C} with geometric contamination akin to the prior π_n^{EG} in J94. Under high sparsity, the prior in J94 is asymptically least favorable for point estimation and has the least Fisher information; its posterior mean is minimax optimal under quadratic loss. However, the corresponding Bayes PRDE $\hat{p}_{\mathsf{EG}}$ is minimax sub-optimal under KL loss. Our proposed π_n^{C} has the same decay rate as π_n^{EG} but lot more atoms. It Bayes PRDE $\hat{p}_{\mathsf{C}}$ is minimax optimal under KL loss.

(b) The proposed prior π_n^{C} is based on the minimax analysis of the posterior predictive relative entropy regret criterion of Sweeting et al. (2006) which differs from traditional reference prior inducing criterion (Bernardo, 1979) as it also considers predictive performance in relation to alternative nondegenerate prior distributions.

(c) As conjectured in MJ15, using the proposed prior π_n^{C} lead to a minimax optimal procedure that do not involve threholding. The optimality in the structure of the proposed prior is established. Among all geometrically contantimated sparse discrete prior having cluster decay rate similar to π_n^{EG} in J94, it has the minimal cardinality as established in Lemmas 1.2 and 1.3.

(d) Compared to the simplier grid priors π_n^{PG} analyzed in MJ17, the geometry of manifold induced by the proposed prior π_n^{C} is significantly different. This necessitates separate analysis and proofs of the risk properties of the Bayes PRDEs from π_n^{C}.

(e) The essential ingredient in the proof is the asymptotic analysis of the terms in the decomposition of the univariate predictive risk function in (2.2). The terms in the right hand side of (2.2) involve exponential Gaussian sums with different means and variances. Minimax analysis of the predictive regret involves asymptotic characterization of the differences between the logarithm of two exponential Gaussian sums. In contrast, minimax optimality in PE involves studying only one exponential Gaussian sum (see Ch 8.5 in Johnstone (2013) and the description in Sec. 2).

1.3. Further result on the asymptotic Bayes risk

Figure 2 shows the numerical evaluation of the predictive risk $\rho(\theta, \hat{p}_{\mathsf{C}}[\eta, r])$ of our proposed Bayes PRDE when $\eta = 0.001$ and $r = 0.225$. Each cluster has size three. The maximum risk $\hat{p}_{\mathsf{C}}[\eta, r]$ crosses the asymptotic theory limit but does not exceed by much. It shows that the asymptotic analysis is fairly reflective in this non-asymptotic regime.

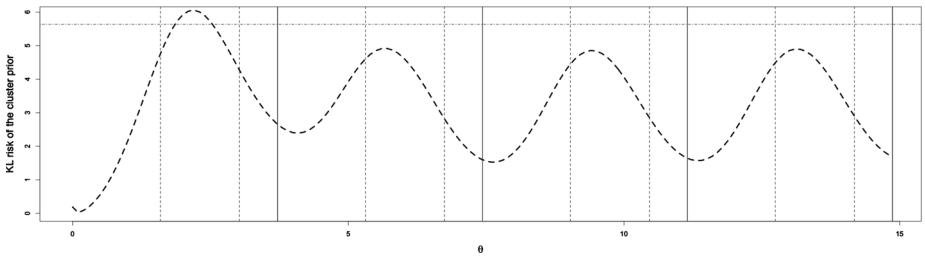

FIG 2. *Plot of the univariate predictive KL risk* $\rho(\theta, \hat{p}_{\mathsf{C}})[\eta, r]$ *as* θ *varies over the* x*-axis. Here,* $\eta = 0.001$ *and* $r = 0.225$*. The horizontal line corresponds to the asymptotic minimax limit* $\lambda_f^2(\eta)/(2r)$*. The dotted vertical lines denotes the location of the non-origin support points of* $\pi_{\mathsf{C}}[\eta, r]$ *with the bold lines marking each cluster boundary.*

The risk function has its peak between μ_{11} and μ_{12} and is approximately periodic barring a few clusters near the origin. As the figure shows, the risk function is much smaller than the asymptotic limit of $\lambda_f^2/(2r)$ for all the points in C_1 barring its first point. As all points in C_1 are equally likely, this implies that the cluster prior is not least favorable. The following result make this observation rigorous by explicitly evaluating the first order asymptotic Bayes risk of the cluster prior. It establishes that when there are two or more points in each cluster (i.e. $r < r_0$) the cluster prior is no longer least favorable.

Theorem 1.4. *If* $\eta_n = s_n/n \to 0$ *as* $n \to \infty$*, then the multivariate cluster prior* $\pi_n^{\mathsf{C}}[\eta_n, r]$ *is not asymptotically least favorable for all* $r < r_0$*. As such, its Bayes*

risk satisfies:

$$\lim_{n\to\infty}\left\{B(\pi_n^{\mathsf{C}}[\eta_n,r])\right\}\Big/R^*(\Theta_0[s_n])=\frac{1}{K_r}\left\{1+r\sum_{i=2}^{\infty}\left(1+r^{-1}-(1+4r)^{2i}\right)_+\right\}, \tag{1.8}$$

where, K_r is defined in (1.6). Additionally, if $\eta_n\to 0$ and $s_n\to\infty$ as $n\to\infty$ then $\pi_n^{\mathsf{C}}[\eta_n,r]$ is asymptotically least favorable for all $r\geq r_0$.

Note that, when $r\geq r_0$, $K_r=1$ and the RHS of (1.8) equals 1. For $r<r_0$, the proposed prior is no longer exactly asymptotically least favorable but its Bayes risk has the same order of the minimax risk as $\eta_n\to 0$.

2. Proof overview

We provide a brief overview of the proof of our main result. Detailed proofs of all the results are provided in section 3. The proof of Theorem 1.1 involves asymptotically upper bounding the risk $\sup_{\theta\in\Theta_0[s_n]}\rho(\theta,\hat{p}_{\mathsf{C}})$ by $R^*(\Theta_0[s_n])$. Then, the asymptotic equality follows as the first term can not be smaller than the minimax risk by definition. Also, note that due to the product structure of the prior, the multivariate maximal risk can be evaluated based on the risk of the univariate Bayes PRDE $\hat{p}_{\mathsf{C}}[\eta_n,r]$ by using the following relation:

$$\sup_{\theta\in\Theta_0[s_n]}\rho(\theta,\hat{p}_{\mathsf{C}})=n(1-\eta_n)\rho(0,\hat{p}_{\mathsf{C}}[\eta_n,r])+n\eta_n\sup_{\theta\in\mathbb{R}\setminus 0}\rho(\theta,\hat{p}_{\mathsf{C}}[\eta_n,r])\ . \tag{2.1}$$

Asymptotic evaluation of the two expressions on the right above is done by using the risk decomposition Lemma 2.1 of MJ17. It reduces the calculation for the univariate predictive risk to finding expectation of functionals involving standard normal random variable Z as

$$\rho(\theta,\hat{p}_{\mathsf{C}}[\eta_n,r])=\frac{\theta^2}{2r}-\mathbb{E}\log N_{\theta,v}(Z)+\mathbb{E}\log D_\theta(Z),\ \text{where,} \tag{2.2}$$

$$N_{\theta,v}(Z)=1+\sum_{i\in\mathbb{Z}\setminus 0}\frac{q_i}{K}\sum_{j=1}^{K}N_{ij}(\theta,Z;v)\ \text{and}\ D_\theta(Z)=N_{\theta,1}(Z)\ .$$

Here, $q_i=(1-\eta_n)^{-1}P(C_i)$ with $P(C_i)$ being the mass of cluster C_i in $\pi_{\mathsf{C}}[\eta_n,r]$; thus $q_i=2^{-1}\exp(-|i|\lambda_{e,n}^2/2)$ with $\lambda_{e,n}=(2\log\eta_n^{-1})^{-1}$ and $\lambda_{f,n}=v^{1/2}\lambda_{e,n}$; N_{ij} is the contribution to the risk of the jth support point $\mu_{ij}(\eta_n,r)$ within the ith cluster.

The risk contributions N_{ij} are exponents of quadratic forms in μ_{ij}, viz,

$$N_{ij}(\theta,Z;v)=\exp\{v^{-1/2}\mu_{ij}Z+v^{-1}\mu_{ij}\theta-(2v)^{-1}\mu_{ij}^2\}.$$

The risk at the origin is well-controlled for this cluster prior based PRDE (see Lemma 3.1) and so, based on (2.1), it is suffices to bound $\sup_\theta\rho(\theta,\hat{p}_{\mathsf{C}}[\eta_n,r])$ by $\lambda_{f,n}^2/(2r)$ to arrive at the desired result. This involves tracing two fundamentally

different risk phenomena depending on the location of θ (a) $\theta \in C_{\pm 1}$ (b) $\theta \notin C_{\pm 1}$. In the former case, $\mathbb{E}\log D_\theta(Z) = O(\lambda_{f,n})$ (and thus the contribution of the third term on the right of (2.2) is not significant. Also, $\mathbb{E}\log N_{\theta,v}(Z) = O(\lambda_{f,n})$ for $|\theta| \le \lambda_{f,n}$ and so, asymptotically $\rho(\theta, \hat{p}_{\mathsf{C}}[\eta_n, r])$ initially increases quadratically in θ and $\rho(\lambda_{f,n}, \hat{p}_{\mathsf{C}}[\eta_n, r]) = \lambda_{f,n}^2/(2r)(1+o(1))$. However, if $|\theta| \in C_1 \setminus [0, \lambda_{f,n}]$, then $\mathbb{E}\log N_{\theta,v}(Z)$ is significantly large and controls the predictive risk below the desired asymptotic limit (see Lemma 3.4).

If $\theta \in C_i$ for any $|i| > 1$, then the risk phenomenon is quite different than the origin adjoining clusters. Now, $\mathbb{E}\log D_\theta(Z)$ is significantly positive. However, an important ingredient of the proof is that its magnitude can be asymptotically well controlled by considering only atoms in C_i or the nearest atom in C_{i-1}. Lemma 3.3 establishes that for $\theta \in C_i$ with $|i| > 1$,

$$\mathbb{E}\log D_\theta(Z) \le \{\mathbb{E}\log D_{i\cdot}(Z)\}_+ + o(\lambda_{f,n}^2) \text{ as } n \to \infty,$$

where, $D_{i\cdot}(Z) = N_{i-1,K}(\theta, Z; 1) + \sum_{j=1}^K N_{ij}(\theta, Z; 1)$. Next, use the naive bound $\mathbb{E}\log N_{\theta,v}(Z) \ge \mathbb{E}\log N_{i\cdot}(Z))$ where $N_{i\cdot} = N_{i-1,K}(\theta, Z; v) + \sum_{j=1}^K N_{ij}(\theta, Z; v)$. Plugging these two bounds in (2.2) we get the desired upper bound in Lemma 3.4.

3. Detailed proofs

3.1. Background and preliminaries

For the technical proofs without loss of generality assume $v_x = 1$. So, $r = v_x/v_y = v_x$. Recall $v = (1 + r^{-1})^{-1}$ and $\eta_n = s_n/n$. As demonstrated in (2.1), the multivariate maximal risk of the Bayes predictive density estimate (PRDE) from the cluster prior can be evaluated by studying the predictive risk of the univariate Bayes PRDE $\hat{p}_{\mathsf{C}}[\eta_n, r]$ based on the univariate cluster prior $\pi_{\mathsf{C}}[\eta_n, r]$:

$$\sup_{\theta \in \Theta_0[s_n]} \rho(\theta, \hat{p}_{\mathsf{C}}) = n(1-\eta_n)\rho(0, \hat{p}_{\mathsf{C}}[\eta_n, r]) + n\eta_n \sup_{\theta \in \mathbb{R}\setminus 0} \rho(\theta, \hat{p}_{\mathsf{C}}[\eta_n, r]) \ . \quad (3.1)$$

Henceforth, unless we explicitly mention, we would concentrate on univariate Bayes predictors and their risk functions. Recall, in the multivariate set-up we consider asymptotically sparse regimes, where $\eta_n \to 0$ as $n \to \infty$. Hereon, for convenience of notation we write η instead of η_n keeping the dependence on n implicit. Recall,

$$\lambda_e := \sqrt{2\log \eta^{-1}}, \quad \text{and} \quad \lambda_f := \sqrt{2\log \eta^{-v}}.$$

Recall from equations (1.3)-(1.6) the univariate clustered discrete prior $\pi_{\mathsf{C}}[\eta, r]$ is the following:

$$\pi_{\mathsf{C}}[\eta, r] = (1-\eta)\delta_0 + \frac{1-\eta}{2}\sum_{j=1}^{\infty} \eta^j \{C_j(\eta, r) + C_{-j}(\eta, r)\}.$$

The point-masses in cluster C_j are denoted by $\{\mu_{jk} : k = 1, \ldots, K\}$ where the common cluster size K is

$$K := K(r) = 1 + \left\lceil \log(1 + r^{-1})/(2\log(1 + 4r)) \right\rceil \cdot 1\{r < r_0\},$$

where, $r_0 = 1/2$. Further, recall that

$$C_j(\eta, r) = \frac{1}{K} \sum_{k=1}^{K} \delta_{\mu_{jk}} \text{ for } j \in \mathbb{Z} \setminus \{0\},$$

where $\mu_{1k} = \lambda_f(1+4r)^{k-1} \wedge \lambda_e$ for $1 \le k \le K$, $\mu_{jk} = (j-1)\mu_{1K} + \mu_{1k}$ for $j \ge 2$, and $\mu_{jk} = \mu_{-jk}$ for $j < 0$. So for $r \ge r_0$, that is, when $K = 1$, the clustered discrete prior only has point-masses $\{j\lambda_f : j \in \mathbb{Z}\}$.

By Lemma 2.1 of MJ17 the predictive KL risk of the univariate cluster prior is given by:

$$\rho(\theta, \hat{p}_{\mathsf{C}}[\eta, r]) = \frac{\theta^2}{2r} - \mathbb{E} \log N_{\theta,v}(Z) + \mathbb{E} \log D_\theta(Z) \tag{3.2}$$

where Z is a standard normal random variable, and

$$N_{\theta,v}(Z) = 1 + \frac{1}{2K} \sum_{j \in \mathbb{Z}\setminus\{0\}} \sum_{k=1}^{K} \exp\left\{ \frac{\mu_{jk} Z}{\sqrt{v}} + \frac{\mu_{jk}\theta}{v} - \frac{\mu_{jk}^2}{2v} - |j| \frac{\lambda_e^2}{2} \right\}, \text{ and}$$

$$D_\theta(Z) = 1 + \frac{1}{2K} \sum_{j \in \mathbb{Z}\setminus\{0\}} \sum_{k=1}^{K} \exp\left\{ \mu_{jk}(Z + \theta) - \frac{\mu_{jk}^2}{2} - |j| \frac{\lambda_e^2}{2} \right\}.$$

3.2. Proof of Theorem 1.1

We first present the proof for $r < r_0$ because the proof is more intricate compared to the case when $r \ge r_0$. In the latter case, by definition $K = 1$, and the proof is is comparatively easier. It uses parts of the proof techniques used for $r < r_0$ case but also involves some fundamentally different attributes. Hence, it is presented afterwards where we also explain the choice of $r_0 = 1/2$.

3.2.1. Notations

For convenience of notation, we shall write the support points of the clustered discrete prior as $\{\mu_p : p \in \mathbb{Z}\}$. The identification is made as follows. Let $\mu_0 = 0$, and for $p > 0$ identify μ_p in the new notation with $\mu_{j_p k_p}$ in the original notation where j_p, k_p are the unique positive integers such that $p = (j_p - 1)K + k_p$ with $k_p \le K$. For $p < 0$ let $\mu_p = -\mu_{-p}$. So essentially μ_p is the k_pth point in the pth cluster. Let $j_0 = 0$ and $j_{-p} = j_p$ for $p < 0$. Let $c_0 = 1$ and $c_p = (2K)^{-1}$ for $p \ne 0$. With these new notations can write

$$D_\theta(Z) = \sum_{p \in \mathbb{Z}} D_{\theta p}(Z), \text{ and } N_\theta(Z) = \sum_{p \in \mathbb{Z}} N_{\theta p}(Z)$$

where

$$D_{\theta p}(Z) := c_p \exp\left\{\mu_p Z + \mu_p \theta - \frac{1}{2}\mu_p^2 - j_p \frac{\lambda_e^2}{2}\right\}, \text{ and}$$
$$N_{\theta p}(Z) := c_p \exp\left\{\frac{\mu_p Z}{\sqrt{v}} + \frac{\mu_p \theta}{v} - \frac{\mu_p^2}{2v} - j_p \frac{\lambda_e^2}{2}\right\}.$$

The above notations will be used for all $r \in (0, \infty)$. But now we define two indexes $l_d(\theta)$ and $l_n(\theta)$ for all $\theta > 0$ specifically for $r < r_0$. If $\theta \in [j\lambda_e, (j+1)\lambda_e)$, then let $l_d(\theta) := jK$. So $l_d(\theta)$ is the number of support points in the cluster prior between $[0, j\lambda_e]$. This is the index of the atom μ_p such that $\mathbb{E}\, D_{0p}(Z)$ is maximized. Note that $j\lambda_e = \mu_{jK} = \mu_{l_d(\theta)}$. Now we define the index $l_n(\theta)$ which is the index of the atom μ_p such that $\mathbb{E}\, N_{0p}(Z)$ is maximized. More precisely, $l_n(\theta)$ is defined as follows:

(i) If $\theta \in [j\lambda_e, j\lambda_e + \lambda_f]$, then let $l_n(\theta) := jK$. Note that, in this case $\mu_{l_n} = \mu_{jK} = j\lambda_e$.
(ii) If $\theta \in (j\lambda_e + \lambda_f(1+4r)^k, \min\{j\lambda_e + \lambda_f(1+4r)^k(1+2r), (j+1)\lambda_e\}]$ for $0 \le k < K$, then let $l_n(\theta) := jK + k + 1$. Note that, in this case $\mu_{l_n} = \mu_{jK+k+1} = j\lambda_e + \lambda_f(1+4r)^k$.
(iii) If $\theta \in (j\lambda_e + \lambda_f(1+4r)^k(1+2r), \min\{j\lambda_e + \lambda_f(1+4r)^{k+1}, (j+1)\lambda_e\}]$ for some $0 \le k < K$, then let $l_n(\theta) := jK + k + 2$. Note that, $\mu_{l_n} = \mu_{jK+k+2} = \min\{j\lambda_e + \lambda_f(1+4r)^{k+1}, (j+1)\lambda_e\}$.

3.2.2. Risk at origin

The risk at the origin for our cluster prior based Bayes PRDE is asymptotically much smaller than the risk for the thresholding based risk diversified PRDE of MJ15. As such, comparing equation (51) in the aforementioned paper with the following result, it follows that any thresholding based minimax optimal PRDE will have much higher risk at the origin than the cluster prior based Bayes PRDE. The Bayes PRDEs based on grid and bi-grids priors such as the π_{EG} prior of J94 and π_{PG} and π_{PG} priors of MJ17 have similar risk to the cluster prior based Bayes PRDE at the origin.

Lemma 3.1. *For any fixed $r \in (0, \infty)$, $\rho(0, \hat{p}_{\mathsf{C}}[\eta, r]) \le \eta(1 + o(1))$ as $\eta \to 0$.*

Proof. By definition $N_{\theta,v}(Z) \ge 1$ for all Z. Using (3.2), we have

$$\rho(0, \hat{p}_{\mathsf{C}}) = -\mathbb{E} \log N_{\theta,v}(Z) + \mathbb{E} \log D_\theta(Z) \le \mathbb{E} \log D_\theta(Z).$$

Note that, for $p \ne 0$, $\mathbb{E}\, D_{0p}(Z) = (2K)^{-1}\eta^{j_p}$. Summing over all $p \ne 0$ and using $D_0 = 1$ along with the inequality $\log(1+x) \le x$ for $x \ge 0$ we get

$$\mathbb{E} \log D_0(Z) \le \sum_{p \ne 0} \mathbb{E}\, D_{0p}(Z) = \sum_{p \ne 0} (2K)^{-1}\eta^{j_p} = \sum_{j=1}^{\infty} \eta^j = \frac{\eta}{1-\eta}.$$

This completes the proof. □

3.2.3. Risk bounds at the non-origin parametric points

Next, we concentrate on the risk at the non-origin points. Our goal is to establish

$$\sup_{\theta\in\mathbb{R}\setminus\{0\}} \rho\left(\theta, \hat{p}_{\mathsf{C}}[\eta, r]\right) \leq \frac{\lambda_f^2}{2r}(1+o(1)) \quad \text{as } \lambda_f \to \infty. \tag{3.3}$$

This along with (3.1) and the above result about the risk bound at the origin will imply that the multivariate maximum risk obeys

$$\begin{aligned}\sup_{\theta\in\Theta_0[s_n]} \rho(\theta, \hat{p}_{\mathsf{C}}[\eta_n, r]) &\leq -n\eta_n(1-\eta_n)^{-1} + n\eta_n \frac{\lambda_f^2}{2r}(1+o(1)) \\ &= n\eta_n \log \eta_n^{-1}(1+r)^{-1}(1+o(1))\end{aligned}$$

which would establish the result in Theorem 1.1. By symmetry, it would be enough to prove the bound in (3.3) for positive θ. Hence, hereon in this subsection we only consider $\theta > 0$. In this case the contribution of $D_{\theta i}$s for $i < 0$ are expected to be negligible. This is formalized in the following result.

Lemma 3.2. *For any $r \in (0, \infty)$ and any fixed $\theta > 0$ we have,*

$$\mathbb{E} \log D_\theta(Z) = \mathbb{E} \log \left(1 + \sum_{i=1}^{\infty} D_{\theta i}(Z)\right) + o(1) \quad \text{as } \lambda_f \to \infty.$$

Proof. Using the the inequality $\log(1+x+y) \leq x + \log(1+y)$ for nonnegative x, y we get,

$$\mathbb{E} \log D_\theta(Z) \leq \sum_{i<0} \mathbb{E}\, D_{\theta i}(Z) + \mathbb{E} \log \left(1 + \sum_{i=1}^{\infty} D_{\theta i}(Z)\right).$$

Using definition of j_i and $D_{\theta i}$ and the fact that $\mu_i < 0$, $\theta > 0$ we get,

$$\begin{aligned}\mathbb{E}\, D_{\theta i}(Z) &= \mathbb{E}\, \frac{1}{2} \exp\left\{\mu_i(Z+\theta) - \frac{\mu_i^2}{2} - j_i \frac{\lambda_e^2}{2}\right\} \\ &= \frac{1}{2} \exp\left\{\mu_i \theta - j_i \frac{\lambda_e^2}{2}\right\} \leq \frac{1}{2} e^{-j_i \frac{\lambda_e^2}{2}}.\end{aligned}$$

As i runs from 0 to $-\infty$, j_i goes from 1 to ∞ with each term repeating K times. Hence, summing over $i < 0$ we get, $\sum_{i<0} \mathbb{E}\, D_{\theta i} = o(1)$ as $\lambda_f \to \infty$. This completes the proof. □

We first provide an upper bound on $\mathbb{E} \log D_\theta(Z)$, which would be substituted in equation (3.2) to get the required upper bound. The following result is crucial as it shows that the infinite sum in the expression of $D_\theta(Z)$ can be asymptotically reduced as a contribution from a single dominant term. This reduction greatly helps in tracking the risk of the cluster prior and is pivotal in the proof of Theorem 1.1.

Lemma 3.3. *For $r < r_0$ and any fixed $\theta > 0$ we have,*

$$\mathbb{E}\log D_\theta(Z) = \mathbb{E}\log D_{\theta l_d}(Z) + O(\lambda_f) \quad \text{as } \lambda_f \to \infty.$$

Proof. By virtue of the previous lemma, we can consider only contributions from $D_{\theta i}(Z)$s with $i > 0$. We suppress the dependence of $D_{\theta i}(Z)$s on θ and Z and simply write D_i. Similarly $D_\theta(Z)$ is written only as D_θ. Note that, $l_d \geq 0$ because $\theta > 0$. First we get an upper bound on $\mathbb{E}\log D_\theta$ by separating the contribution from μ_{l_d} as follows

$$\mathbb{E}\log\left(1+\sum_{i=1}^{\infty} D_{\theta i}(Z)\right) \leq \mathbb{E}\log D_{l_d} + \mathbb{E}\log\left(1+\sum_{i=l_d+1}^{\infty}\frac{D_i}{D_{l_d}}\right) + \mathbb{E}\log\left(1+\sum_{i=0}^{l_d-1}\frac{D_i}{D_{l_d}}\right). \quad (3.4)$$

In the right hand side of the above equation, the second term compares contribution of μ_{l_d} with that of the succeeding terms. We split it further by separating out the contribution of points in the next cluster from the rest in the following manner

$$\mathbb{E}\log\left(1+\sum_{i=l_d+1}^{\infty}\frac{D_i}{D_{l_d}}\right) \leq \sum_{i=l_d+1}^{l_d+K}\mathbb{E}\log\left(1+\frac{D_i}{D_{l_d}}\right) + \sum_{i=l_d+K}^{\infty}\mathbb{E}\,\frac{D_{i+1}}{D_i}. \quad (3.5)$$

Note that by definition of l_d, $\mu_{l_d} < \theta \leq \mu_{l_d+K}$. Take i such that $l_d + 1 \leq i \leq l_d + K$. Let $d_i := \mu_i - \mu_{l_d}$. Using the inequality $\log(1+x) \leq \log 2 + (\log x)_+$ we get

$$\begin{aligned}
\mathbb{E}\log\left(1+\frac{D_i}{D_{l_d}}\right) &= \mathbb{E}\log\left(1+\exp\left\{d_i\left(Z+\theta-\mu_{l_d}-\frac{d_i}{2}\right)-\frac{\lambda_e^2}{2}\right\}\right)\\
&\leq \mathbb{E}\log\left(1+\exp\left\{d_i Z-\frac{1}{2}(\lambda_e-d_i)^2\right\}\right)\\
&\leq \log 2 + \mathbb{E}(d_{il}Z-(\lambda_e-d_i)^2/2)_+\\
&= O(\lambda_f). \quad (3.6)
\end{aligned}$$

Summing over i in the range $l_d + 1 \leq i \leq l_d + K$ we get the fist term in the right-hand side of equation (3.5) is $O(\lambda_f)$.

Now let us consider the second term in the right-hand side of (3.5). Let $i \geq l_d + K$ so that $\theta \geq \mu_i$. Using $\mathbb{E}((\mu_{i+1}-\mu_i)Z) = (\mu_{i+1}-\mu_i)^2/2$ we get

$$\mathbb{E}\,\frac{D_{i+1}}{D_i} \leq \mathbb{E}\exp\left\{(\mu_{i+1}-\mu_i)\left(Z+\theta-\frac{\mu_{i+1}+\mu_i}{2}\right)\right\} = \exp\left\{(\mu_{i+1}-\mu_i)(\theta-\mu_i)\right\}. \quad (3.7)$$

Since i runs from $l_d + K$ to ∞ and $\theta \leq \mu_{l_d+K}$, if we take a sum over i, we get a geometrically decaying sum so that the second term in the right-hand side of equation (3.5) is $O(1)$.

Hence, the second term in the right-hand side of equation (3.4) is $O(\lambda_f)$.

Now we consider the third term in the right-hand side of equation (3.4) is also $O(\lambda_f)$. We split the sum as

$$\mathbb{E}\log\left(1+\sum_{i=0}^{l_d-1}\frac{D_i}{D_{l_d}}\right) \le \sum_{i=l_d-K+1}^{l_d-1}\mathbb{E}\log\left(1+\frac{D_i}{D_{l_d}}\right) + \mathbb{E}\log\left(1+\frac{D_{l_d-K}}{D_{l_d}}z\right)+\sum_{i=0}^{l_d-K-1}\mathbb{E}\,\frac{D_i}{D_{l_d-K}}. \quad (3.8)$$

To consider the first term in the right-hand side above, take $l_d-K+1 \le i \le l_d-1$. Then $\theta \ge \mu_{l_d} \ge (\mu_{l_d}+\mu_i)/2$ and because of the structure of the atoms in the clusters, $\theta-(\mu_{l_d}+\mu_i)/2 = O(\lambda_f)$. Note that, i and l_d belong to the same cluster. Using symmetry of the distribution of Z we get

$$\mathbb{E}\log\left(1+\frac{D_i}{D_{l_d}}\right) = \mathbb{E}\log\left(1+\exp\left((\mu_i-\mu_{l_d})\left(Z+\theta-\frac{\mu_{l_d}+\mu_i}{2}\right)\right)\right)$$
$$= \mathbb{E}\log\left(1+\exp\left((\mu_{l_d}-\mu_i)\left(Z-\theta+\frac{\mu_{l_d}+\mu_i}{2}\right)\right)\right) = O(1).$$

Summing over $l_d-K+1 \le i \le l_d-1$ we get the first term in the right-hand side of equation (3.8) is $O(1)$.

Now let us consider the second term in the right-hand side of (3.8). Since $\mu_{l_d}-\mu_{l_d-K}=\lambda_e$ and $\theta \ge \mu_{l_d}$, we get $(\mu_{l_d}+\mu_{l_d-K})/2-\theta \le -\lambda_e/2$. Therefore

$$\begin{aligned}
&\mathbb{E}\log\left(1+\frac{D_{l_d-K}}{D_{l_d}}\right)\\
&=\mathbb{E}\log\left(1+\exp\left((\mu_{l_d}-\mu_{l_d-K})\left(Z+\frac{\mu_{l_d}+\mu_{l_d-K}}{2}-\theta\right)+\frac{\lambda_e^2}{2}\right)\right)\\
&=\mathbb{E}\log\left(1+\exp\left(\lambda_e\left(Z+\frac{\mu_{l_d}+\mu_{l_d-K}}{2}-\theta\right)+\frac{\lambda_e^2}{2}\right)\right)\\
&\le \mathbb{E}\log\left(1+\exp\left(\lambda_e Z\right)\right)\\
&\le \log 2+\lambda_e\,\mathbb{E}\,Z_+\\
&=O(\lambda_f).
\end{aligned}$$

This shows that the second term in the right-hand side of (3.8) is $O(\lambda_f)$. Here we see how η, which is the probability decay rate from cluster to cluster, goes with the cluster length λ_e.

Finally, for each $0 \le i < l_d-K$ define $b_i = \lfloor (l_d-K-i)/K \rfloor$. Then

$$\mathbb{E}\log\left(1+\frac{D_i}{D_{l_d-K}}\right) \le \mathbb{E}\,\frac{D_i}{D_{l_d-K}} \le 2K\exp\left((\mu_{l_d-K}-\mu_i)(\mu_{l_d-K}-\theta)+b_i\frac{\lambda_e^2}{2}\right). \quad (3.9)$$

Note that, $\theta - \mu_{l_d-K} \geq \lambda_e$ and $\mu_{l_d-K} - \mu_i \geq b_i\lambda_e$. Thus, we have

$$(\mu_{l_d-K} - \mu_i)(\mu_{l_d-K} - \theta) + b_i\frac{\lambda_e^2}{2} \leq -b_i\lambda_e^2 + b_i\frac{\lambda_e^2}{2} = -b_i\frac{\lambda_e^2}{2}.$$

Using (3.9) and summing over i in the range $0 \leq i \leq l_d - K$ we get

$$\sum_{0\leq i\leq l_d-K} \mathbb{E}\log\left(1 + \frac{D_i}{D_{l_d-K}}\right) \leq \sum_{0\leq i\leq l_d-K} \mathbb{E}\frac{D_i}{D_{l_d-K}} \leq K\sum_{p=0}^{b_0} e^{-p\lambda_e^2/2} = O(1).$$

This shows that the third term in the right-hand side of (3.8) is $O(1)$. Thus we have proved that the second and third term in the right-hand side of (3.4) are $O(\lambda_f)$ as $\lambda_f \to \infty$. This completes the proof. □

The previous lemma essentially shows that to get an upper bound on $\mathbb{E}\log D_\theta(Z)$ it is enough to consider only $D_{\theta l_d}(Z)$ because asymptotically the contribution of the other terms are negligible. To prove (3.3) using (3.2) we need a lower bound on $\mathbb{E}\log N_{\theta,v}(Z)$, which we get by the straightforward inequality $\mathbb{E}\log N_\theta(Z) \geq \mathbb{E}\log N_{\theta l_n}(Z)$. Of course the novelty is in choice of $l_n(\theta)$ and in the next result we see that these bounds are enough to prove (3.3).

Lemma 3.4. *For $r < r_0$ and for any $\theta > 0$, with $l_n(\theta)$, $l_d(\theta)$, $N_{\theta l_n}$, $D_{\theta l_d}$ defined in Subsection 3.2.1 we have*

$$\frac{\theta^2}{2r} - \mathbb{E}\log N_{\theta l_n}(Z) + \mathbb{E}\log D_{\theta l_d}(Z) \leq \frac{\lambda_f^2}{2r}(1 + o(1)) \quad \text{as } \lambda_f \to \infty.$$

Proof. For convenience we write $l_d(\theta)$ and $l_n(\theta)$ as l_d and l_n respectively. Note that from the definition of l_d it follows $\mu_{l_d} \leq \theta \leq \mu_{l_d+K}$. Let

$$A_\theta := \mu_{l_d}\theta - \frac{\mu_{l_d}^2}{2} - j_{l_d}\frac{\lambda_e^2}{2}, \quad \text{and} \quad B_\theta := \frac{\mu_{l_n}\theta}{v} - \frac{\mu_{l_n}^2}{2v} - j_{l_n}\frac{\lambda_e^2}{2}.$$

From definitions it follows $D_{l_d} = c_{l_d}\exp(\mu_{l_d}Z + A_\theta)$ and $N_{l_n} = c_{l_n}\exp(\mu_{l_n}Z + B_\theta)$. Hence,

$$-\mathbb{E}\log N_{\theta l_n}(Z) \leq -B_\theta + O(1) \text{ as } \lambda_f \to \infty. \tag{3.10}$$

Using $\theta \geq \mu_{l_d} = j_{l_d}\lambda_e$ we get

$$A_\theta = j_{l_d}\theta - \frac{j_{l_d}^2\lambda_e^2}{2} - j_{l_d}\frac{\lambda_e^2}{2} \geq \left(j_{l_d}^2 - j_{l_d}\right)\frac{\lambda_e^2}{2} \geq 0.$$

Using this, we derive the upper bound

$$\begin{aligned}\mathbb{E}\log(1 + D_{\theta l_d}(Z)) &= \mathbb{E}\log(1 + c_\lambda\exp(\mu_{l_d}Z + A_\theta)) \\ &= A_\theta + \mathbb{E}\log(c_\lambda + \exp(\mu_{l_d}Z - A_\theta)) \leq A_\theta + \mathbb{E}(\mu_{l_d}Z - A_\theta)_+ + O(1).\end{aligned}$$

Since $\mu_{l_d} = j_{l_d}\lambda_e$ and $A_\theta \geq (j_{l_d}^2 - j_{l_d})\lambda_e^2/2$, we see that $\mathbb{E}(\mu_{l_d}Z - A_\theta)_+ = O(\lambda_f)$. Hence,

$$\mathbb{E}\log(1 + D_{\theta l_d}(Z)) \leq A_\theta + O(\lambda_f). \tag{3.11}$$

Combining equations (3.10) and (3.11) we get

$$\frac{\theta^2}{2r} - \mathbb{E}\log N_{\theta l_n}(Z) + \mathbb{E}\log D_{\theta l_d}(Z) \le \frac{\theta^2}{2r} - A_\theta + B_\theta + O(\lambda_f).$$

We will show that $\theta^2/(2r) + A_\theta - B_\theta \le \lambda_f^2/(2r)$. First consider the case $l_n = l_d$ which means $\theta \in [j_{l_d}\lambda_e, j_{l_d}\lambda_e + \lambda_f]$. Observe in this case

$$\frac{\lambda_f^2}{2r} - \frac{\theta^2}{2r} - A_\theta + B_\theta = \frac{\lambda_f^2}{2r} - \frac{\theta^2}{2r} + \frac{\mu_{l_d}\theta}{r} - \frac{\mu_{l_d}^2}{2r} = \frac{1}{2r}\left(\lambda_f^2 - (\mu_{l_d} - \theta)^2\right) \ge 0.$$

Now consider the case $l_n \ne l_d$, that is, $l_d + 1 \le l_n \le l_d + K$. In this case, using $j_{l_n} = j_{l_d} + 1$, we get

$$\frac{\lambda_f^2}{2r} - \frac{\theta^2}{2r} - A_\theta + B_\theta = -\frac{\lambda_f^2}{2} - \frac{\theta^2}{2r} + \frac{\mu_{l_n}\theta}{v} - \frac{\mu_{l_n}^2}{2v} - \mu_{l_d}\theta + \frac{\mu_{l_d}^2}{2}. \tag{3.12}$$

If we fix values of l_n and l_d then this is a quadratic in θ with roots $\alpha(l_n, l_d)$ and $\beta(l_n, l_d)$ where

$$\alpha(p,q) := \mu_p + r\,(\mu_p - \mu_q) - r\left(\frac{(\mu_p - \mu_q)^2}{v} - \frac{\lambda_f^2}{r}\right)^{1/2}, \tag{3.13}$$

$$\beta(p,q) := \mu_p + r\,(\mu_p - \mu_q) + r\left(\frac{(\mu_p - \mu_q)^2}{v} - \frac{\lambda_f^2}{r}\right)^{1/2}. \tag{3.14}$$

To show that the right-hand side of (3.12) is nonnegative for all θ, we need to verify that for the range of θ for which $l_n(\theta) = p$ and $l_d(\theta) = q$ is in the interval $[\alpha(p,q), \beta(p,q)]$ for all feasible p, q ($l_n(\theta)$ determines $l_d(\theta)$). Because of the periodicity of the clusters, it's enough to consider the first cluster. So we fix $q = 0$ and consider p going from 1 to K.

Consider the case $p = 1$. Using $\mu_0 = 0$ and $\mu_1 = \lambda_f$ we get $\alpha(1,0) = \lambda_f$ and $\beta(1,0) = \lambda_f(1 + 2r)$. By definition, $l_n = 1$ for $\theta \in [\alpha(1,0), \beta(1,0)]$ as required. With some calculations we can see that if $\alpha(2,0) = \lambda_f(1 + 2r)$ then $\mu_2 = \lambda_f(1 + 4r)$. In general for any $p > 1$ we can verify $\alpha(p,0) < \beta(p-1,0)$ proving that the expression in the right-hand side of (3.12) is always non-negative. □

3.2.4. Proof of Theorem 1.1 for $r \ge r_0$

The proof follows essentially the same ideas of the proof in the case $r < r_0$ but there are some technical differences. The analysis of risk at the origin is unchanged because Lemma 3.1 holds for all r. So now, we analyze risk at non-origin points and basically prove (3.6).

As before, we use the decomposition of risk in (3.2). Our strategy is the same, that is, showing that contribution of $\mathbb{E}\,D_{\theta l_d(\theta)}(Z)$ for one particular index $l_d(\theta)$ is dominant in $\mathbb{E}\log D_\theta(Z)$ and using a naive lower bound on $\mathbb{E}\log N_{\theta v}(Z)$ considering $\mathbb{E}\,N_{\theta l_n(\theta)}$ for one particular index $l_n(\theta)$.

The choices of the indexes in this case are slightly different. Recall that each cluster C_j of $\pi_{\mathsf{C}}[\eta, r]$ consists of only one point. The atoms are at $\mu_p = p\lambda_f$ for all $p \in \mathbb{Z}$. By symmetry we only consider $\theta > 0$. By Lemma 3.2, which didn't depend on value of r, we can ignore all $D_{\theta p}$ with $p < 0$. Suppose $\theta \in [\mu_l, \mu_{l+1})$ for some $l \geq 0$. The contribution of $D_{\theta i}$ for all $i > l$ is negligible compared to $D_{\theta l}$ and the proof is exactly same as done in the beginning of Lemma 3.3, c.f., equations (3.4), (3.5), (3.6) and (3.7). The crucial difference from the sub-critical case arises now. We will see that, if $l \geq 1$, then unlike the sub-critical case $D_{\theta l}$ is not always the dominant term. Instead $D_{\theta,l-1}$ dominates $D_{\theta l}$ for some θ if $r > r_0$. To see this, note that,

$$\mathbb{E}\log\left(1 + \frac{D_{\theta,l-1}}{D_{\theta l}}\right) = \mathbb{E}\log\left(1 + \frac{c_{l-1}}{c_l}\exp\left\{\lambda_f\left(Z + \frac{\mu_l + \mu_{l-1}}{2} - \theta\right) + \frac{\lambda_e^2}{2}\right\}\right).$$

Hence, $\mathbb{E}\, D_{\theta l}(Z)$ dominates $\mathbb{E}\, D_{\theta,l-1}(Z)$ if $\lambda_f((\mu_l + \mu_{l-1})/2 - \theta) + \lambda_e^2/2 \leq 0$, which simplifies to $\theta \geq \mu_l + \lambda_f/(2r)$. For $\theta \in [\mu_l, \mu_l + \lambda_f/(2r)]$, it can be shown that $D_{\theta,l-1}$ is dominant. Also note that for $l \geq 2$

$$\mathbb{E}\log\left(1 + \frac{D_{\theta,l-2}}{D_{\theta,l-1}}\right) = \mathbb{E}\log\left(1 + \frac{c_{l-2}}{c_{l-1}}\exp\left\{\lambda_f\left(Z + \frac{\mu_{l-2} + \mu_{l-1}}{2} - \theta\right) + \frac{\lambda_e^2}{2}\right\}\right).$$

Using $r \geq 1/2$

$$\lambda_f\left(\frac{\mu_{l-2} + \mu_{l-1}}{2} - \theta\right) + \frac{\lambda_e^2}{2} \leq \frac{\lambda_e^2}{2} - \frac{3\lambda_f^2}{2} \leq 0.$$

Hence, $D_{\theta,l-2}$ and the preceding $D_{\theta i}$s are not dominant.

Now if $D_{\theta l}$ is dominant, that is, $\theta \in [\mu_l + \lambda_f/(2r), \mu_{l+1})$ then using the naive lower bound $\mathbb{E}\log N_{\theta,v}(Z) \geq \mathbb{E}\log N_{\theta l}(Z)$ we get

$$\rho(\theta, \hat{p}_{\mathsf{C}}[\eta, r]) = \frac{\theta^2}{2r} - \mathbb{E}\log N_{\theta,v}(Z) + \mathbb{E}\log D_\theta(Z) \leq \frac{\lambda_f^2}{2r}(1 + o(1)).$$

We skip the details of the proof because it's exactly similar to the case $l_d = l_n$ in Lemma 3.4.

On the other hand, if $D_{\theta,l-1}$ is dominant, that is, $\theta \in [\mu_l, \mu_l + \lambda_f/(2r))$, then we use $\mathbb{E}\log N_{\theta l}(Z)$ as a lower bound of $\mathbb{E}\log N_{\theta,v}(Z)$. We end up with a quadratic in θ similar to equation (3.12), which is nonnegative in $[\mu_l, \mu_l + 2r\lambda_f]$. Since this interval covers the interval $[\mu_l, \mu_l + \lambda_f/(2r)]$ for $r \geq r_0$ we get the the above equation.

3.3. Proof of Lemma 1.2

Similar to Lemma 3.4, here also, since $l > \lambda_e$, we can show that in the risk decomposition in (2.2) $\mathbb{E}\log N_{\theta,v}(Z)$ can be replaced with $\mathbb{E}\log N_{\theta,l_n(\theta)}(Z)$ for some $l_n(\theta)$, and similarly $\mathbb{E}\log D_\theta(Z)$ can be replaced with $\mathbb{E}\log D_{\theta,l_d(\theta)}(Z)$ for some $l_d(\theta)$ ($D_{\theta,p}$ and $N_{\theta,p}$ defined in Subsection 3.2.1), in the sense that the

difference is negligible, that is $O(\lambda_f)$, asymptotically. Consider $\theta \in [\nu + (p-1)l, \nu + pl]$ for some $p > 1$. By calculations similar to Lemma 3.4, we can show for $\theta \in [\nu + (p-1)l, \nu + (p-1)l + \lambda_f]$ if we choose $l_n = l_d = p$ then the risk is below the threshold $\lambda_f^2/(2r)$ asymptotically and similarly $l_n = l_d = p+1$ works for $\theta \in [\nu+pl-\lambda_f, \nu+pl]$. But for $r \leq 1/3$ we have $\nu+pl-\lambda_f > \nu+(p-1)l+\lambda_f$, so that for $\theta = \nu + (p-1)l + \lambda_f + \epsilon$ for small $\epsilon > 0$, we must choose $l_n = p+1$ and $l_d = p$. This makes the risk go above the threshold $\lambda_f^2/(2r)$ leading to sub-optimality.

3.4. Proof of Lemma 1.3

From calculations of Lemma 3.4 it is clear that the for any $r < r_0$ the choice of $\gamma = \gamma_r = (1+4r)$ cannot be improved upon for asymptotic minimax optimality. So if we have $\gamma > \gamma_r$ then we will have asymptotic minimax suboptimality. Since the common ration γ determines the cluster size K_γ, the cluster size cannot be improved if we want to maintain minimax optimality. Also if we keep $\gamma = \gamma_r$ and then dropping points from the geometrically spaced grid also causes suboptimality. To see this, suppose the first point we drop is the m'th point in the cluster. If $m = 1$, i.e., we drop λ_f then as we have already seen that this creates suboptimality. Let $m > 1$. This implies that (3.13)-(3.14) defined in the proof of Lemma 3.4 must statisfy the constraint: $\beta(m+1, 0) > \alpha(m-1, 0)$. Writing down the constraint in terms of r and letting $r \to 0$ we get a contradiction, which shows that we cannot drop support points from our prescribed prior.

3.5. Proof of Theorem 1.4

The Bayes risk of the multivariate cluster prior $B(\pi_n^{\mathsf{C}}) = nB(\pi_{\mathsf{C}})$ and the univariate Bayes risk is given by

$$B(\pi_{\mathsf{C}}) = (1-\eta_n)\rho(0, \hat{p}_{\mathsf{C}}[\eta_n, r]) + \frac{1-\eta_n}{2K}\sum_{i=1}^{\infty}\sum_{j=1}^{K}\eta_n^{|i|}\rho(\mu_{ij}, \hat{p}_{\mathsf{C}}[\eta_n, r])$$

where K is defined in (1.6). From the risk calculations in Lemmas 3.2, 3.3, 3.4 it is clear that the first order asymptotic risk as $\eta_n \to 0$ can be reduced to just concentrating on the origin adjoining clusters $C_{\pm 1}$ and thereafter by symmetry:

$$B(\pi_{\mathsf{C}}) = \frac{(1-\eta_n)\eta_n}{K}\sum_{j=1}^{K}\rho(\mu_{1j}, \hat{p}_{\mathsf{C}}[\eta_n, r])(1+o(1)) \ .$$

Now, by (3.2) and Lemma 3.4, for each $1 \leq j \leq K$ we have:

$$\rho(\mu_{1j}, \hat{p}_{\mathsf{C}}[\eta_n, r]) = \mu_{1j}^2/(2r) - \mathbb{E}\log N_{\mu_{1j},v}(Z) + O(\lambda_{f,n}) \ \text{ as } \eta_n \to 0 \ .$$

Also, following exactly the similar asymptotic analysis as in Lemma 3.4 abet now with $N_{\mu_{1j},v}(Z)$ we can establish $\mathbb{E}\log N_{\mu_{1j},v}(Z) = (2v)^{-1}(\mu_{1j}^2 - \lambda_{f,n}^2)(1+$

$o(1)$). By construction, $\mu_{1j} \geq \lambda_{f,n}$ with strict equality only when $j = 1$ and so each of the terms barring the first one has some positive contributions. Thus, $\rho(\mu_{11}, \hat{p}_{\mathsf{C}}[\eta_n, r]) = \lambda_{f,n}^2/(2r)(1 + o(1))$. For all $j > 1$, recalling $\mu_{1j}/\lambda_{f,n} = (1 + 4r)^j \wedge v^{-1/2}$ and $v = (1 + r^{-1})^{-1}$ we have,

$$\rho(\mu_{1j}, \hat{p}_{\mathsf{C}}[\eta_n, r]) = 2^{-1}\lambda_{f,n}^2\big\{1 + r^{-1} - (1 + 4r)^{2j}\big\}_+ + O(\lambda_{f,n})$$

where, the first term in the right side above is 0 only when $j = K$. Thus, the maximal risk is only attained at $\mu_{11} = \lambda_{f,n}$. Thereafter, the risk decays and finally at $j = K$, the risk is negligible compared to the asymptotic minimax risk. Figure 3 shows the numerical evaluation of the risk of the cluster prior at the different support point of the first cluster. The figure shows the risk profile when $\eta_n = 10^{-15}$ which well captures the asymptotic analysis and the aforementioned decay in the risk function is evident from the figure.

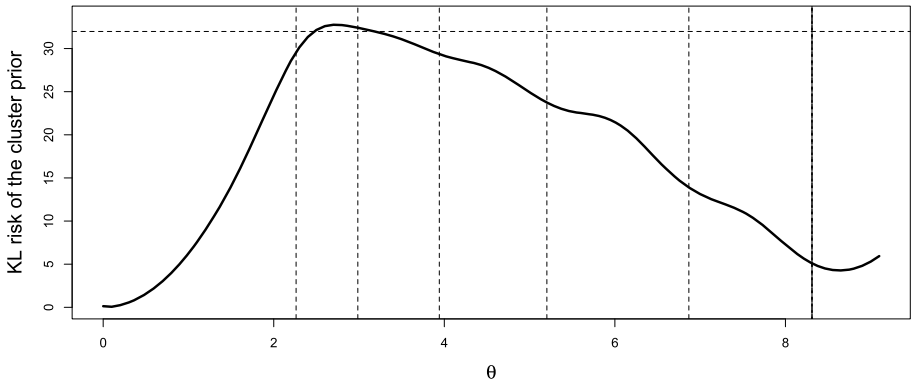

FIG 3. *Plot of the univariate predictive KL risk $\rho(\theta, \hat{p}_{\mathsf{C}}[\eta_n, r]$ as θ varies over the first cluster spanning $[0, \lambda_e(\eta_n)]$. Here, $\eta_n = 10^{-15}$ and $r = 0.08$. The horizontal line corresponds to the asymptotic theoretical limit $\lambda_f^2(\eta_n)/(2r)$. The dotted vertical lines denotes the location of the the support points in cluster C_1 of $\pi_{\mathsf{C}}[\eta_n, r]$.*

Noting that the multivariate minimax risk is $n\eta_n\lambda_{f,n}^2/(2r)(1 + o(1))$ as $\eta_n \to 0$, the result follows from the above display. When $r > r_0$, then $K = 1$ and so, the above result directly imply $B(\pi_n^{\mathsf{C}})/R^*(\Theta_0[s_n]) \to 1$ as $n \to \infty$. The condition $s_n \to \infty$ ensures that the prior concentrates on the parametric space $\Theta_0[s_n]$ (see Theorem 1B of MJ15 for details) and thus is least favorable in this case.

4. Simulations

We introspect the performance of the aforementioned PRDEs across different sparsity regimes. The product structure of our estimation framework allows us to concentrate on the maximal risk of the corresponding univariate PRDEs. (2.1) shows that the multivariate maximum risk of $\hat{p}_{\mathsf{C}}$ over $\Theta_0[s_n]$ is a function of sparsity level η_n and the univariate risk of $\hat{p}_{\mathsf{C}}$. In table 2, we report the

TABLE 2

Numerical evaluation of the maximum risk for the different univariate predictive density estimates as the degree of sparsity (η) and predictive difficulty r varies. The asymptotic theory based minimax risk value is reported in 'A-Theory' and the subsequent columns report the maximum risk of the estimators as quotients of 'A-Theory' values.

Sparsity	**r**	**A-Theory**	**Plugin**	$\hat{p}_{\mathsf{T}}$	$\hat{p}_{\mathsf{EG}}$	$\hat{p}_{\mathsf{PG}}$	$\hat{p}_{\mathsf{C}}$
0.2	1	0.8047	1.3626	0.9934	0.8321	1.0384	1.1019
	0.4	1.1496	2.2073	1.1302	1.2182	1.3824	1.5500
	0.2	1.3412	3.5794	1.5878	1.6444	1.6730	1.9128
	0.1	1.4631	6.2714	1.9686	2.8799	2.0024	2.5775
0.10	1	1.1513	1.2037	0.8250	0.6908	0.8700	0.9364
	0.4	1.6447	1.9999	0.9464	0.9962	1.1108	1.1568
	0.2	1.9188	3.3046	1.2179	1.4599	1.3572	1.4608
	0.1	2.0933	5.8811	1.4644	2.6143	1.6369	1.9706
0.05	1	1.4979	1.1319	0.7434	0.6473	0.7845	0.8573
	0.4	2.1398	1.9102	0.8907	0.9767	0.9880	1.0055
	0.2	2.4964	3.1917	1.0344	1.4607	1.1814	1.2722
	0.1	2.7234	5.7314	1.1756	2.5838	1.4082	1.6359
0.01	1	2.3026	1.0841	0.7065	0.6279	0.7287	0.7537
	0.4	3.2894	1.8602	0.8643	0.9827	0.9211	0.9170
	0.2	3.8376	3.1433	0.9636	1.5069	1.0680	1.1132
	0.1	4.1865	5.6949	1.1083	2.6074	1.2092	1.3617
0.001	1	3.4539	1.0913	0.7022	0.6849	0.7081	0.6966
	0.4	4.9341	1.8849	0.8792	1.0408	0.9104	0.9086
	0.2	5.7565	3.2018	0.9493	1.5674	1.0309	1.0746
	0.1	6.2798	5.8275	1.0696	2.7097	1.1471	1.1883
0.0001	1	4.6052	1.1130	0.7193	0.7164	0.7177	0.7118
	0.4	6.5788	1.9292	0.8990	1.0548	0.8918	0.9111
	0.2	7.6753	3.2857	0.9512	1.5981	1.0142	1.0378
	0.1	8.3730	5.9936	1.0502	2.7587	1.1083	1.1301
0.00001	1	5.7565	1.1371	0.7364	0.7437	0.7248	0.7267
	0.4	8.2235	1.9752	0.9102	1.0939	0.8940	0.9132
	0.2	9.5941	3.3695	0.9609	1.6501	1.0108	1.0262
	0.1	10.4663	6.1542	1.0429	2.7887	1.0995	1.1000
10^{-10}	1	11.5129	1.2390	0.7954	0.8304	0.7888	0.7872
	0.4	16.4470	2.1615	0.9321	1.1669	0.9000	0.9298
	0.2	19.1882	3.6981	0.9680	1.7358	1.0109	1.0184
	0.1	20.9326	6.7701	1.0185	2.9049	1.0916	1.0253

maximum risk of our proposed clustered prior based Bayes PRDE $\hat{p}_{\mathsf{C}}$ (in last column) as the degree of sparsity η and predictive difficulty r varies. Using (2.2), we evaluate the univariate risk of $\hat{p}_{\mathsf{T}}$ for any fixed θ with high precision by using Monte Carlo integration. Thereafter, the maximum risk is found by a conducting a univariate grid search as θ varies over $\mathbb{R}$.

The performance of the following related PRDEs (a) hard thresholding based

plugin estimator (b) thresholding based risk diversified PRDE $\hat{p}_{\mathsf{T}}$ of MJ15 (c) Bayes PRDE $\hat{p}_{\mathsf{EG}}$ based on π_{EG} prior of J94 (d) Bayes PRDE $\hat{p}_{\mathsf{PG}}$ based on π_{PG} prior of MJ17 are respectively reported in columns 4 to 7 in table 2. The risk of $\hat{p}_{\mathsf{EG}}$ and $\hat{p}_{\mathsf{PC}}$ are evaluated by looking at their maximum univariate risk and analogous version of (2.1) which follows from Lemma 2.1 of MJ17. The maximum risk of $\hat{p}_{\mathsf{T}}$ is numerically evaluated by combining (24), (34), (43) and (47) in MJ15.

Under moderate sparsity, the maximal values of the PRDEs exceed the minimax value specified by the asymptotic theory in (1.2). The exceedance is higher for lower values of r. From the table, it seems that the numeric results are in accordance with the asymptotic theory as $\eta_n \leq 10^{-3}$. As expected the plug-in PRDE is highly sub-optimal for lower values of r across all regimes. Once the asymptotic behaviour sets in, the maximum risk of the proposed PRDE $\hat{p}_{\mathsf{C}}$ is near optimal among the concerned PRDEs across all r regimes; under moderate sparsity its maximum risk is little worse but for $\eta_n \leq 10^{-3}$ it has lower maximum risk than $\hat{p}_{\mathsf{EG}}$, $\hat{p}_{\mathsf{PG}}$, and similar risks as $\hat{p}_{\mathsf{T}}$. However, due to the presence of the countably infinite univariate discrete prior in $\hat{p}_{\mathsf{C}}$, the asymptotic approximation to its maximum risk as described by Theorem 1.1 comes into effect at relatively smaller η_n values than in the risk of $\hat{p}_{\mathsf{T}}$.

5. Discussion and future work

The results developed here assume that the variances v_x and v_y are known. If $v_y = rv_x$ where r is known but v_x is unknown, a simple approach would be to substitute an estimate $\hat{v}_x$ of v_x in the PRDEs discussed here. For $\hat{v}_x$ we can use the median absolute deviation from zero which is used for PE in the `EbayesThresh` package of Johnstone and Silverman (2005). For other good candidates for $\hat{v}_x$, see Xing et al. (2020) and the references therein. However, as shown in Kato (2009b) such plug-in approach will not be optimal. Recently, Maruyama et al. (2020) developed a decision theoretic framework under repeated sampling for studying point estimation efficiency in Gaussian models with unknown scale. As future work, it will be interesting to study PRDE under Kullback-leibler loss in such a framework.

Also, if the sparsity level is unknown, it can be estimated from the data using the empirical Bayes maximum likelihood approach in Johnstone and Silverman (2005) and the estimated sparsity level can be plugged in the form of the Bayes PRDEs discussed in this paper. The PRDEs discussed here are based on spike-and-slab priors with the slab being infinite discrete priors. PRDEs based on continuous slabs (Rockova and George, 2018) are preferred for practical implementation. A manuscript studying adaptivity of such spike-and-slab PRDEs to unknown sparsity is forthcoming.

Acknowledgments

GM is indebted to Professor Iain Johnstone for numerous stimulating discussions which led to many of the ideas in this paper.

References

James Aitchison. Goodness of prediction fit. *Biometrika*, 62(3):547–554, 1975. ISSN 0006-3444. MR0391353

James Aitchison and I. R. Dunsmore. *Statistical prediction analysis*. Cambridge University Press, 1975. MR0408097

James O. Berger and José M Bernardo. Reference priors in a variance components problem. In *Bayesian analysis in statistics and econometrics*, pages 177–194. Springer, 1992. MR1194392

Jose M. Bernardo. Reference posterior distributions for bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2):113–128, 1979. MR0547240

Peter J. Bickel. Minimax estimation of the mean of a normal distribution when the parameter space is restricted. *The Annals of Statistics*, 9(6):1301–1309, 1981. MR0630112

P.J. Bickel. Minimax estimation of the mean of a normal distribution subject to doing well at a point. In *Recent Advances in Statistics*, pages 511–528. Elsevier, 1983. MR0736544

Lawrence D. Brown, Edward I. George, and Xinyi Xu. Admissible predictive density estimation. *Ann. Statist.*, 36(3):1156–1170, 2008. ISSN 0090-5364. 10.1214/07-AOS506. URL http://dx.doi.org/10.1214/07-AOS506.

Merlise Clyde and Edward I George. Flexible empirical bayes estimation for wavelets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):681–698, 2000. MR1796285

Dominique Fourdrinier, Éric Marchand, Ali Righi, and William E. Strawderman. On improved predictive density estimation with parametric constraints. *Electron. J. Stat.*, 5:172–191, 2011. ISSN 1935-7524. 10.1214/11-EJS603. URL http://dx.doi.org/10.1214/11-EJS603. MR2792550

Seymour Geisser. *Predictive inference*, volume 55 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, New York, 1993. ISBN 0-412-03471-9. An introduction. MR1252174

Edward George, Eric Marchand, Gourab Mukherjee, and Debashis Paul. New and evolving roles of shrinkage in large-scale prediction and inference. *BIRS Workshop Report*, 2019.

Edward I. George, Feng Liang, and Xinyi Xu. Improved minimax predictive densities under Kullback-Leibler loss. *Ann. Statist.*, 34(1):78–91, 2006. ISSN 0090-5364. 10.1214/009053606000000155. URL http://dx.doi.org/10.1214/009053606000000155.

Edward I. George, Feng Liang, and Xinyi Xu. From minimax shrinkage estimation to minimax shrinkage prediction. *Statist. Sci.*, 27(1):82–94, 2012. ISSN 0883-4237. 10.1214/11-STS383. URL http://dx.doi.org/10.1214/11-STS383. MR2953497

Malay Ghosh, Victor Mergel, and Gauri Sankar Datta. Estimation, prediction and the Stein phenomenon under divergence loss. *J. Multivariate Anal.*, 99(9):1941–1961, 2008. ISSN 0047-259X. 10.1016/j.jmva.2008.02.002. URL http://dx.doi.org/10.1016/j.jmva.2008.02.002. MR2466545

I. M. Johnstone. Gaussian estimation: Sequence and wavelet models. 2013. Version: 11 June, 2013. Available at http://www-stat.stanford.edu/~imj.

Iain M. Johnstone. On minimax estimation of a sparse normal mean vector. *Ann. Statist.*, 22(1):271–289, 1994a. ISSN 0090-5364. 10.1214/aos/1176325368. URL http://dx.doi.org/10.1214/aos/1176325368.

Iain M. Johnstone. Minimax Bayes, asymptotic minimax and sparse wavelet priors. In *Statistical decision theory and related topics, V (West Lafayette, IN, 1992)*, pages 303–326. Springer, New York, 1994b. MR1286310

Iain M. Johnstone and Bernard W. Silverman. Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences. *Ann. Statist.*, 32(4):1594–1649, 2004. ISSN 0090-5364. 10.1214/009053604000000030. URL http://dx.doi.org/10.1214/009053604000000030.

Iain M. Johnstone and Bernard W Silverman. Ebayesthresh: R and s-plus programs for empirical bayes thresholding. *J. Statist. Soft*, 12:1–38, 2005. MR2364426

Kengo Kato. Improved prediction for a multivariate normal distribution with unknown mean and variance. *Ann. Inst. Statist. Math.*, 61(3):531–542, 2009a. ISSN 0020-3157. 10.1007/s10463-007-0163-z. URL http://dx.doi.org/10.1007/s10463-007-0163-z.

Kengo Kato. Improved prediction for a multivariate normal distribution with unknown mean and variance. *Annals of the Institute of Statistical Mathematics*, 61(3):531–542, 2009b. MR2529965

Peter J. Kempthorne. Numerical specification of discrete least favorable prior distributions. *SIAM Journal on Scientific and Statistical Computing*, 8(2):171–184, 1987. MR0879409

Fumiyasu Komaki. A shrinkage predictive distribution for multivariate normal observables. *Biometrika*, 88(3):859–864, 2001. ISSN 0006-3444. 10.1093/biomet/88.3.859. URL http://dx.doi.org/10.1093/biomet/88.3.859. MR1859415

Tatsuya Kubokawa, Éric Marchand, William E Strawderman, and Jean-Philippe Turcotte. Minimaxity in predictive density estimation with parametric constraints. *Journal of Multivariate Analysis*, 116:382–397, 2013. MR3049911

S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statistics*, 22:79–86, 1951. ISSN 0003-4851. MR0039968

Feng 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. MR2703233

Eric Marchand, William E Strawderman, et al. Estimation in restricted parameter spaces: A review. In *A Festschrift for Herman Rubin*, pages 21–44. Institute of Mathematical Statistics, 2004. MR2126884

Yuzo Maruyama and Toshio Ohnishi. Harmonic bayesian prediction under alpha-divergence. *IEEE Transactionson Information Theory*, 65:5352–53666, 2019. MR4009238

Yuzo Maruyama, William E. Strawderman, et al. Admissible bayes equivariant estimation of location vectors for spherically symmetric distributions with unknown scale. *Annals of Statistics*, 48(2):1052–1071, 2020. MR4102687

G. Mukherjee and I. M. Johnstone. Exact minimax estimation of the predictive density in sparse gaussian models. *Annals of Statistics*, 2015. MR3346693

Gourab Mukherjee. *Sparsity and Shrinkage in Predictive Density Estimation*. PhD thesis, Stanford University, 2013. MR4187552

Gourab Mukherjee and Iain M. Johnstone. On minimax optimality of sparse bayes predictive density estimates. *arXiv preprint 1707.04380*, 2017.

Veronika Rockova and Edward I. George. The spike-and-slab lasso. *Journal of the American Statistical Association*, 113(521):431–444, 2018. MR3803476

Trevor J. Sweeting, Gauri S. Datta, and Malay Ghosh. Nonsubjective priors via predictive relative entropy regret. *The Annals of Statistics*, pages 441–468, 2006. MR2275249

Zhengrong Xing, Peter Carbonetto, and Matthew Stephens. Flexible signal denoising via flexible empirical bayes shrinkage. *Journal of Machine Learning Research*, 2020.

Xinyi Xu and Feng Liang. Asymptotic minimax risk of predictive density estimation for non-parametric regression. *Bernoulli*, 16(2):543–560, 2010. ISSN 1350-7265. 10.3150/09-BEJ222. URL `http://dx.doi.org/10.3150/09-BEJ222`. MR2668914

Z. Zhang. Discrete non-informative priors. In *Ph.D. Dissertation*. Department of Statistics, Yale University, 1994. MR2692257