Minimax Adaptive Predictive Density Estimation for Non-parametric Regression

Gourab Mukherjee

University of Southern California

March 11, 2021

Abstract

We consider the problem of estimating the predictive density of future observations from a non-parametric regression model. The density estimators are evaluated under the global Kullback-Leibler divergence loss. We adapt to the unknown smoothness of the function classes by adjusting our estimator for possibly sparse and dense coefficient sequences in the transformed basis. We study the risk properties of a gamut of popularly used *spike-and-slab* predictive density estimates which are based on mixture priors with an atom of probability at zero in the transformed parametric space. We consider a wide range of function classes which corresponds to exact ℓ_0 sparsity as well strong and weak ℓ_p, $0 < p < \infty$ constrained parameters in the transformed basis. We establish that empirical Bayes predictive density estimates based on a mixture of an atom of probability at zero and any heavy-tailed density starting from exponential to fat-tailed quadratic decays such as the quasi-cauchy density are minimax adaptive when the mixing weights are chosen by marginal maximum likelihood. For all the classes considered, these predictors attain exact minimax optimality for sparse coefficients, and in this sense adapts automatically to the sparseness of the underlying signal and in turn to the smoothness of the function. Additionally, the risk of these predictors is uniformly bounded over all functions suggesting that their usage is not detrimental even if the function classes were mis-specified. We highlight non-parametric regimes where plug-in density estimators are comparatively very sub-optimal. We also show that *spike-and-slab* predictive density estimators based on Gaussian tails are minimax sub-optimal. Our results show that despite glaring differences in minimax decision theoretic phenomena between sparse point estimation and predictive

density estimation, popularly used spike-and-slab priors in sparse estimation enjoy desired decision theoretic guarantees in the predictive regime and can be used for efficient predictive density estimation in non-parametric regression.

Keywords: adaptivity, predictive density, minimax, sparsity, empirical Bayes, non-parametric regression, sparse mixture priors, cauchy prior.

1 Introduction

Predictive density estimation is a fundamental problem in statistical prediction analysis with applications in game theory, econometrics, information theory, machine learning and mathematical finance [2, 14, 34]. Recently, intricate connections with point estimation theory [4, 13, 17, 19, 28, 29, 44] as well as new decision theoretic phenomena [30, 31, 33, 35, 36] have been established in the predictive density estimation problem (See George *et al.* [18] for review). Here, we consider the problem of predictive density estimation in the non-parametric regression setup. This is a flexible framework and is widely used in a host of modern applications where the predictor do not have any predetermined form [12]. Asymptotic optimality of estimators in such frameworks is usually studied through the minimax risk associated with estimation over different function spaces. A huge body of literature has been devoted to the evaluation of minimax risks and optimal estimators over varied function spaces. An excellent survey of the literature in this area can be found in Johnstone [23]. In contrast, few results have been obtained on predictive density estimation for non-parametric models. Till date the most comprehensive result in this direction has been by Xu and Liang [43] who evaluated the minimax risk of the predictive density estimation problem over ellipsoids and showed that linear predictive density estimators (prde s) are minimax optimal. However, for function estimation under quadratic loss it has been witnessed that for sufficiently smooth function classes which translate to estimation over relatively sparser parametric spaces than ellipsoids, linear estimators are sub-optimal and sparsity inducing mixture priors or threshold estimators are used for attaining minimax optimality [see 8, Ch 9.5 of 23 and the references therein]. A rich literature has been developed for constructing effective estimators that adapt to the unknown smoothness of the function classes [See 7, 9, 10, 27, Ch 11 of 20] However, no such results exist for the predictive density estimation problem.

In this paper we first evaluate the minimax risk of predictive density estimation over a wide range of smooth function classes that are governed by exact, strong as well as weak ℓ_p constraints on the transformed sequence model and present computationally tractable minimax optimal prdes that depend on the smoothness the function

classes. Thereafter, after we propose a class of *Spike-and-Slab* priors based *adaptive predictive density estimates* (SASA) that are tuned by empirical Bayes maximum likelihood estimation (EBMLE) method. We show that prdes in SASA work well over a wide range of function classes without any prior knowledge of their smoothness pattern and level. SASA attains minimax optimality over a large collection sparse spaces (and hence is minimax adaptive) and has well controlled predictive error over dense spaces. We next present our predictive framework, adaptive density estimators and the main results.

1.1 Predictive Framework and Minimax Risk

Non-parametric regression setup. Following Xu and Liang [43], we consider the canonical non-parametric regression predictive model, where

$$\mathfrak{d}(\mathfrak{t}_i) = \mathfrak{f}(\mathfrak{t}_i) + \sigma\,\epsilon_i\ , \quad i = 1, 2, 3 \cdots, n \tag{1}$$

and, $\mathfrak{f}$ is an unknown function in $\mathcal{L}^2[0,1]$, ϵ_i's are i.i.d. standard normal noise, σ is known and $0 \leq \mathfrak{t}_1 < \mathfrak{t}_2 \cdots < \mathfrak{t}_n = 1$ is an equi-spaced grid on $[0,1]$ with $\mathfrak{t}_i = i/n$ for $i = 1, 2, \cdots, n$. We would like to get the predictive density of a future observation from the above model that is sampled at equidistant m points $\mathfrak{s}_i = i/m$ for $i = 1, \cdots, m$. As such,

$$\tilde{\mathfrak{d}}(\mathfrak{s}_i) = \mathfrak{f}(\mathfrak{s}_i) + \sigma\,\tilde{\epsilon}_i, \quad i = 1, 2, \cdots, m$$

where $\tilde{\epsilon_i}$'s and ϵ_i's are independent Gaussian noise and $\mathfrak{f}$ is the same function as (1). Define $\mathbb{D} = (\mathfrak{d}(\mathfrak{t}_1), \cdots, \mathfrak{d}(\mathfrak{t}_n))$ and $\tilde{\mathbb{D}} = (\tilde{\mathfrak{d}}(\mathfrak{s}_1), \cdots, \tilde{\mathfrak{d}}(\mathfrak{s}_m))$. Based on observing $\mathbb{D}$, we seek estimators $\hat{\mathfrak{p}}(\tilde{\mathbb{D}}|\mathbb{D})$ of the future observation density $\mathfrak{p}(\tilde{\mathbb{D}}|\mathfrak{f}) = \prod_{i=1}^{m} \phi(\tilde{\mathfrak{d}}_i; f(\mathfrak{s}_i), \sigma^2)$, where $\phi(x; \mu, \sigma^2)$ denotes the normal density function at x for mean μ and variance σ^2. Any predictive density estimator $\hat{\mathfrak{p}}(\tilde{\mathbb{D}}|\mathbb{D})$ is a non-negative function of $\tilde{\mathbb{D}}$ that integrates to 1 with respect to $\tilde{\mathbb{D}}$. We recall two natural ways of generating large classes of estimators. A naive approach is the "plugin" or estimative rule that simply substitutes an estimate $\hat{\mathfrak{f}}$ for $\mathfrak{f}$ in $p(\tilde{\mathbb{D}}|\mathfrak{f})$. The comparatively subtler approach is to consider the Bayes predictive density $\hat{\mathfrak{p}}_\pi(\tilde{\mathbb{D}}|\mathbb{D}) = \int \mathfrak{p}(\tilde{\mathbb{D}}|f)\,\pi(d\mathfrak{f}\,|\,\mathbb{D})$ for any given prior measure π, proper or improper, such that the posterior $\pi(d\mathfrak{f}|\,\mathbb{D})$ is well defined. Here, we evaluate the performance of a predictive density estimator $\hat{\mathfrak{p}}$ by the global divergence measure of Kullback-Leibler (KL) [32] averaged over the m predictive points:

$$\bar{\mathcal{R}}_{n,m}(\mathfrak{f}, \hat{\mathfrak{p}}) = \frac{1}{m}\,\mathbb{E}_{(\mathbb{D}, \tilde{\mathbb{D}}|\mathfrak{f})}\,\log\left(\frac{\mathfrak{p}(\tilde{\mathbb{D}}|\mathfrak{f})}{\hat{\mathfrak{p}}(\tilde{\mathbb{D}}|\mathbb{D})}\right),$$

where, the expectation is over the true joint density $\mathfrak{p}(\mathbb{D}|\mathfrak{f}) \cdot \mathfrak{p}(\tilde{\mathbb{D}}|\mathfrak{f})$. For the function space $\mathcal{F} \subseteq \mathcal{L}^2[0,1]$, the minimax KL risk for estimation over $\mathcal{F}$ is,

$$\bar{\mathcal{R}}^{\mathsf{M}}_{n,m}(\mathcal{F}) = \inf_{\hat{\mathfrak{p}}} \sup_{\mathfrak{f} \in \mathcal{F}} \bar{\mathcal{R}}_{n,m}(\mathfrak{f}, \hat{\mathfrak{p}}) \ .$$

The predictive distribution function assigns probabilities to all possible outcomes and provides a complete description of the uncertainty associated with the data generation process. Compared to estimating $\mathfrak{f}$ itself, predictive density estimates are more useful when we are interested in making inferences about future observations from the same process. In particular, plug-in approaches can be far out-performed by efficient prde s when the prediction grid $\{\mathfrak{s}_i\}_{i=1}^m$ is denser than $\{\mathfrak{t}_i\}_{i=1}^n$. Henceforth, we assume $m \geq n$ and study the efficacy of different predictive density estimation methods.

Transformed Sequence Model. Let $\{\mathfrak{u}_i\}_{i=1}^\infty$ be an orthonormal basis of $\mathcal{L}^2[0,1]$. Then, $\mathfrak{f} = \sum_{i=1}^\infty \mu_i \, \mathfrak{u}_i$ where $\mu_i = \int_0^1 \mathfrak{f}(t)\mathfrak{u}_i(t)\, dt$. As in [27], we tie the notion of smoothness of the class $\mathcal{F}$ of functions in $\mathcal{L}^2[0,1]$ by the sparseness of $\{\mu_i : 1 \leq i \leq n\}$. Let $\sigma_n = n^{-1/2}\sigma$, $\theta_i = \sigma_n^{-1}\mu_i$ and $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_n)$. Consider matrices $\mathbb{A}$ and $\mathbb{B}$ where $\mathbb{A}_{i,j} = \mathfrak{u}_j(\mathfrak{t}_i)$ and $\mathbb{B}_{k,j} = \mathfrak{u}_j(\mathfrak{s}_k)$ where $1 \leq i, j \leq n$ and $1 \leq k \leq m$. Define $\boldsymbol{X} = n^{-1}\sigma_n^{-1}\mathbb{A}'\mathbb{D}$ and $\boldsymbol{Y} = m^{-1}\sigma_n^{-1}\mathbb{B}'\tilde{\mathbb{D}}$. Then,

$$\boldsymbol{X} \sim N_n(\boldsymbol{\theta}, \mathrm{I}) \quad \text{and} \quad \boldsymbol{Y} \sim N_n(\boldsymbol{\theta}, r_{m,n}\, \mathrm{I}) \tag{2}$$

where $r_{m,n} = n/m \in (0,1]$. Note that, Fraktur font is used in the non-parametric regression model but not for the sequence model where bold symbols denotes vectors. Consider estimating the true predictive density $p(\boldsymbol{Y}|\boldsymbol{\theta}, \mathrm{I})$ in model (2) by any density estimator $\hat{p}(\boldsymbol{Y}|\boldsymbol{X})$. Its KL risk is:

$$\rho_{n,m}(\boldsymbol{\theta}, \hat{p}) = \int p(\boldsymbol{x}|\boldsymbol{\theta}, \mathrm{I})\, p(\boldsymbol{y}|\boldsymbol{\theta}, r_{n,m}\, \mathrm{I}) \, \log\left(\frac{p(\boldsymbol{y}|\boldsymbol{\theta}, r_{n,m}\, \mathrm{I})}{\hat{p}(\boldsymbol{y}|\boldsymbol{x})}\right) d\boldsymbol{y}\, d\boldsymbol{x} \ ,$$

and the corresponding minimax error for estimation over the parametric space Θ_n is

$$\rho^{\mathsf{M}}_{n,m}(\Theta_n) = \inf_{\hat{p}} \sup_{\boldsymbol{\theta} \in \Theta_n} \rho_{n,m}(\boldsymbol{\theta}, \hat{p}).$$

Sparsity Constraints on Θ. We first consider a natural notion of sparsity [23] where that the number of non-zero coordinates $\mu_{\mathfrak{i}}$ is bounded. This results in ℓ_0-sparse parametric spaces:

$$\ell_0\left[\eta_n\right] := \left\{ \boldsymbol{\theta} : n^{-1} \sum_{i=1}^n I[\theta_i \neq 0] \leq \eta_n \right\}.$$

We next consider strong ℓ_p constrained parametric space, which do not require μ_i to be exactly zero, but still constrain its strength to be concentrated on a few of the μ_i, by placing bounds on the p-norm of $\boldsymbol{\mu}$ for $p > 0$. For any $p > 0$, define strong ℓ_p-sparse parametric space as,

$$\ell_p[\eta_n] := \left\{ \boldsymbol{\mu} : n^{-1} \sum_{i=1}^{n} |\mu_i|^p \leq (\eta_n/\sigma_n)^p \right\} = \left\{ \boldsymbol{\theta} : n^{-1} \sum_{i=1}^{n} |\theta_i|^p \leq \eta_n^p \right\}.$$

We extend the aforementioned notion of sparsity to a more subtler characterization that uses the rate at which the individual magnitudes $|\mu_i|$ decay. Define Marcinkiewicz weak ℓ_p-sparsity as

$$\ell_p^*[\eta_n] = \left\{ \boldsymbol{\mu} : (k/n)^{1/p} |\mu|_{(k)} \leq \eta_n/\sigma_n \right\} = \left\{ \boldsymbol{\theta} : (k/n)^{1/p} |\theta|_{(k)} \leq \eta_n \right\}.$$

Asymptotic Framework. We consider a sequence of experiments where $m := m_n$ and $r_{n,m} = n/m_n := r_n$ is such that $r_n \to r$ as $n \to \infty$. The undersampling ratio $r \in (0, 1]$ is an important parameter. The minimax risk increases as r decreases as we need to estimate the future observation density based on relatively undersampled past observations (compared to the future), and so the difficulty of the density estimation problem increases [35].

By Theorem 2.1 of [43] it follows that in such regimes under ℓ_0 (along with boundary constraints) or strong or weak ℓ_p constraints on Θ_n for any $p > 0$, the minimax average KL risk for non-parametric function estimation is

$$\bar{\mathcal{R}}_{n,m}^{\mathsf{M}}(\mathcal{F}) = r\, n^{-1} \rho_{n,m}^{\mathsf{M}}(\Theta_n)(1 + o(1)) \text{ as } n \to \infty. \tag{3}$$

Also, if $\hat{p}(\boldsymbol{Y}|\boldsymbol{X})$ is an asymptotically minimax optimal prde in model (2), then the corresponding minimax optional prde in model (1) is:

$$\hat{\mathfrak{p}}(\tilde{\mathbb{D}}|\mathbb{D}) = \hat{p}(\boldsymbol{Y}|\boldsymbol{X}) \cdot p(\boldsymbol{Z}|\boldsymbol{0}, r_{n,m})\, r_{n,m}^{-m/2}\, \sigma^{-m}\ , \tag{4}$$

where, $m = m_n$, $\boldsymbol{Z}$ is a $(m - n)$ length vector and $\boldsymbol{Z} = \sigma_n^{-1} m^{-1} \mathbb{C}' \tilde{\mathbb{D}}$, where $\mathbb{C}_{ij} = \mathfrak{u}_{j+n}(\mathfrak{s}_k)$, $j = 1, \cdots, m - n$; $k = 1, 2, \cdots, m$. Due to this asymptotic connection between predictive density estimation in the non-parametric regression model (1) and the Gaussian sequence model (2), we concentrate on minimax adaptive estimation in (2). Define the asymptotic minimax risk $\rho_{n,m}^{\mathsf{M}}(\Theta_n)$ over $\Theta_n = \ell_0[\eta_n]$, $\ell_p[\eta_n]$ and $\ell_p^*[\eta_n]$ by $\rho_0(r, \eta_n)$, $\rho_p(r, \eta_n)$ and $\rho_p^*(r, \eta_n)$ respectively.

Minimax Risk. We establish the exact first order asymptotic minimax risk over

strong and weak ℓ_p balls with $p \in [0, \infty)$. These results supplement Theorem 1A of [35] which evaluated the minimax risk under exact sparsity and now follows as a special case of Theorem 1 below. As witnessed in sparse point estimation theory [8, 10], here too we observe phase transition in the minimax risk under asymptotically high sparsity levels as shape of ℓ_p ball varies over 0 to ∞. For quadratically convex parametric sets (ℓ_p balls with $p \geq 2$) the asymptotic minimax risk equals the linear predictive minimax risk in [43], but it is much lower than the linear minimax risk when $p < 2$.

Unlike the point estimation (PE) (see Theorem 13.3 of [23]), predictive minimax risk here heavily depend on the under sampling ratio r. In PE, the multivariate least favorable prior is exchangeable block prior (across dimensions) with a spike of length t_n in each block where $t_n^2 = (2 \log \eta_n^{-p}) \wedge (2 \log n)$. Under high sparsity, the number of blocks κ_n can be finite which complicates exact enumeration as a fractional spike is used in one of the blocks instead of a complete t_n length spike. In the predictive regime, the length of the spikes depend on the under sampling ratio r and is reduced to $v^{-1/2} t_n$ where $v = 1/(1 + r^{-1})$. Correspondingly the number of such spikes in the least favorable prior is $\kappa_n = n v^{-p/2} t_n^{-p} \eta_n^p$. Following Zhang [45] define the highly sparse (HS) regime where the mean radius η_n of the parametric space Θ_n is not large enough such that $\eta_n / (\log^3 n/n)^{1/p} \to \infty$ as $n \to \infty$. Note that, whenever $\Theta_n \notin$ HS, $\kappa_n \to \infty$. The HS regime needs separate analysis. Define $R(x) = [x] + \{x\}^{2/p}$ where $[x]$ is the integral and $\{x\}$ is the positive fractional part of x. This is involved in predictive minimax risk for HS regimes as shown below.

Theorem 1.1 (Minimax Risk)**.** *The minimax predictive risk when $\eta_n \to 0$ and $r_n \to r \in (0, \infty)$ as $n \to \infty$ in the non-parametric regression framework of* (1)-(4) *is*

A. Under Exact ℓ_0 Sparsity. $\rho_0(r, \eta_n) \sim (1+r)^{-1} \, n \, \eta_n \log \eta_n^{-1}$

B. Under Strong ℓ_p sparsity. For $p \geq 2$, $\rho_p(r, \eta_n) \sim (2r)^{-1} n \eta_n^2$ and for all $p \in (0, 2)$, $\rho_p(r, \eta_n) \sim (1 + r)^{-1} R(\kappa_n) \log(n/\check{k}_n)$ where $\check{\kappa}_n = \kappa_n \vee 1$. As such when $\kappa_n \to \infty$,

$$\rho_p(r, \eta_n) \sim n \eta_n^p \left(\frac{\log \eta_n^{-1}}{1 + r} \right)^{1 - \frac{p}{2}} \left(\frac{1}{2r} \right)^{\frac{p}{2}} \quad \text{for } p \in (0, 2).$$

C. Under weak ℓ_p sparsity. For all $p > 2$, $\rho_p^*(r, \eta_n) \sim p(p-2)^{-1} \rho_p(r, \eta_n)$. For $p \in (0, 2)$, $\bar{\rho}_p^*(r, \eta_n) \sim \sum_{i=1}^n \min(\eta_n^2 (n/i)^{2/p}, v t_n^2)$. Additionally, if η_n is such that we are not in the HS regime, then it further simplifies as $\rho_p^*(r, \eta_n) \sim 2 \, (2 - p)^{-1} \, \rho_p(r, \eta_n)$ for any $0 < p < 2$.

The minimax risk increases as r decreases. However, the rate of convergence of

the predictive risk with n does not depend on r, and so exact determination of the constants is needed to show the role of under sampling in this prediction problem. For example, if Θ_n is contained in a ℓ_p ball with $p \in [0, \infty)$ and mean radius $\eta_n \to 0$ but not fast enough so that we are in a HS regime, then the optimal risk of plug-in prdes is $(1 + r^{-1})^{1-p/2}$ higher than the minimax risk. Note that when $p = 0$ we get back the plug-in inefficiency calculated in page 4 of [35].

The proof of the theorem is presented in the Appendix. The proof uses the technical arguments developed for PE in Ch 13.2 of [23], [22] and [45] along the connecting equation in Theorem 2 of George *et al.* [17] that provides a tractable evaluation of the KL predictive risk of Bayes prde through an integral representation of the quadratic risk of the posterior mean.

1.2 Risk properties of Spike-and-Slab prdes

Spike-and-Slab prdes. We construct prdes in the sequence model and would transform them via (4) for usage in the regression set-up. Consider a univariate prior which a mixture of an atom of probability at 0 (spike) and a density γ (slab),

$$\pi[\gamma, \omega](\theta) = (1 - \omega)\delta_0(\theta) + \omega\gamma(\theta) \ . \tag{5}$$

Denote its prde by $\hat{p}[\gamma, \omega]$. Consider multivariate priors $\pi[\gamma, \omega](d\boldsymbol{\theta})$ based on i.i.d. components of such spike-and-slab priors: $\pi[\gamma, \omega](d\boldsymbol{\theta}) = \prod_{i=1}^n \pi[\gamma, \omega](d\theta_i)$ and the resultant prde $\hat{p}[\gamma, \omega](\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^n \hat{p}[\gamma, \omega](y_i|x_i)$.

We consider three explicit choices of γ: the Gaussian, the exponential and the fat tailed quasi-cauchy density

$$\gamma_\tau^{\mathsf{G}}(\theta) = \phi(\theta; 0, \tau),\ \gamma_a^{\mathsf{E}}(\theta) = \frac{a}{2}\exp(-a|\theta|) \quad \text{and } \gamma^{\mathsf{C}}(\theta) = \phi(0)\left(1 - |\theta| \cdot \frac{\tilde{\Phi}(|\theta|)}{\phi(\theta)}\right) \ .$$

The scale parameters σ and τ of the Gaussian and exponential slabs are positive. γ^{C} is the scale mixture of normals whose scale follows beta-prime distribution (see section 2 of [38]) and eqn (6) of [25]). As such, $\gamma^{\mathsf{C}}(\theta) = \int_0^\infty \sigma^{-1}\phi(\sigma^{-1}\theta)h(\sigma)\, d\sigma$ where h has Beta-prime$(1, 1/2)$ density. The tails of γ has the same weight as those of the Cauchy density. Mixture priors related to (5), particularly those with Gaussian tails [1, 6] are very popular in wavelet regression [42]. Let $\hat{p}_{\mathsf{S}}[\gamma, \omega]$ be the univariate prde based on $\pi[\gamma, \omega]$. Then,

$$\hat{p}[\gamma, \omega](Y|X) = \phi(X)\,\phi(Y; 0, r)\,\frac{(1 - \omega) + \omega\,\{\phi(0)\}^{-1}e^{W^2/2}\int_{-\infty}^{\infty}\phi(W - v^{-1/2}\theta)\,\gamma(\theta)\,d\theta}{(1 - \omega) + \omega\,\phi(X)g(X)/\phi(0)}$$

where, $W = (1+1/r)^{-1/2}(X + Y/r)$ and $v = (1+r^{-1})^{-1}$; g is the marginal density: $g(x) = \int \phi(x;\theta,1)\gamma(\theta)\,d\theta$. The marginals for the aforementioned slabs have closed form expressions (Sec. 2.3 of 27) with $g_\tau^{\mathsf{G}}(x) = \phi(x;0,1+\tau)$, $g_a^{\mathsf{E}}(x) = 2^{-1}a(e^{-ax}\Phi(x-a) + e^{ax}\tilde{\Phi}(x+a))$ and $g^{\mathsf{C}}(x) = \phi(0)\,x^{-2}(1-e^{-x^2/2})$ where Φ is the standard normal cdf and $\tilde{\Phi} = 1-\Phi$. Note that, while $\phi(x;\mu,\sigma^2)$ is used to denote the normal density function at x for mean μ and variance σ^2 we use $\phi(x)$ dropping the extra suffixes for standard normal density.

A class of Spike-and-Slab prdes. Consider symmetric slabs γ with exponential or heavier tails, i.e.,

$$\sup_{u>0}\left|\frac{d}{du}\log\gamma(u)\right| = \Lambda < \infty\ .$$

Also, assume that the tails of γ are no heavier than Cauchy: $\sup_{u>0} u^2\gamma(u) < \infty$ and its tail probabilities obey the following benign condition with $\kappa \in [1,2]$:

$$\int_y^\infty \gamma(u)\,du \asymp y^{\kappa-1}\gamma(y) \text{ as } y \to \infty$$

If γ has asymptotically exponential tails, then $\kappa = 1$ and any Pareto tail behavior corresponds to $\kappa = 2$. Denote the set of slabs obeying the above three properties by $\mathcal{A}$. Let $\mathcal{P}_{\mathcal{A}}$ be the class of all spike-and-slab prdes based on slabs in $\mathcal{A}$. Before proceeding further, we state the following notations that are used through out the paper.

Notations. The notation $a_n \sim b_n$ denotes $a_n/b_n \to 1$ as $n \to \infty$. We have used 'Big O' and 'Small o' in accordance to their standard definition. $a_n \gtrsim b_n$ denotes $a_n \geq b_n(1+o(1))$. $a_n \asymp b_n$ denotes $c_1 a_n \leq b_n \leq c_2 a_n$ for all large n where c_1 and c_2 are constants. Φ denotes standard normal cdf and $\tilde{\Phi} = 1-\Phi$. Also, recall $v = (1+r^{-1})^{-1}$ where $r = \lim_{n\to\infty} n/m_n$. Fraktur font is used in the non-parametric regression model but not for the sequence model (2) on the transformed basis.

Minimax optimality of $\mathcal{P}_{\mathcal{A}}$ under known sparsity. Our next result shows that if the nature of sparsity p and its level η_n are known, then any spike-and-slab prde in $\mathcal{P}_{\mathcal{A}}$ when appropriately tuned is asymptotically minimax optimal for all sparse regimes. To facilititate combined presentation of exact and approximate sparsity results, we define $\tilde{p} = p \wedge 2$ if $p > 0$, and 1 if $p = 0$. For optimality, we set the non-zero mass hyper-parameter for any prde $\hat{p} \in \mathcal{P}_{\mathcal{A}}$, at $\eta_n^{\tilde{p}}$.

Theorem 1.2 (Asymptotic minimaxity of $\mathcal{P}_{\mathcal{A}}$ under known sparsity)**.** *If $\eta_n \to 0$ as*

$n \to \infty$ in non-parametric regression framework of (1)-(4) *with $r_n \to r \in (0, \infty)$, for all spike-and-slab* prde *$\hat{p} \in \mathcal{P}_{\mathcal{A}}$ and $0 \leq p < \infty$ we have,*

$$\sup_{\gamma \in \mathcal{A}} \sup_{\boldsymbol{\theta} \in \ell_p[\eta_n]} \rho(\boldsymbol{\theta}, \hat{p}[\gamma, \eta_n^{\tilde{p}}]) \sim \rho_p(r, \eta_n) \quad \text{as } n \to \infty,$$

and $\hat{p}[\gamma, \eta_n^{\tilde{p}}]$ is also minimax optimal over $\ell_p^*[\eta_n]$ for all γ in $\mathcal{A}$.

Mukherjee and Johnstone [36] showed that compared to sparse point estimation (PE) the geometry of minimax optimal discrete priors vastly differed in the predictive density estimation probalem. Also, spike-and-slab priors with uniform slabs fail to obtain minimaxity in ℓ_0 sparse unbounded parametric spaces. Theorem 1.2 above shows that despite such glaring differences in minimax decision theoretic phenomena between PE and prde, popularly used spike-and-slab priors in PE [5, 26] when appropriately tuned enjoy desired decision theoretic guarantees in sparse predictive regime.

Risk of Gaussian slab based prdes. Note that, spike-and-slab prdes $\hat{p}_{\mathsf{G}}$ using Gaussian slabs are not included in $\mathcal{P}_{\mathcal{A}}$. The following result shows spike and slab prde $\hat{p}_{\tau}^{\mathsf{G}}$ based on Gaussian slabs with any fixed variance τ is minimax sub-optimal. For large τ, they are however much efficient than all plug-in predictors. Also, if the variance of the slab is allowed to diverge based on the sparsity level such as $\tau_n = -\log \eta_n$ then $\hat{p}_{\tau_n}^{\mathsf{G}}$ can attain minimax optimality.

Theorem 1.3 (Gaussian Slabs)**.** *If $\eta_n \to 0$ as $n \to \infty$ for any fixed $\tau > 0$ the maximal risk of any spike and slab* prde *based on a Gaussian slab γ of mean 0 and variance τ is*

$$\sup_{\boldsymbol{\theta} \in \ell_0[\eta_n]} \inf_{w \in (0,1)} \rho(\boldsymbol{\theta}, \hat{p}[\gamma, w]) \sim \left(1 + \frac{1}{(1+r)^{-1} + \tau/r}\right) \rho_0(r, \eta_n) \text{ as } n \to \infty \ .$$

Consider $\gamma_n = N(0, \tau_n)$ with $\tau_n = \log \eta_n^{-1}$ then for any $p \in [0, 2)$

$$\sup_{\boldsymbol{\theta} \in \ell_p[\eta_n]} \rho(\boldsymbol{\theta}, \hat{p}[\gamma_n, \eta_n^{\tilde{p}}]) \sim \rho_p(r, \eta_n) \text{ as } n \to \infty \ .$$

1.3 Adaptive Predictive Density Estimation using SASA

For adapting to unknown sparsity, we calibrating any member of $\mathcal{P}_{\mathcal{A}}$ by estimating the hyper-parameter ω as follows. SASA is the class of all such calibrated prdes.

Empirical Bayes tuning. We estimate the hyper-parameter ω in our mixture prior

by EBMLE $\hat{\omega}_{\mathsf{EB}}$ which maximizes the marginal log likelihood

$$\ell(\omega) = \sum_{i=1}^{n} \log\{(1-\omega)\phi(x_i) + \omega g(x_i)\} \quad \text{such that} \quad 2\log n/n \le \omega \le 1. \tag{6}$$

Note that, there always exists a unique solution $\hat{\omega}_{\mathsf{EB}}(\gamma)$ for every $\gamma \in \mathcal{A}$. Perhaps the most natural calibration is to use $\hat{p}_{\mathsf{EB}}[\gamma] = \hat{p}[\gamma, \hat{\omega}_{\mathsf{EB}}]$. Other approaches for such hyper-parameter selection has been studied in [1, 5, 15]. In particular, for function estimation, Johnstone and Silverman [25] suggested using thresholds induced by the posterior median of (5). Consider $\zeta = \beta^{-1}(\omega^{-1})$ where $\beta(x) = g(x)/\phi(x) - 1$ and β^{-1} is the positive inverse of β defined on $[\beta(0), \infty)$. Then, $\hat{\zeta}(\gamma) = \beta^{-1}(\hat{\omega}_{\mathsf{EB}}^{-1})$ is the asymptotic threshold choice of [25] (defined as the pseudo-threshold in sec 5.4 of [25]). Define $\hat{\omega}_{\mathsf{JS}}(\gamma) = \exp(-\hat{\zeta}^2/2)$. We consider using this calibration for any γ in $\mathcal{A}$ and propose using $\hat{p}_{\mathsf{JS}}[\gamma] = \hat{p}[\gamma, \hat{\omega}_{\mathsf{JS}}]$. We establish asymptotic optimality of such prdes in the next section. Note that, for the class $\mathcal{A}$, $\hat{\omega}_{\mathsf{JS}}$ is a conservative estimate of sparsity compared to $\hat{\omega}_{\mathsf{EB}}$. Though we do not prescribe to use it for tuning spike-and-slab, it should be noted that $\hat{\omega}_{\mathsf{JS}}$ is no longer conservative for Gaussian slabs. Also, for different member in $\mathcal{A}$, the difference between $\hat{\omega}_{\mathsf{JS}}$ and $\hat{\omega}_{\mathsf{EB}}$ may vary substantially with the gap being asymptotically much closer for fatter tails than exponential tails.

Lemma 1.1. *There exists $\epsilon > 0$ such that for all $\hat{\omega}_{\mathsf{EB}} < \epsilon$,*

$$\hat{\omega}_{\mathsf{JS}} < \hat{\omega}_{\mathsf{EB}} \text{ for any } \gamma \in \mathcal{A} \ .$$

In particular, as $\hat{\omega}_{\mathsf{EB}} \to 0$, for any γ_a^{E} with $a > 0$, we have $\hat{\omega}_{\mathsf{JS}} \asymp \exp(-2^{-1}\hat{\omega}_{\mathsf{EB}}^{-2})$ and for γ^{C} we have $\hat{\omega}_{\mathsf{JS}}/\hat{\omega}_{\mathsf{EB}} \asymp 1/\log(\hat{\omega}_{\mathsf{JS}})$.
However for any Gaussian slab γ^{G} there exists $\epsilon > 0$ such that for all $\hat{\omega}_{\mathsf{EB}} < \epsilon$, $\hat{\omega}_{\mathsf{JS}} > \hat{\omega}_{\mathsf{EB}}$.

Note that, other hyper-parameters such as the scale parameter a in γ^{E} can be simultaneously optimized in (6). Also, once we have a calibrated prde $\hat{p}$ it will be transformed via (4) and the corresponding $\hat{\mathfrak{p}}$ will be used in non-parametric regression examples.

We next show that prdes in SASA are minimax adaptive. The theorem below shows that apart from a poly-log error term, for all $0 \le p < \infty$ and any $r \in (0, \infty)$ these predictors uniformly attain exact minimax optimality in the sense that their inefficiency compared to the minimax rates is bounded and independent of r. Note that, the minimax predictive rates heavily depend on the undersampling ratio r (see Theorem 1.1) and even plugin estimators are rate-optimal (but not optimal in r).

Thus, it is important to show that the proposed predictors has maximal error of the order of minimax risk with constants being independent of r. Thus, prdes in SASA automatically adapt to the sparseness of the underlying signal and in turn to the smoothness of the function. Also, the risk of these predictors is uniformly bounded over all functions suggesting that the they can be used without much harm even if the function classes were mis-specified.

Theorem 1.4 (Adaptive Minimaxity of SASA). *For any $\gamma \in \mathcal{A}$ there exists constants $C_0 := C_0(\gamma)$, $C_1 := C_1(\gamma, p)$ and $C_2 := C_2(\gamma, p)$ independent of r such that,*
[a] Uniformly bounded risk.

$$\sup_{\boldsymbol{\theta}} n^{-1}\rho(\boldsymbol{\theta}, \hat{p}[\gamma, \hat{\omega}_{\mathsf{JS}}]) \leq C_0 \quad \text{as } n \to \infty.$$

[b] Adaptivity to sparse signals. For all $p \in [0, \infty)$ and $\eta_n \leq \eta_0(\gamma, p)$, $n \geq n_0(\gamma, p)$

$$\sup_{\boldsymbol{\theta} \in \ell_p[\eta_n]} \rho(\boldsymbol{\theta}, \hat{p}[\gamma, \hat{\omega}_{\mathsf{JS}}]) \leq C_1\rho_p(r, \eta_n) + C_2(\log n)^{3-\tilde{p}/2}.$$

1.4 Discussion

In sparse sequence models Mukherjee and Johnstone [35] prescribed a thresholding based minimax optimal prde that uses a discrete cluster prior below the threshold. Theorem 1.4 of Mukherjee and Johnstone [36] suggests that thresholded prde based on a spike and continuous uniform-slab prior is also minimax optimal. However, thresholding will be needed as uniform slabs are highly suboptimal for parametric spaces with unrestricted growth. Non-parametric regression is connected with the sequence set-up through the inversion step in (4) and so, computations in non-parametric model based on thresholded prde in sequence models are difficult. The adaptive Bayes procedures prescribed here ameliorate thresholding related computational challenges. The prdes presented here are based on mixture priors with an atom of probability at 0. These priors are popularly referred to as *point-mass* spike-and-slab priors [39]. Recently, continuous approximations to such priors, which further reduces compuational burden, are increasingly being used for sparse estimation [3, 21, 37, 39–41]. It will be interesting to intropect the predictive properties of these continuous spike-and-slab procedures. Also, the results presented here are for non-parametric regression based on non-local basis. Extending these predictive results for multi-scale basis functions [5, 27] will be useful.

1.5 Supplement

The proofs of Theorems 1.1, 1.2, 1.3 and 1.4 and Lemma 1.1 are provided in the supplement.

References

[1] Abramovich, F., Sapatinas, T., and Silverman, B. W. (1998). Wavelet thresholding via a bayesian approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **60**(4), 725–749.

[2] Aitchison, J. and Dunsmore, I. R. (1975). *Statistical prediction analysis*. Cambridge University Press.

[3] Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015). Dirichlet–laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, **110**(512), 1479–1490.

[4] Brown, L. D., George, E. I., and Xu, X. (2008). Admissible predictive density estimation. *Ann. Statist.*, **36**(3), 1156–1170.

[5] Clyde, M. and George, E. I. (2000). Flexible empirical bayes estimation for wavelets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **62**(4), 681–698.

[6] Clyde, M., Parmigiani, G., and Vidakovic, B. (1998). Multiple shrinkage and subset selection in wavelets. *Biometrika*, **85**(2), 391–401.

[7] Donoho, D. L. and Johnstone, I. M. (1994a). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**(3), 425–455.

[8] Donoho, D. L. and Johnstone, I. M. (1994b). Minimax risk over l_p-balls for l_q-error. *Probab. Theory Related Fields*, **99**(2), 277–303.

[9] Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.*, **90**(432), 1200–1224.

[10] Donoho, D. L. and Johnstone, I. M. (1996). Neo-classical minimax problems, thresholding and adaptive function estimation. *Bernoulli*, pages 39–62.

[11] Donoho, D. L., Johnstone, I. M., Hoch, J. C., and Stern, A. S. (1992). Maximum entropy and the nearly black object. *J. Roy. Statist. Soc. Ser. B*, **54**(1), 41–81. With discussion and a reply by the authors.

[12] Efromovich, S. (1999). *Nonparametric curve estimation*. Springer Series in Statistics. Springer-Verlag, New York. Methods, theory, and applications.

[13] 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.

[14] Geisser, S. (1993). *Predictive inference*, volume 55 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, New York. An introduction.

[15] George, E. and Foster, D. P. (2000). Calibration and empirical bayes variable selection. *Biometrika*, **87**(4), 731–747.

[16] George, E. I. and Xu, X. (2008). Predictive density estimation for multiple regression. *Econometric Theory*, **24**(2), 528–544.

[17] George, E. I., Liang, F., and Xu, X. (2006). Improved minimax predictive densities under Kullback-Leibler loss. *Ann. Statist.*, **34**(1), 78–91.

[18] George, E. I., Liang, F., and Xu, X. (2012). From minimax shrinkage estimation to minimax shrinkage prediction. *Statist. Sci.*, **27**(1), 82–94.

[19] 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.

[20] Härdle, W., Kerkyacharian, G., Picard, D., and Tsybakov, A. (2012). *Wavelets, approximation, and statistical applications*, volume 129. Springer Science & Business Media.

[21] Ishwaran, H. and Rao, J. S. (2005). Spike and slab gene selection for multigroup microarray data. *Journal of the American Statistical Association*, **100**(471), 764–780.

[22] Johnstone, I. M. (1994). 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.

[23] Johnstone, I. M. (2013). Gaussian estimation: Sequence and wavelet models. Version: 11 June, 2013. Available at `"http://www-stat.stanford.edu/~imj"`.

[24] Johnstone, I. M. and Silverman, B. W. (1997). Wavelet threshold estimators for data with correlated noise. *Journal of the royal statistical society: series B (statistical methodology)*, **59**(2), 319–351.

[25] Johnstone, I. M. and Silverman, B. W. (2004a). Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences. *Ann. Statist.*, **32**(4), 1594–1649.

[26] Johnstone, I. M. and Silverman, B. W. (2004b). Needles and straw in haystacks: Empirical bayes estimates of possibly sparse sequences. *Annals of Statistics*, pages 1594–1649.

[27] Johnstone, I. M. and Silverman, B. W. (2005). Empirical Bayes selection of wavelet thresholds. *Ann. Statist.*, **33**(4), 1700–1752.

[28] Kato, K. (2009). Improved prediction for a multivariate normal distribution with unknown mean and variance. *Annals of the Institute of Statistical Mathematics*, **61**(3), 531–542.

[29] Komaki, F. (2001). A shrinkage predictive distribution for multivariate normal observables. *Biometrika*, **88**(3), 859–864.

[30] 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.

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

[32] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Statistics*, **22**, 79–86.

[33] Maruyama, Y. and Ohnishi, T. (2016). Harmonic bayesian prediction under alpha-divergence. *arXiv preprint arXiv:1605.05899*.

[34] Mukherjee, G. (2013). *Sparsity and Shrinkage in Predictive Density Estimation*. Ph.D. thesis, Stanford University. Available at "`http://purl.stanford.edu/gm306wz2890`".

[35] Mukherjee, G. and Johnstone, I. M. (2015). Exact minimax estimation of the predictive density in sparse gaussian models. *Annals of Statistics*.

[36] Mukherjee, G. and Johnstone, I. M. (2017). On minimax optimality of sparse bayes predictive density estimates. *arXiv preprint arXiv:1707.04380*.

[37] Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, **103**(482), 681–686.

[38] Polson, N. G., Scott, J. G., *et al.* (2012). On the half-cauchy prior for a global scale parameter. *Bayesian Analysis*, **7**(4), 887–902.

[39] Rocková, V. (2015). Bayesian estimation of sparse signals with a continuous spike-and-slab prior. *Submitted manuscript*, pages 1–34.

[40] Rocková, V. and George, E. (2014). The spike-and-slab lasso. *Manuscript in preparation*.

[41] Ročková, V. and George, E. I. (2014). Negotiating multicollinearity with spike-and-slab priors. *Metron*, **72**(2), 217–229.

[42] Vidakovic, B. (1998). Wavelet-based nonparametric bayes methods. In *Practical nonparametric and semiparametric Bayesian statistics*, pages 133–155. Springer.

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

[44] Xu, X. and Zhou, D. (2011). Empirical bayes predictive densities for high-dimensional normal models. *J. Multivariate Analysis*, **102**(10), 1417–1428.

[45] Zhang, C.-H. (2012). Minimax ℓ_q risk in ℓ_p balls. In *Contemporary Developments in Bayesian Analysis and Statistical Decision Theory: A Festschrift for William E. Strawderman*, pages 78–89. Institute of Mathematical Statistics.