Asymptotically minimax empirical Bayes estimation of a sparse normal mean vector

Ryan Martin

Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago, 851 S. Morgan St., Chicago, IL 60607, USA
e-mail: rgmartin@uic.edu

and

Stephen G. Walker

Department of Mathematics, University of Texas at Austin, 2515 Speedway Stop C1200, Austin, TX 78712
e-mail: s.g.walker@math.utexas.edu

Abstract: For the important classical problem of inference on a sparse high-dimensional normal mean vector, we propose a novel empirical Bayes model that admits a posterior distribution with desirable properties under mild conditions. In particular, our empirical Bayes posterior distribution concentrates on balls, centered at the true mean vector, with squared radius proportional to the minimax rate, and its posterior mean is an asymptotically minimax estimator. We also show that, asymptotically, the support of our empirical Bayes posterior has roughly the same effective dimension as the true sparse mean vector. Simulation from our empirical Bayes posterior is straightforward, and our numerical results demonstrate the quality of our method compared to others having similar large-sample properties.

1. Introduction

High-dimensional problems, where the parameter is effectively lower-dimensional, are now commonplace in statistical applications. Examples include variable selection in regression (Fan and Lv, 2010), covariance matrix estimation (Cai, Zhang and Zhou, 2010; Cai and Zhou, 2012; Lam and Fan, 2009), large-scale multiple testing (Bogdan *et al.*, 2011; Cai and Jin, 2010), and function estimation (Cai, 2012; Johnstone and Silverman, 2005). The canonical example, which we shall consider here, is that of estimating a sparse high-dimensional normal mean vector. Let $X_1, \dots, X_n$ be independent observations, with $X_i \sim \mathsf{N}(\theta_i, 1)$,

$i = 1, \ldots, n$, and the goal is to estimate the mean vector $\theta = (\theta_1, \ldots, \theta_n)$ under squared-error loss $\|\hat{\theta} - \theta\|^2$, where $\| \cdot \|$ is the usual ℓ_2-norm on $\mathbb{R}^n$ (e.g., Abramovich *et al.*, 2006; Brown and Greenshtein, 2009; Castillo and van der Vaart, 2012; Donoho and Johnstone, 1994; Donoho *et al.*, 1992; Jiang and Zhang, 2009). With only a single observation X_i for each θ_i, accurate estimation is not possible without some structure. Assuming θ is sparse, in the sense that most of the θ_i's are zero, makes the effective dimension relatively small so that reasonably accurate estimation becomes possible.

This normal means model is by now a classic one which has been widely studied from both a mathematical and applied point of view. Despite the extent to which the many-normal-means model has been studied, it is still a practically important model in a variety of problems. For example, the sparse normal mean model is the cornerstone for many modern Bayes and empirical Bayes multiple testing procedures, e.g., Scott and Berger (2006), Jin and Cai (2007), Bogdan, Ghosh and Tokdar (2008), Efron (2008), and Martin and Tokdar (2012). More recently, Scott *et al.* (2013) have presented a novel use of the same classical model considered here but in the regression setting. Clearly, research on this classical model is alive and well, and the results provided by our unique approach, namely, asymptotically minimax concentration rates and superior finite-sample performance compared to many existing methods, are useful contributions.

Recently, Castillo and van der Vaart (2012) have considered the performance of several Bayesian methods for this problem. They focus on frequentist properties of a Bayesian posterior distribution, and the corresponding Bayes estimators, for priors with a two-groups structure. In sparse estimation problems, a two-groups prior puts positive probability on θ vectors with some exact zero entries, so the marginal prior for each component is a mixture of a continuous distribution and a point-mass at zero. Castillo and van der Vaart (2012) show that, for a suitably chosen two-groups prior, the posterior concentrates around the true signal at the asymptotically optimal minimax rate. From this, concentration properties of posterior quantities, such as the posterior mean, can be derived. An important message in their paper is that care is needed in choosing the prior for the non-zero θ entries. In particular, they show that priors with too light tails, e.g., Gaussian, give sub-optimal concentration properties. The results presented herein provide similar guidance, though our perspective is different.

Here we take a novel empirical Bayes approach. In particular, we present a hierarchical two-groups prior where, given a weight ω, the θ_i's are modeled as independent, with $\theta_i = 0$ with probability $g_i(\omega)$, and $\theta_i \sim h_i(\theta \mid \omega)$ with probability $1 - g_i(\omega)$, where the functions g_i and h_i depend on data X_i. These functions are defined explicitly in Section 2. To complete the hierarchy, ω is assigned a prior concentrated near 1. We argue that the effect of the data-dependent prior is mitigated by preventing the posterior from tracking the data too closely. This approach provides some new insights, which we compare to those coming from the approach of Castillo and van der Vaart (2012).

In Section 3 we present our theoretical framework. First, we show that our empirical Bayes posterior concentrates, with probability 1, around the true mean

vector at the optimal minimax rate (with respect to square error loss) for the assumed sparsity class. Concentration rate theorems for empirical Bayes posteriors are relatively scarce in the literature, and our technique for handling the challenges that arise from data appearing in both the likelihood and prior might be useful in other problems; one possible extension is discussed briefly in Section 5. We then show that our empirical Bayes posterior mean is an asymptotically minimax estimator of θ. Finally, we show that, asymptotically, the support of our empirical Bayes posterior has, up to a logarithmic factor, the same effective dimension as the true sparse θ. An interesting observation is that the particular form of the prior on ω is the main catalyst for concentration of our empirical Bayes posterior.

Section 4 describes computation of our empirical Bayes posterior mean via a straightforward Markov chain Monte Carlo. Simulation results are presented to show that our empirical Bayes posterior mean generally outperforms those Bayesian and non-Bayesian competitors with comparable large-sample properties. In particular, we compare our method with a two hard thresholding estimators (Donoho and Johnstone, 1994), Bayes and empirical Bayes estimators based on priors with a two-groups structure (Castillo and van der Vaart, 2012; Johnstone and Silverman, 2004), and a new estimator based on the one-group Dirichlet–Laplace prior (Bhattacharya *et al.*, 2014). Our proposed empirical Bayes estimator is competitive in all cases considered here, and is often strikingly better than the others. Concluding remarks are given in Section 5.

2. An empirical Bayes model

For the independent normal mean model, $X_i \sim \mathsf{N}(\theta_i, 1)$, $i = 1, \ldots, n$, let $p_{\theta_i}(x_i)$ denote the density of X_i, and, for $x = (x_1, \ldots, x_n)$, let $p_\theta^n(x) = \prod_{i=1}^n p_{\theta_i}(x_i)$ denote the corresponding joint density of $X = (X_1, \ldots, X_n)$. Define a data-dependent hierarchical prior Π_X for $\theta = (\theta_1, \ldots, \theta_n)$ as follows. Introduce a parameter $\omega \in (0, 1)$, and take the joint prior distribution for $(\theta_1, \ldots, \theta_n, \omega)$, under Π_X, to have density proportional to

$$\omega^{\alpha n - 1} \prod_{i=1}^n \left\{ \omega e^{\frac{1}{2}(1-\kappa)X_i^2} \delta_0(d\theta_i) + (1 - \omega)\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left[\frac{1-(1-\kappa)\sigma^2}{\sigma^2}\right](\theta_i - X_i)^2} d\theta_i \right\}, \quad (1)$$

where $\alpha > 0$, $\kappa \in (0, 1)$, and $\sigma^2 > 0$ are parameters to be discussed further in Sections 3–4. A representation of this as a genuine empirical Bayes plug-in prior is given in Section 3.1. The dependence of the prior on $(\alpha, \kappa, \sigma^2)$ will not be reflected in our notation. Hence, $g_i(\omega)$ and $h_i(\omega)$, mentioned in Section 1, are given by

$$g_i(\omega) = \frac{\omega e^{-\frac{1}{2}(1-\kappa)X_i^2}}{\omega e^{-\frac{1}{2}(1-\kappa)X_i^2} + (1 - \omega)\{1 - (1 - \kappa)\sigma^2\}^{-1}}$$

and

$$h_i(\cdot \mid \omega) = \mathsf{N}\big(\cdot \mid X_i, \sigma^2\{1 - (1 - \kappa)\sigma^2\}^{-1}\big),$$

respectively. If we denote by $\Pi_X(\theta_1, \ldots, \theta_n, \omega)$ the density (1), then by $\Pi_X(\theta)$ we mean the marginal prior for $\theta = (\theta_1, \ldots, \theta_n)$, i.e.,

$$\Pi_X(\theta) = \int_0^1 \Pi_X(\theta_1, \ldots, \theta_n, \omega) \, d\omega.$$

Observe that if $\sigma^2 < (1-\kappa)^{-1}$, then the prior for θ_i is proper, a mixture of a point mass and a Gaussian centered at X_i. When $\sigma^2 > (1-\kappa)^{-1}$, the prior is improper. In any case, the posterior is proper, so this possible impropriety of the prior is not a concern. In fact, $\sigma^2 = (1-\kappa)^{-1}$ is a critical boundary, corresponding to an improper uniform prior for the non-zero θ_i's; see Section 3.3. The term $\omega^{\alpha n - 1}$ in the joint density, which resembles a beta density, turns out to be critical to the success of our proposed method, both in theory and in implementation.

Given data $X = (X_1, \ldots, X_n)$ from the normal mean model and the empirical Bayes prior distribution Π_X for θ, we could combine these to form an empirical Bayes posterior distribution via Bayes theorem. That is, for a suitable set A in the θ-space, define the probability measure

$$Q_n(A) \equiv Q_{n,X}(A) \propto \int_A p_\theta^n(X) \, \Pi_X(d\theta).$$

We will investigate concentration properties of the empirical Bayes posterior in Section 3. In particular, we show that the empirical Bayes posterior mean derived from Q_n is an asymptotically minimax estimator of θ.

It might seem that our apparent double-use of the data—once in the prior and again in the likelihood—could lead to a posterior Q_n that tracks the data too closely. To see that this is not the case, note that if $|X_i|$ is large, then the prior probability for $\theta_i = 0$, under Π_X, would be rather large. Thus, the prior has an unexpected shrinkage effect, pushing θ_i corresponding to X_i with large magnitude towards zero. On the other hand, an X_i with large magnitude shifts the prior on the non-zero part further from zero, effectively making the tails heavier, to accommodate large signals. These two phenomena suggest that using data in both the prior and the likelihood will not result in a posterior that tracks data too closely. In fact, our theoretical and numerical results demonstrate that the posterior is doing the right thing, namely, concentrating on the true θ.

3. Empirical Bayes posterior asymptotics

3.1. A fractional likelihood perspective

To start, it will help to look at the proposed model from a different perspective. For mathematical convenience, we shift our focus and rewrite the empirical Bayes posterior Q_n using a fractional likelihood. That is, we write $p_\theta^n(X) = p_\theta^n(X)^\kappa p_\theta^n(X)^{1-\kappa}$ and move the $1 - \kappa$ fraction into the prior Π_X defined above. The effect of this is an alternative prior for (θ, ω) of a very simple form:

$$\begin{aligned} \theta_i \mid \omega &\stackrel{\text{ind}}{\sim} \omega \delta_0 + (1 - \omega) \mathsf{N}(X_i, \sigma^2), \quad i = 1, \ldots, n, \\ \omega &\sim \mathsf{Beta}(\alpha n, 1). \end{aligned} \tag{2}$$

To provide some further intuition for the prior (1) in Section 2, we may consider a data-free version of the prior in (2), where the X_i's are replaced by hyperparameters μ_i. The marginal likelihood for $\mu = (\mu_1, \ldots, \mu_n)$, given ω, is

$$\prod_{i=1}^{n} \{ \omega \mathsf{N}(X_i \mid 0, 1) + (1 - \omega) \mathsf{N}(X_i \mid \mu_i, 1) \}.$$

and X_i is clearly the maximum marginal likelihood estimate of μ_i. The use of plug-in estimates for mean hyperparameters was considered in Babenko and Belitser (2010) though in a slightly different context. We get the empirical Bayes prior (1) by plugging in X_i for μ_i and undoing the fractional likelihood.

To make this explicit, we have

$$\exp\left\{ -\frac{\kappa(X_i - \theta_i)^2}{2} \right\} \left[\omega \delta_0(\theta_i) + (1 - \omega) \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{(\theta_i - \mu_i)^2}{2\sigma^2} \right\} \right],$$

which is equivalent to

$$\exp\left\{ -\frac{(X_i - \theta_i)^2}{2} \right\} \left[\omega \exp\left\{ -\frac{(\kappa - 1)(\theta_i - X_i)^2}{2} \right\} \delta_0(\theta_i) \right.$$
$$\left. + (1 - \omega) \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{[1 - (1 - \kappa)\sigma^2](\theta_i - X_i)^2}{2\sigma^2} \right\} \right],$$

with $\mu_i = X_i$.

Within this alternative setup, we introduce independent binary latent variables $I_1, \ldots, I_n$, where $I_i = 1$ if and only if $\theta_i = 0$. Then, given ω, the indicators $I_1, \ldots, I_n$ are independent $\mathsf{Ber}(\omega)$ variables. These indicators characterize the support of the vector θ; in particular, $\sum_{i=1}^{n}(1 - I_i)$ is the number of non-zero θ_i and is distributed as $\mathsf{Bin}(n, 1 - \omega)$. The beta prior for ω is concentrated near 1 for n large, so the support size will tend to be small, consistent with the assumption of sparsity. Castillo and van der Vaart (2012), on the other hand, focus primarily on priors directly on the support size, though this kind of beta–binomial prior is considered in their Example 2.2. We find that direct use of the weight ω is both theoretically and computationally convenient; see Remark 1.

Write this new version (2) of the prior as $\tilde{\Pi}_X$ and express the posterior as

$$Q_n(A) \propto \int_A p_\theta^n(X)^\kappa \, \tilde{\Pi}_X(d\theta).$$

This version of the empirical Bayes posterior is amenable for our asymptotic analysis; see, also Walker and Hjort (2001). The use of pseudo-posteriors, where an inverse temperature parameter plays the role of κ, has been considered in the statistics and machine learning literature (e.g., Dalalyan and Tsybakov, 2008; Jiang and Tanner, 2008; Zhang, 2006), but our context is different.

3.2. Lower bound on the denominator

In the normal mean model, let $\theta^\star$ denote the true mean vector. Assume that $\theta^\star$ is sparse in the sense that most of its entries are zero. To make this more precise,

let $\mathcal{S}^\star \subset \{1, 2, \ldots, n\}$ denote the support of $\theta^\star$, i.e., $\theta_i^\star \neq 0$ if and only if $i \in \mathcal{S}^\star$. Let $s_n = \#\mathcal{S}^\star$ be the cardinality of $\mathcal{S}^\star$, and say that $\theta^\star$ is s_n-sparse. Then by sparse we mean that $s_n \to \infty$ but $s_n = o(n)$ as $n \to \infty$. That is, although $\theta^\star$ is n-dimensional, its effective dimension is actually much smaller.

Start by rewriting the empirical Bayes posterior Q_n once more as

$$Q_n(A) = \frac{\int_A \{p_\theta^n(X)/p_{\theta^\star}^n(X)\}^\kappa \, \tilde{\Pi}_X(d\theta)}{\int_{\mathbb{R}^n} \{p_\theta^n(X)/p_{\theta^\star}^n(X)\}^\kappa \, \tilde{\Pi}_X(d\theta)}. \tag{3}$$

Our overall goal is to show that Q_n concentrates its mass near $\theta^\star$ with $\mathsf{P}_{\theta^\star}$-probability 1. The strategy is to show that the denominator of Q_n is not too small, and the numerator, for sets A_n away from $\theta^\star$, is not too large.

Our first result gives a bound on the denominator of Q_n, like that which obtains from the familiar Kullback–Leibler property (e.g., Barron, Schervish and Wasserman, 1999; Ghosal, Ghosh and Ramamoorthi, 1999; Ghosal, Ghosh and van der Vaart, 2000; Schwartz, 1965; Shen and Wasserman, 2001). This lower bound will be used in Section 3.3 to derive vanishing upper bounds on the Q_n-probability assigned to complements of balls around $\theta^\star$. But besides as a tool for proving other things, the following lemma suggests that our empirical Bayes-style prior is sufficiently concentrated around $\theta^\star$. As Castillo and van der Vaart (2012) show, without suitable prior concentration, the desired posterior concentration is not possible. Therefore, if we associate lower bounds on the denominator of Q_n in (3) with adequate prior concentration, then Lemma 1 says that our prior is sufficiently concentrated around $\theta^\star$.

Lemma 1. *Let D_n be the denominator in (3). If $\theta^\star$ is s_n-sparse, then there exists $\eta \in \mathbb{R}$, depending on $(\kappa, \alpha, \sigma^2)$, such that $D_n > \frac{\alpha}{1+\alpha} \exp\{\eta s_n - 2s_n \log(n/s_n) + o(s_n)\}$ with $\mathsf{P}_{\theta^\star}$-probability 1.*

Proof. Write D_n in terms of the conditional prior $(\theta_1, \ldots, \theta_n) \mid \omega \sim \tilde{\Pi}_{X,\omega}$ and the marginal prior $\omega \sim \tilde{\pi}$ for ω under $\tilde{\Pi}_X$. That is,

$$D_n = \int_0^1 \int_{\mathbb{R}^n} \left\{ \frac{p_\theta^n(X)}{p_{\theta^\star}^n(X)} \right\}^\kappa \tilde{\Pi}_{X,\omega}(d\theta) \, \tilde{\pi}(d\omega)$$

$$= \int_0^1 \prod_{i=1}^n \int_{\mathbb{R}} \left\{ \frac{p_{\theta_i}(X_i)}{p_{\theta_i^\star}(X_i)} \right\}^\kappa \tilde{\Pi}_{X,\omega}(d\theta_i) \, \tilde{\pi}(d\omega).$$

For given ω, the inner expectation involves an average over all configurations of the indicators $(I_1, \ldots, I_n)$ defined in Section 3.1. This average is clearly larger than just the case where the indicators exactly match up with the support $\mathcal{S}^\star$ of $\theta^\star$, times the probability of that configuration. That is,

$$D_n > \int_0^1 \omega^{n-s_n}(1-\omega)^{s_n} \, \tilde{\pi}(d\omega)$$

$$\times \prod_{i \in \mathcal{S}^\star} \int_{\mathbb{R}} e^{\frac{\kappa}{2}\{(X_i - \theta_i^\star)^2 - (X_i - \theta_i)^2\}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(X_i - \theta_i)^2} \, d\theta_i.$$

The term $\omega^{s_n-n}(1-\omega)^{s_n}$ corresponds to the probability for the configuration of $(I_1, \ldots, I_n)$ matching the support $\mathcal{S}^\star$. The integral for $i \in \mathcal{S}^\star$ is the expectation of the normal density ratio for non-zero θ_i with respect to the $\mathsf{N}(X_i, \sigma^2)$ prior. Finally, the product over $i \notin \mathcal{S}^\star$ disappears because $p_0(X_i) = p_{\theta_i^\star}(X_i)$ for $i \notin \mathcal{S}^\star$. To further bound this quantity, first pull out the terms $\exp\{\frac{\kappa}{2}(X_i - \theta_i^\star)^2\}$ in the latter integrand that do not depend on θ_i. Since, by the law of large numbers, $s_n^{-1} \sum_{i \in \mathcal{S}^\star}(X_i - \theta_i^\star)^2 \to 1$, as $n \to \infty$, with $\mathsf{P}_{\theta^\star}$-probability 1, this part contributes a factor $\exp\{\frac{\kappa}{2}s_n + o(s_n)\}$ to the lower bound for D_n. Next,

$$\int_{\mathbb{R}} e^{-\frac{\kappa}{2}(X_i-\theta_i)^2} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\theta_i-X_i)^2} \, d\theta_i = \frac{1}{(1+\kappa\sigma^2)^{1/2}}.$$

So, the remaining product over $i \in \mathcal{S}^\star$ equals $(1+\kappa\sigma^2)^{-s_n/2}$, and we can conclude that the entire product over $i \in \mathcal{S}^\star$ in the lower bound for D_n is itself lower bounded by

$$\exp\left[\frac{s_n}{2}\{\kappa - \log(1 + \kappa\sigma^2)\} + o(s_n)\right].$$

It remains to bound the integral over ω. Since $\pi(d\omega) = \alpha n\omega^{\alpha n-1}\, d\omega$, we have

$$\int_0^1 \omega^{n-s_n}(1-\omega)^{s_n}\, \tilde{\pi}(d\omega) > \alpha n \int_0^{1-s_n/n} \omega^{n-s_n+\alpha n-1}(1-\omega)^{s_n}\, d\omega$$

$$> \left(\frac{s_n}{n}\right)^{s_n} \frac{\alpha n}{n - s_n + \alpha n}\left(1 - \frac{s_n}{n}\right)^{n-s_n+\alpha n}$$

$$> \frac{\alpha}{1+\alpha}\left(\frac{s_n}{n}\right)^{2s_n}\left(1 - \frac{s_n}{n}\right)^{\alpha n}. \tag{4}$$

The last inequality follows since $(1 - b)^{1-b} > b^b$ for small $b > 0$. Next, if we write

$$(1 - s_n/n)^{\alpha n} = \exp[-\alpha n\{-\log(1 - s_n/n)\}],$$

and use the approximation $-\log(1 - x) = x + o(x)$, for $x \approx 0$, then we get a lower bound on the ω-integral of the form:

$$c\exp\{-2s_n \log(n/s_n) - \alpha s_n + o(s_n)\}, \quad n \to \infty,$$

for $c = \alpha/(1 + \alpha) > 0$. Putting these pieces together, gives the lower bound

$$D_n > c\exp\left[\frac{s_n}{2}\{\kappa - 2\alpha - \log(1 + \kappa\sigma^2) + o(1)\} - 2s_n \log(n/s_n)\right].$$

Set $\eta = \frac{1}{2}\{\kappa - 2\alpha - \log(1 + \kappa\sigma^2)\} \in \mathbb{R}$ to complete the proof. $\qquad\square$

3.3. Concentration

In the frequentist problem of estimating a s_n-sparse vector θ under squared ℓ_2-error loss, it is known that the minimax rate is proportional to $\varepsilon_n := s_n \log(n/s_n)$; see Donoho *et al.* (1992). Following Castillo and van der Vaart (2012), our goal
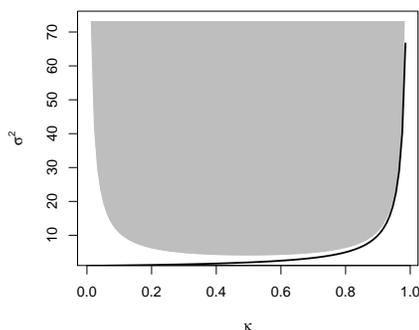
FIG 1. *Portion of the feasible region R_β in (5), with $\beta = 200$, for (κ, σ^2). Solid black line corresponds to the curve $\sigma^2 = (1 - \kappa)^{-1}$.*

here is to show that Q_n concentrates asymptotically on n-balls, centered at $\theta^\star$, with square radius proportional to ε_n. More precisely, for a constant $M > 0$, let

$$A_{M\varepsilon_n} = \{\theta \in \mathbb{R}^n : \|\theta - \theta^\star\|^2 > M\varepsilon_n\};$$

then we will demonstrate that $Q_n(A_{M\varepsilon_n}) \to 0$ with $\mathsf{P}_{\theta^\star}$-probability 1.

The theorem below requires a restriction on (κ, σ^2). In particular, we require that, for some $\beta > 1$, (κ, σ^2) reside in the feasible region

$$R_\beta = \left\{ (\kappa, \sigma^2) : \frac{1}{\sigma^2(1 + \beta/\sigma^2)^{1/\beta}} - \frac{1}{\sigma^2 + \beta} < \frac{\kappa[(1 - \kappa)\beta - 1]}{\beta - 1} \right\}. \qquad (5)$$

We are particularly interested in large β, so that κ arbitrarily close to 1 can be included. Figure 1 displays a portion of the region R_β, for $\beta = 200$. The condition $\sigma^2 = (1 - \kappa)^{-1}$ discussed in Section 2 defines the boundary of R_β, for large β and $\kappa \approx 1$.

Theorem 1. *For any fixed $\beta > 1$, take (κ, σ^2) in the feasible set R_β. If $\theta^\star$ is s_n-sparse, then there exists $M > 0$ such that $Q_n(A_{M\varepsilon_n}) \to 0$ with $\mathsf{P}_{\theta^\star}$-probability 1.*

Proof. Let N_n be the numerator for $Q_n(A_{M\varepsilon_n})$ in (3), i.e.,

$$N_n = \int_0^1 \int_{A_{M\varepsilon_n}} \prod_{i=1}^n \left(\frac{p_{\theta_i}(X_i)}{p_{\theta_i^\star}(X_i)}\right)^\kappa \tilde{\Pi}_{X_i,\omega}(d\theta_i)\, \tilde{\pi}(d\omega).$$

Taking expectation of N_n, with respect to $\mathsf{P}_{\theta^\star}$, we get

$$\mathsf{E}_{\theta^\star}(N_n) = \int_0^1 \int_{A_{M\varepsilon_n}} \prod_{i=1}^n \int_{\mathbb{R}} \left(\frac{p_{\theta_i}(x_i)}{p_{\theta_i^\star}(x_i)}\right)^\kappa \tilde{\Pi}_{x_i,\omega}(d\theta_i) p_{\theta_i^\star}(x_i)\, dx_i\, \tilde{\pi}(d\omega).$$

Write $J_\omega(d\theta_i)$ for the measure defined in the i-th product term. Split this into discrete and continuous pieces:

$$J_\omega(d\theta_i) = \int \left(\frac{p_{\theta_i}(x_i)}{p_{\theta_i^\star}(x_i)}\right)^\kappa \tilde{\Pi}_{x_i,\omega}(d\theta_i) p_{\theta_i^\star}(x_i)\, dx_i$$

$$= \omega\Big\{ \int \Big(\frac{p_{\theta_i}(x_i)}{p_{\theta_i^\star}(x_i)}\Big)^\kappa p_{\theta_i^\star}(x_i)\,dx_i \Big\}\,\delta_0(d\theta_i)$$

$$+ (1-\omega)\Big\{ \int \Big(\frac{p_{\theta_i}(x_i)}{p_{\theta_i^\star}(x_i)}\Big)^\kappa \frac{p_{\theta_i/\sigma}(x_i/\sigma)}{\sigma} p_{\theta_i^\star}(x_i)\,dx_i \Big\}\,d\theta_i.$$

For clarity, we shall work with the discrete and continuous parts separately.

Discrete part. Using the Renyi divergence formula for normal distributions, the discrete term simplifies to $\omega \exp\{-\frac{\kappa(1-\kappa)}{2}(\theta_i - \theta_i^\star)^2\}\,\delta_0(d\theta_i)$.

Continuous part. An application of Hölder's inequality, with coefficients $\frac{\beta}{\beta-1}$ and β, whose reciprocals sum to one, gives

$$\int \Big(\frac{p_{\theta_i}(x_i)}{p_{\theta_i^\star}(x_i)}\Big)^\kappa \frac{p_{\theta_i/\sigma}(x_i/\sigma)}{\sigma} p_{\theta_i^\star}(x_i)\,dx_i$$

$$\leq \Big\{ \int \Big(\frac{p_{\theta_i}(x_i)}{p_{\theta_i^\star}(x_i)}\Big)^{\frac{\kappa\beta}{\beta-1}} p_{\theta_i^\star}(x_i)\,dx_i \Big\}^{\frac{\beta-1}{\beta}} \Big\{ \int \Big(\frac{p_{\theta_i/\sigma}(x_i/\sigma)}{\sigma}\Big)^\beta p_{\theta_i^\star}(x_i)\,dx_i \Big\}^{\frac{1}{\beta}}.$$

For $(\kappa, \sigma^2) \in R_\beta$, we have $\frac{\kappa\beta}{\beta-1} < 1$. Then the same Renyi divergence formula used above gives $\exp\{-\frac{\kappa}{2}\frac{\beta(1-\kappa)-1}{\beta-1}(\theta_i - \theta_i^\star)^2\}$. The second term in the upper bound equals

$$\frac{1}{\sqrt{2\pi\sigma^2}}\Big\{ \frac{\sigma}{(\sigma^2+\beta)^{1/2}} e^{-\frac{\beta}{2(\sigma^2+\beta)}(\theta_i-\theta_i^\star)^2} \Big\}^{1/\beta}.$$

After some tedious algebra, this can be rewritten as

$$\exp\Big\{ \frac{1}{2}\Big(\frac{1}{\sigma^2(1+\beta/\sigma^2)^{1/\beta}} - \frac{1}{\sigma^2+\beta}\Big)(\theta_i - \theta_i^\star)^2 \Big\} \mathsf{N}(\theta_i \mid \theta_i^\star, \sigma^2(1+\beta/\sigma^2)^{1/\beta}).$$

Combining the two terms in the upper bound, ignoring the normal density, gives

$$\exp\Big[-\frac{1}{2}\Big\{ \frac{\kappa[(1-\kappa)\beta - 1]}{\beta-1} - \Big(\frac{1}{\sigma^2(1+\beta/\sigma^2)^{1/\beta}} - \frac{1}{\sigma^2+\beta}\Big) \Big\}(\theta_i - \theta_i^\star)^2 \Big].$$

For (κ, σ^2) in the feasible region R_β in (5), the coefficient on $(\theta_i - \theta_i^\star)^2$ in the exponential term above is negative.

We can now find a constant $c > 0$, depending on $(\kappa, \sigma^2, \beta)$, such that

$$J_\omega(d\theta_i) \leq e^{-c(\theta_i - \theta_i^\star)^2}\{\omega\delta_0(d\theta_i) + (1-\omega)\mathsf{N}(\theta_i \mid \theta_i^\star, \sigma^2(1+\beta/\sigma^2)^{1/\beta})\,d\theta_i\}.$$

Then $J_\omega^n(d\theta) := \prod_{i=1}^n J_\omega(d\theta_i)$ is upper bounded by $\exp\{-c\|\theta - \theta^\star\|^2\}$ times a probability measure in θ on $\mathbb{R}^n$. Therefore, by definition of $A_{M\varepsilon_n}$,

$$\mathsf{E}_{\theta^\star}(N_n) = \int_0^1 \int_{A_{M\varepsilon_n}} J_\omega^n(d\theta)\,\tilde\pi(d\omega) \leq e^{-cM\varepsilon_n}.$$

Next, take M such that $cM > 2$, and then take $K \in (2, cM)$. Then Markov's inequality gives the upper bound

$$\mathsf{P}_{\theta^\star}(N_n > e^{-K\varepsilon_n}) \leq Le^{-(cM-K)\varepsilon_n}.$$

This upper bound has a finite sum over $n \geq 1$, so the Borel–Cantelli lemma gives that $N_n \leq e^{-K\varepsilon_n}$, with $\mathsf{P}_{\theta^\star}$-probability 1 for all large n. Putting together this bound on N_n and the one on D_n from Lemma 1, we get

$$\frac{N_n}{D_n} \leq \frac{1+\alpha}{\alpha} e^{-(K-2)\varepsilon_n - \eta s_n + o(s_n)}. \tag{6}$$

Since $s_n = o(\varepsilon_n)$, the exponent diverges to $-\infty$ regardless of the sign on η. Therefore, $Q_n(A_{M\varepsilon_n}) \to 0$ as $n \to \infty$ with $\mathsf{P}_{\theta^\star}$-probability 1. $\qquad\square$

Remark 1. The ε_n concentration rate is driven primarily by the beta prior on the weight ω. In particular, it comes from the term $(s_n/n)^{2s_n}$ in the lower bound (4) in Lemma 1. This means that the prior for θ, given ω, should be selected so that it does not interfere with the correct rate coming from the lower bound on the denominator of Q_n.

Remark 2. Castillo and van der Vaart (2012) show that the minimax concentration rate will not hold if the prior on non-zero θ has too light of tails, e.g., Gaussian. A way to understand this point, from our perspective, is that the Gaussian conditional prior interferes with what the beta prior for the weight ω is doing. As we have demonstrated, this does not necessarily mean that Gaussian is wrong, but that some adjustments should be made to prevent this interference.

3.4. *Asymptotic minimaxity of the posterior mean*

Since the empirical Bayes posterior concentrates around the right place and the right rate, it ought to produce an estimator of θ with good properties. For this problem, perhaps the most natural choice of estimator is the empirical Bayes posterior mean,

$$\hat{\theta}_n = \int \theta \, Q_n(d\theta)$$

Next we show that $\hat{\theta}_n$ is a minimax estimator if $\theta^\star$ is s_n-sparse.

Theorem 2. *Take (κ, σ^2) as in Theorem 1. If $\theta^\star$ is s_n-sparse, then there exists a universal constant $M' > 0$ such that $\mathsf{E}_{\theta^\star}\|\hat{\theta}_n - \theta^\star\|^2 \leq M'\varepsilon_n$ for all large n.*

Proof. Start by considering the quantity $\int \|\theta - \theta^\star\|^2 \, Q_n(d\theta)$. Split this integral into two via the partition $\mathbb{R}^n = A_{M\varepsilon_n} \cup A_{M\varepsilon_n}^c$ for M as in Theorem 1. On $A_{M\varepsilon_n}^c$, $\|\theta - \theta^\star\|^2$ is bounded above by $M\varepsilon_n$, and $Q_n(A_{M\varepsilon_n}^c) \leq 1$ trivially. So, we immediately get

$$\int_{A_{M\varepsilon_n}^c} \|\theta - \theta^\star\|^2 \, Q_n(d\theta) \leq M\varepsilon_n.$$

For the integration over $A_{M\varepsilon_n}$, we again look at the numerator and denominator of Q_n separately, as in the previous subsection. The denominator has the same lower bound as in Lemma 1. Take n large enough that, with $\mathsf{P}_{\theta^\star}$-probability 1, the lower bound in the lemma holds; then the expectation of the ratio can be bounded by upper bounding the expectation of the numerator, together with the

lower bound on the denominator. Expectation of the numerator, with respect to $\mathsf{P}_{\theta^\star}^n$, proceeds just like in the proof of Theorem 1. This time, we get

$$\int_0^1 \int_{A_{M\varepsilon_n}} \|\theta - \theta^\star\|^2 J_\omega^n(d\theta)\,\tilde{\pi}(d\omega),$$

where $J_\omega^n(d\theta)$ is $\exp(-c\|\theta - \theta^\star\|^2)$ times a probability measure for θ in $\mathbb{R}^n$, just as in the proof of Theorem 1. Since the function $x \mapsto xe^{-cx}$ is monotonically decreasing for large enough x, we can see that, for large n, $\|\theta - \theta^\star\|^2 \exp(-c\|\theta - \theta^\star\|^2) < M\varepsilon_n \exp(-cM\varepsilon_n)$ on $A_{M\varepsilon_n}$. Therefore, the expectation is eventually bounded by $M\varepsilon_n \exp(-cM\varepsilon_n)$. Combining this with the lemma's lower bound, we can find $\nu > 0$ such that, for large n,

$$\mathsf{E}_{\theta^\star} \int_{A_{M\varepsilon_n}} \|\theta - \theta^\star\|^2\, Q_n(d\theta) \leq M\varepsilon_n e^{-\nu\varepsilon_n}.$$

But $\|\hat{\theta}_n - \theta^\star\|^2 \leq \int \|\theta - \theta^\star\|^2\, Q_n(d\theta)$ by Jensen's inequality, so $\mathsf{E}_{\theta^\star}\|\hat{\theta} - \theta^\star\|^2 \leq M\varepsilon_n(1 + e^{-\nu\varepsilon_n})$. Take $M' = 2M$ to complete the proof. $\qquad\square$

3.5. *Effective posterior dimension*

Besides posterior concentration around $\theta^\star$ at the minimax rate, it is desirable if the majority of the posterior mass is concentrated in a roughly s_n-dimensional subspace of $\mathbb{R}^n$, where it is presumed that $\theta^\star$ resides. Castillo and van der Vaart (2012) show that their fully Bayes posteriors have effective dimension proportional to s_n. An interesting question, therefore, is if a similar result obtains for our empirical Bayes posterior. In this section we show that, under the conditions of Theorems 1–2, the posterior distribution for $1-\omega$ puts vanishingly small mass above $s_n n^{-1}$ (up to a logarithmic factor), so that ω tends to concentrate around $1 - s_n n^{-1}$. That this provides some information about the effective dimension of the posterior can be seen from the following expression:

$$\mathsf{E}(\omega \mid X) = \frac{\alpha}{\alpha + 1 + n^{-1}} + \frac{1}{\alpha + 1 + n^{-1}} \frac{\mathsf{E}(D_\theta \mid X)}{n}, \tag{7}$$

where $D_\theta = \#\{i : \theta_i = 0\}$; this fact derives from the full conditionals in Section 4.1 below. So, if α is not too large, and ω concentrates around $1 - s_n n^{-1}$, then D_θ concentrates around $n - s_n$. Therefore, the posterior distribution for θ must reside on a space with effective dimension proportional to s_n

Theorem 3. *Let $\delta_n = K\varepsilon_n n^{-1}$, where $\varepsilon_n = s_n \log(n/s_n)$ as before, and $K > 0$ is a suitably large constant. Then, under the conditions of Theorem 1,*

$$\mathsf{E}_{\theta^\star}\{\mathsf{P}(1 - \omega > \delta_n \mid X)\} \to 0 \quad as\ n \to \infty.$$

Proof. Write the numerator of $\mathsf{P}(1 - \omega > \delta_n \mid X)$ as

$$N_n = \int_0^{1-\delta_n} \int_{\mathbb{R}^n} \prod_{i=1}^n \Big(\frac{p_{\theta_i}(X_i)}{p_{\theta_i^\star}(X_i)}\Big)^\kappa \tilde{\Pi}_{X_i,\omega}(d\theta_i)\,\tilde{\pi}(d\omega).$$

This is similar to the first display in the proof of Theorem 1. Just as in that proof, we get the following bound on the expectation:

$$\mathsf{E}_{\theta^\star}(N_n) \leq \int_0^{1-\delta_n} \prod_{i=1}^n \int_{\mathbb{R}} e^{-\frac{c}{2}(\theta_i - \theta_i^\star)^2} \{\omega\delta_0(d\theta_i) + (1-\omega)\mathsf{N}(d\theta_i \mid \theta_i^\star, v)\}\, \tilde{\pi}(d\omega),$$

where $c > 0$ is a constant and $v = v(\sigma^2, \beta)$ is a variance term that depends on σ^2 and β. Each integral in the inside product is bounded above by 1, so we get

$$\mathsf{E}_{\theta^\star}(N_n) \leq \int_0^{1-\delta_n} \tilde{\pi}(d\omega) = \alpha n \int_0^{1-\delta_n} \omega^{\alpha n - 1}\, d\omega \leq e^{-\alpha n \delta_n} = e^{-K\alpha\varepsilon_n}.$$

From Lemma 1, we have that the denominator of $\mathsf{P}(1 - \omega > \delta_n \mid X)$ is lower bounded by $\exp\{-2\varepsilon_n + O(s_n)\}$ with probability 1 for large n. So, for large n, we get

$$\mathsf{E}_{\theta^\star}\{\mathsf{P}(1 - \omega > \delta_n \mid X)\} \leq \mathsf{E}_{\theta^\star}(N_n)e^{2\varepsilon_n + O(s_n)} \leq e^{-(K\alpha-2)\varepsilon_n + O(s_n)}.$$

If we pick K such that $K\alpha > 2$, then the fact that $s_n = o(\varepsilon_n)$ implies that this upper bound approaches zero as $n \to \infty$, proving the claim. □

Since the logarithmic term $\log(n/s_n)$ is small, the practical implication of this result is that the posterior distribution of ω concentrates around $1 - s_n n^{-1}$. The simulation results displayed in Figure 3 below confirm this.

4. Numerical results

4.1. Computational considerations

Computation of the empirical Bayes posterior mean can be carried out via a simple Gibbs sampler for ω and $\theta = (\theta_1, \ldots, \theta_n)$ based on the full conditionals:

$$\theta_i \mid \omega, X \overset{\text{ind}}{\sim} \begin{cases} \delta_0 & \text{with prob. } \propto \omega e^{-\frac{\kappa}{2}X_i^2} \\ \mathsf{N}(X_i, \frac{\sigma^2}{1+\kappa\sigma^2}) & \text{with prob. } \propto \frac{1-\omega}{\sqrt{1+\kappa\sigma^2}}, \end{cases} \quad i = 1, \ldots, n \qquad (8a)$$

$$\omega \mid \theta, X \sim \mathsf{Beta}(\alpha n + D_\theta, 1 + n - D_\theta), \qquad (8b)$$

where $D_\theta = \#\{i : \theta_i = 0\}$. That is, first sample from the $\theta \mid \omega$ conditional posterior in (8a), then from the $\omega \mid \theta$ conditional posterior in (8b). Repeat this process to obtain a sample from the full posterior. R code for this Gibbs sampling procedure is available at http://www.math.uic.edu/~rgmartin. Once the posterior sample is available, the empirical Bayes estimator $\hat{\theta}$, the posterior mean, is obtained by computing a coordinate-wise average of the posterior θ samples. Besides the posterior mean, many other quantities of interest can be calculated. For example, inclusion probabilities, $\mathsf{P}(\theta_i \neq 0 \mid X)$, $i = 1, \ldots, n$, can be easily calculated. Also, in a function estimation problem, where $\theta_1, \ldots, \theta_n$ are coefficients attached to the fixed basis functions, the posterior samples of the unknown functions are readily available.

Mean square errors, based on 100 replications, sampling X of dimension $n = 200$. First three rows are from Bhattacharya et al. (2014); last two rows corresponds to the proposed empirical Bayes methods. Boldface font indicates the column winner

s_n	10		20		40	
A	7	8	7	8	7	8
$\mathrm{DL}_{1/n}$	16	14	33	31	66	60
EBMed	26	26	57	56	119	119
PMed1	23	22	49	48	102	102
$\mathrm{EBM}_{\alpha=0.25}$	**13**	**13**	**25**	**25**	**47**	**48**
$\mathrm{EBM}_{\hat{\alpha}}$	14	14	26	27	52	50

Theory and experience suggest that good numerical results are obtained for large κ and large σ^2. Throughout, we use $\kappa = 0.99$ and $\sigma^2 = (1 - 0.99)^{-1} = 100$, on the boundary of the feasible region. Other values of (κ, σ^2) like these give similar numerical results. Equation (7) suggests that relatively small values of α are appropriate, so that the ω posterior can learn from X through D_θ. Here we consider two different strategies for handling α.

- Based on experience, we have found that α decreasing in n works well. (This has no consequence on the results in Theorems 1–3.) In particular, in the three examples below, with $n = 200, 500, 1000$ we take $\alpha = 0.25, 0.10, 0.05$, respectively.
- Alternatively, one could use the data to choose α. For example, a method-of-moments estimator of α can be obtained as follows. First, estimate $D = D_\theta$ via universal hard thresholding, i.e., $\hat{D}$ equals the number of X_i such that $|X_i| \leq (2 \log n)^{1/2}$. Under the assumed prior, D has a beta–binomial distribution, with expectation $n^2 \alpha / (n\alpha + 1)$. If we set this expectation equal to $\hat{D}$, then solving for α gives a method-of-moments estimator, in particular, $\hat{\alpha} = \hat{D}\{n(n - \hat{D})\}^{-1}$.

4.2. Simulation studies

For illustration, we first reproduce a simulation study presented in Bhattacharya *et al.* (2014). In particular, we take samples $X = (X_1, \ldots, X_n)$ of dimension $n = 200$ from the normal mean model $X_i \sim \mathsf{N}(\theta_i^\star, 1)$. Recall the sparsity level s_n is the number of non-zero $\theta_i^\star$'s. In this case, we consider $s_n = 10, 20, 40$, and the signals are fixed at values $A = 7, 8$. Table 1 displays estimates of the mean squared error obtained from 100 replications of X. In addition to the proposed empirical Bayes posterior mean estimator (EBM), based on $\kappa = 0.99$, $\sigma^2 = 100$, and either $\alpha = 0.25$ or $\hat{\alpha}$, the methods being compared are a Dirichlet–Laplace estimator (DL) of Bhattacharya *et al.* (2014), an empirical Bayes median estimator (EBMed) of Johnstone and Silverman (2004), and a fully Bayes posterior median estimator (PMed1) of Castillo and van der Vaart (2012). A few other methods have been considered in the literature recently, and some comments on why they are omitted from comparison here are given in Remark 3 below.
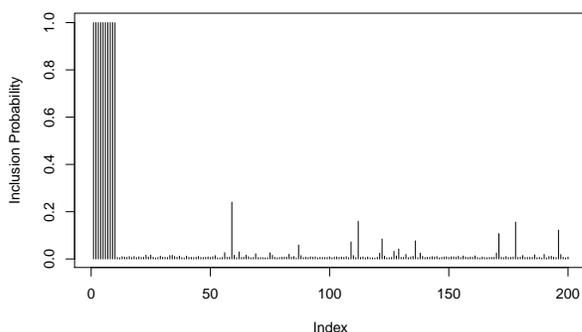
FIG 2. *Plot of the empirical Bayes posterior inclusion probability* $\mathsf{P}(\theta_i \neq 0 \mid X)$ *for* $i = 1, \ldots, n$. *Here* $n = 200$, $s_n = 10$, *and* $\theta_1^\star = \cdots = \theta_{10}^\star = 7$.

Here, we find that our proposed empirical Bayes estimator, with fixed $\alpha = 0.25$, is the top performer overall, with the data-driven α case close behind.

Consider a single sample X under the simulation setting described above, with $n = 200$, where the first $s_n = 10$ entries in $\theta^\star$ equal $A = 7$, and the remaining entries are zero. For the given X, the Gibbs sampler is run to obtain a sample from our empirical Bayes posterior distribution of θ. In Figure 2 we plot the posterior inclusion probability $\mathsf{P}(\theta_i \neq 0 \mid X)$ as a function of the indices $i = 1, \ldots, n$. It is evident that the empirical Bayes posterior is able to clearly identify the correct model.

As a second example, we reproduce a simulation study presented in Castillo and van der Vaart (2012). In this case, we look at $n = 500$, $s_n = 25, 50, 100$, and signals fixed at $A = 3, 4, 5$. Table 2 displays estimates of the mean squared error based on 100 replications. This time, the methods are two fully Bayes

TABLE 2

Mean square errors, based on 100 replications, sampling X of dimension $n = 500$. First eight rows are from Castillo and van der Vaart (2012); last two rows corresponds to the proposed empirical Bayes methods. Boldface font indicates the column winner

s_n	25			50			100		
A	3	4	5	3	4	5	3	4	5
PM1	111	96	94	176	165	154	267	302	307
PM2	106	92	82	169	165	152	269	280	274
EBM	**103**	96	93	166	177	174	271	312	319
PMed1	129	83	73	205	149	130	255	279	283
PMed2	125	86	68	187	**148**	129	273	254	245
EBMed	110	**81**	72	**162**	**148**	142	**255**	294	300
HT	175	142	70	339	284	135	676	564	252
HTO	136	92	84	206	159	139	306	261	245
EBM$_{\alpha=0.10}$	139	99	54	237	159	**89**	386	245	**152**
EBM$_{\hat{\alpha}}$	141	97	**48**	232	153	90	379	**235**	155

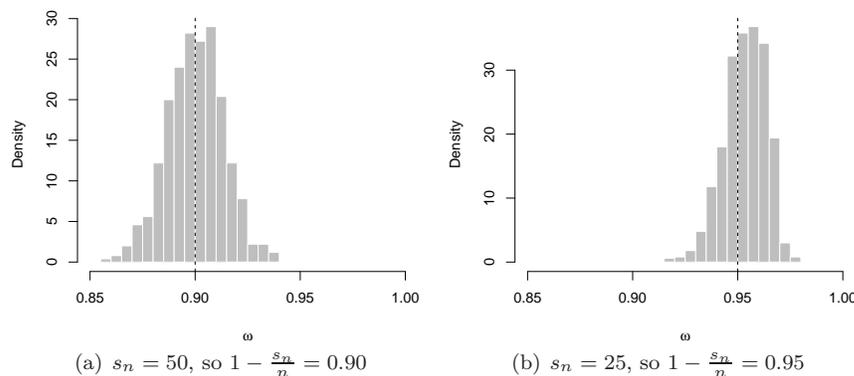(a) $s_n = 50$, so $1 - \frac{s_n}{n} = 0.90$ (b) $s_n = 25$, so $1 - \frac{s_n}{n} = 0.95$

FIG 3. *Posterior distributions for ω when $n = 500$ and $A = 5$ for two values of s_n. In each case, $\kappa = 0.99$, $\sigma^2 = 100$, and $\alpha = 0.10$.*

posterior mean estimates (PM1 and PM2), two fully Bayes component-wise posterior medians (PMed1 and PMed2), Johnstone and Silverman (2004) empirical Bayes mean (EBM) and median (EBMed), and hard thresholding (HT) and hard thresholding oracle (HTO) rules. Our proposed empirical Bayes estimator, based on both $\alpha = 0.10$ and $\hat{\alpha}$, is competitive when $A = 4$, and clearly dominates when $A = 5$, just like in the previous illustration. Interestingly, the empirical Bayes estimators are the better performers overall in this case.

One rather unusual observation is that some of the methods have, for given s_n, a mean square error increasing in the signal size A. We find this behavior to be counterintuitive, since it should be easier to detect stronger signals. The two thresholding estimators have decreasing mean square error as A increases, as does our proposed estimator.

To follow up on the mean square error results in Table 2, we also display the posterior distribution of ω for two separate runs. As indicated from Theorem 3, the posterior distribution for ω should concentrate around $1 - s_n n^{-1}$. For both cases in Figure 3, the posterior concentrates where we expect it would.

As a final example, consider a $n = 1000$ dimensional mean vector, with the first 10 entries of $\theta^\star$ equal 10, the next 90 entries equal A, and the remaining 900 entries equal zero. Mean square errors for two Dirichlet–Laplace estimators in Bhattacharya *et al.* (2014) and our empirical Bayes estimator, based on $\alpha = 0.05$, are displayed in Table 3. Here we consider a range of A, from $A = 2$ to $A = 7$. For the smaller signals, $A \leq 4$, the Dirichlet–Laplace estimator, with prior weight $1/2$ is the best, but our estimator, with either $\alpha = 0.05$ or $\hat{\alpha}$, is better for larger signals, $A > 4$. The larger weight Dirichlet–Laplace prior estimator is dominated by our empirical Bayes estimator.

Remark 3. There are a number of existing methods available for this problem besides those included in our comparisons here. These include the lasso (Tibshirani, 1996), the Bayesian lasso (Park and Casella, 2008), the horseshoe prior

TABLE 3

Mean square errors, based on 100 replications, sampling X of dimension $n = 1000$. First two rows are from Bhattacharya et al. (2014); last row corresponds to the proposed empirical Bayes posterior mean. Boldface font indicates the column winner

A	2	3	4	5	6	7
$\mathrm{DL}_{1/2}$	368	679	671	374	214	160
$\mathrm{DL}_{1/n}$	**268**	**316**	**267**	213	193	177
$\mathrm{EBM}_{\alpha=0.05}$	320	416	291	**172**	137	**129**
$\mathrm{EBM}_{\hat{\alpha}}$	323	415	282	176	**136**	**129**

estimator (Carvalho, Polson and Scott, 2010), the empirical Bayes estimators of Jiang and Zhang (2009), Brown and Greenshtein (2009), and, most recently, Koenker and Mizera (2014). Some of these methods, including a version of ours, are compared more extensively in Koenker (2014). Those estimators without minimax guarantees, such as Koenker–Mizera, can only be motivated by finite-sample simulation studies which are necessarily narrow. On the other hand, our estimator has the desired minimax property and also has the best overall finite-sample performance among those provably minimax competitors.

5. Discussion

The paper has considered a classical problem of estimating a sparse high-dimensional normal mean vector, and we have proposed a novel empirical Bayes solution. Though the stated prior itself may seem overly informative, we show that the prior induces a sort of shrinkage effect, preventing the posterior from tracking the data too closely. We go on to prove that the empirical Bayes posterior concentrates around $\theta^\star$ at the minimax rate, that its mean is an asymptotic minimax estimator, and that its effective dimension agrees with that of the true sparse mean vector.

The mathematical device used in our asymptotic analysis is an alternative representation of the empirical Bayes model with a fractional likelihood. As in Walker and Hjort (2001), this fractional likelihood posterior is a powerful tool, though our concentration results do not follow immediately from theirs. With this adjustment, the prior changes to a very simple one, which we have called $\tilde{\Pi}_X$. The key to success of our empirical Bayes posterior in the asymptotic framework is the particular beta prior on ω, under $\tilde{\Pi}_X$. From this prior, and the lower bound derived in Lemma 1, the minimax rate $\varepsilon_n = s_n \log(n/s_n)$ drops out almost automatically. As we indicated, to push through the minimax concentration result, we only need the conditional prior on θ, given ω, under $\tilde{\Pi}_X$, to not interfere with the dynamics induced by the prior on ω. Intuitively, there should be many priors that would accomplish this. We showed that an empirical Bayes prior obtained by centering a Gaussian prior at the observations, under $\tilde{\Pi}_X$, minimax concentration follows relatively easily. Castillo and van der Vaart (2012) have similar results, e.g., they make sure the prior for θ does not interfere by requiring suitably heavy tails.

In addition to the good large-sample properties, our empirical Bayes procedure is easy to compute, and, in a number of cases, the finite-sample performance of our empirical Bayes posterior mean is considerably better than that of existing methods with comparable large-sample properties (Remark 3). Since our method admits a full posterior distribution, any other feature, such as the inclusion probabilities displayed in Figure 2, useful in the signal detection problem, can be readily calculated.

A possible extension of the method presented herein is as follows. Suppose that each X_i and θ_i are r-vectors, where $r = r_n$ possibly depends on n. Collecting some of the variables together in vectors introduce a group structure. This structure appears in a variety of applications, and this has motivated developments in model selection and estimation with grouped variables (e.g., Yuan and Lin, 2006). Abramovich and Grinshtein (2013) prove asymptotic minimaxity of a Bayes method in this grouped setting, and we expect that similar results can be derived based on the ideas presented here.

Finally we mention that the techniques introduced in this paper for dealing with sparsity have already been applied to the high-dimensional $(p \gg n)$ regression problem; see Martin, Mess and Walker (2014).

Acknowledgements

References

Abramovich, F. and Grinshtein, V. (2013). Estimation of a sparse group of sparse vectors. *Biometrika* **100** 355–370. MR3068439

Abramovich, F., Benjamini, Y., Donoho, D. L. and Johnstone, I. M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.* **34** 584–653. MR2281879

Babenko, A. and Belitser, E. (2010). Oracle convergence rate of posterior under projection prior and Bayesian model selection. *Math. Methods Statist.* **19** 219–245. MR2742927

Barron, A., Schervish, M. J. and Wasserman, L. (1999). The consistency of posterior distributions in nonparametric problems. *Ann. Statist.* **27** 536–561. MR1714718

Bhattacharya, A., Pati, D., Pillai, N. S. and Dunson, D. B. (2014). Dirichlet–Laplace priors for optimal shrinkage. Unpublished manuscript, arXiv:1212.6088.

Bogdan, M., Ghosh, J. K. and Tokdar, S. T. (2008). A comparison of the Benjamini-Hochberg procedure with some Bayesian rules for multiple testing. In *Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen* (N. Balakrishnan, E. Peña and M. Silvapulle, eds.) 211–230. IMS, Beachwood, OH. MR2462208

BOGDAN, M., CHAKRABARTI, A., FROMMLET, F. and GHOSH, J. K. (2011). Asymptotic Bayes-optimality under sparsity of some multiple testing procedures. *Ann. Statist.* **39** 1551–1579. MR2850212

BROWN, L. D. and GREENSHTEIN, E. (2009). Nonparametric empirical Bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *Ann. Statist.* **37** 1685–1704. MR2533468

CAI, T. T. (2012). Minimax and adaptive inference in nonparametric function estimation. *Statist. Sci.* **27** 31–50. MR2953494

CAI, T. T. and JIN, J. (2010). Optimal rates of convergence for estimating the null density and proportion of nonnull effects in large-scale multiple testing. *Ann. Statist.* **38** 100–145. MR2589318

CAI, T. T., ZHANG, C.-H. and ZHOU, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.* **38** 2118–2144. MR2676885

CAI, T. T. and ZHOU, H. H. (2012). Optimal rates of convergence for sparse covariance matrix estimation. *Ann. Statist.* **40** 2389–2420.

CARVALHO, C. M., POLSON, N. G. and SCOTT, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97** 465–480. MR2650751

CASTILLO, I. and VAN DER VAART, A. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *Ann. Statist.* **40** 2069–2101. MR3059077

DALALYAN, A. S. and TSYBAKOV, A. B. (2008). Aggregation by exponential weighting, sharp PAC-Bayesian bounds, and sparsity. *Machine Learning* **72** 39–61.

DONOHO, D. L. and JOHNSTONE, I. M. (1994). Minimax risk over l_p-balls for l_q-error. *Probab. Theory Related Fields* **99** 277–303. MR1278886

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** 41–81. With discussion and a reply by the authors. MR1157714

EFRON, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statist. Sci.* **23** 1–22. MR2431866

FAN, J. and LV, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statist. Sinica* **20** 101–148. MR2640659

GHOSAL, S., GHOSH, J. K. and RAMAMOORTHI, R. V. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *Ann. Statist.* **27** 143–158. MR1701105

GHOSAL, S., GHOSH, J. K. and VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531. MR1790007

JIANG, W. and TANNER, M. A. (2008). Gibbs posterior for variable selection in high-dimensional classification and data mining. *Ann. Statist.* **36** 2207–2231. MR2458185

JIANG, W. and ZHANG, C.-H. (2009). General maximum likelihood empirical Bayes estimation of normal means. *Ann. Statist.* **37** 1647–1684. MR2533467

JIN, J. and CAI, T. T. (2007). Estimating the null and the proportional of nonnull effects in large-scale multiple comparisons. *J. Amer. Statist. Assoc.* **102** 495–506. MR2325113

JOHNSTONE, I. M. and SILVERMAN, B. W. (2004). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Ann. Statist.* **32** 1594–1649. MR2089135

JOHNSTONE, I. M. and SILVERMAN, B. W. (2005). Empirical Bayes selection of wavelet thresholds. *Ann. Statist.* **33** 1700–1752. MR2166560

KOENKER, R. (2014). A Gaussian compound decision bakeoff. *Stat.* **3** 12–16.

KOENKER, R. and MIZERA, I. (2014). Convex optimization, shape constraints, compound decisions, and empirical Bayes rules. *J. Amer. Statist. Assoc.* **109** 674–685. MR3223742

LAM, C. and FAN, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.* **37** 4254–4278. MR2572459

MARTIN, R., MESS, R. and WALKER, S. G. (2014). Empirical Bayes posterior concentration in sparse high-dimensional linear models. Unpublished manuscript, arXiv:1406.7718.

MARTIN, R. and TOKDAR, S. T. (2012). A nonparametric empirical Bayes framework for large-scale multiple testing. *Biostatistics* **13** 427–439.

PARK, T. and CASELLA, G. (2008). The Bayesian lasso. *J. Amer. Statist. Assoc.* **103** 681–686. MR2524001

SCHWARTZ, L. (1965). On Bayes procedures. *Z. Wahrs. verw. Geb.* **4** 10–26. MR0184378

SCOTT, J. G. and BERGER, J. O. (2006). An exploration of aspects of Bayesian multiple testing. *J. Statist. Plann. Inference* **136** 2144–2162. MR2235051

SCOTT, J. G., KELLY, R. C., SMITH, M. A. and KASS, R. E. (2013). False discovery rate regression: An application to neural synchrony detection in primary visual cortex. Unpublished manuscript, arXiv:1307.3495.

SHEN, X. and WASSERMAN, L. (2001). Rates of convergence of posterior distributions. *Ann. Statist.* **29** 687–714. MR1865337

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

WALKER, S. and HJORT, N. L. (2001). On Bayesian consistency. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **63** 811–821. MR1872068

YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68** 49–67. MR2212574

ZHANG, T. (2006). From ϵ-entropy to KL-entropy: Analysis of minimum information complexity density estimation. *Ann. Statist.* **34** 2180–2210. MR2291497