Empirical Bayes posterior concentration in sparse high-dimensional linear models

RYAN MARTIN[1,*], RAYMOND MESS[1,**] and STEPHEN G. WALKER[2]

[1]*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;* [**]*rmess1@uic.edu*
[2]*Department of Mathematics, University of Texas at Austin, 2525 Speedway Stop C1200, Austin, TX 78712, USA. E-mail:* *s.g.walker@math.utexas.edu*

We propose a new empirical Bayes approach for inference in the $p \gg n$ normal linear model. The novelty is the use of data in the prior in two ways, for centering and regularization. Under suitable sparsity assumptions, we establish a variety of concentration rate results for the empirical Bayes posterior distribution, relevant for both estimation and model selection. Computation is straightforward and fast, and simulation results demonstrate the strong finite-sample performance of the empirical Bayes model selection procedure.

Keywords: data-dependent prior; fractional likelihood; minimax; regression; variable selection

1. Introduction

In this paper, we consider the Gaussian linear regression model, given by

$$Y = X\beta + \varepsilon, \tag{1}$$

where Y is a $n \times 1$ vector of response variables, X is a $n \times p$ matrix of predictor variables, β is a $p \times 1$ vector of slope coefficients, and ε is a $n \times 1$ vector of i.i.d. $\mathsf{N}(0, \sigma^2)$ random errors. Recently, there has been considerable interest in the high-dimensional case, where $p \gg n$, driven primarily by challenging applications. Indeed, in genetic studies, where the response variable corresponds to a particular observable trait, the number of subjects, n, may be of order 10^3, while the number of genetic features, p, in consideration can be of order 10^5. Despite the large number of features, usually only a few have a genuine association with the trait. For example, the Wellcome Trust Case Control Consortium [37] has confirmed that only seven genes have a non-negligible association with Type I diabetes. Therefore, it is reasonable to assume that β is sparse, that is, only a few non-zero entries.

Given the practical importance of the high-dimensional regression problem, there is now a substantial body of literature on the subject. In the frequentist setting, a variety of methods are available based on minimizing loss functions, equipped with a penalty on the complexity of the model. This includes the lasso (Tibshirani [34]), the smoothly clipped absolute deviation (Fan and Li [13]), the adaptive lasso (Zou [41]), and the Dantzig selector (Candes and Tao [7], James and Radchenko [22], James, Radchenko and Lv [23]). Fan and Lv [14] give a selective overview of these and other frequentist methods. From a Bayesian perspective, popular methods for variable selection in high-dimensional regression include stochastic search variable selection (George and McCullogh [16]) and the methods based on spike-and-slab priors (Ishwaran and

Rao [20,21]). These methods and others are reviewed in Clyde and George [11] and Heaton and Scott [19]. More recently, Bondell and Reich [5], Johnson and Rossell [26], and Narisetty and He [28] propose Bayesian variable selection methods and establish model selection consistency.

Any Bayesian approach to the regression problem (1) yields a posterior distribution on the high-dimensional parameter β. It is natural to ask under what conditions will the β posterior distribution concentrate around the true value at an appropriate or optimal rate. Recently, Castillo, Schmidt-Hieber and van der Vaart [8] show that, with a suitable Laplace-like prior for β, similar to those in Park and Casella [29], and under conditions on the design matrix X, the posterior distribution concentrates around the truth at rates that match those for the corresponding lasso estimator (e.g., Bühlmann and van de Geer [6]). These results leave room for improvement in at least two directions; first, the rates associated with the lasso estimator are not optimal, so a break from the Laplace priors (and perhaps even the standard Bayesian setup itself) is desirable; second, and perhaps most importantly, posterior computation with these inconvenient non-conjugate priors is expensive and non-trivial. In this paper, we develop a new approach, motivated by computational considerations, which leads to improvements in both directions, simultaneously.

Towards a model that leads to more efficient computation, it is natural to consider a conjugate normal prior for β. However, Theorem 2.8 in Castillo and van der Vaart [9] says that if the prior has normal tails, then the posterior concentration rates can be suboptimal, motivating a departure from the somewhat rigid Bayesian framework. Following Martin and Walker [27], we consider a new empirical Bayes approach, motivated by the very simple idea that the tails of the prior are irrelevant as long as its center is chosen informatively. So, our proposal is to use the data to provide an informative center for the normal prior for β, along with an extra regularization step to prevent the posterior from tracking the data too closely. Details of our proposed empirical Bayes model are presented in Section 2. It turns out that this new empirical Bayes posterior is both easy to compute and has desirable asymptotic concentration properties. Section 3 presents a variety of concentration rate results for our empirical Bayes posterior. For example, under almost no conditions on the model or design matrix, a concentration rate relative to prediction error loss is obtained which is, at least in some cases, minimax optimal; the optimal rate can be achieved in all cases, but at a cost (see Remark 1). Furthermore, we provide a model selection consistency result which says that, under optimal conditions, the empirical Bayes posterior can asymptotically identify those truly non-zero coefficients in the linear model. Our approach has some similarities with the exponential weighting methods in, for example, Rigollet and Tsybakov [31,32] and Arias-Castro and Lounici [2]; in fact, ours can be viewed as a generalization of these approaches, defining a full posterior that, when suitably summarized, corresponds essentially to their estimators. In Section 4, we propose a simple and efficient Markov chain Monte Carlo method to sample from our empirical Bayes posterior, and we present several simulation studies to highlight both the computational speed the superior finite-sample performance of our method compared to several others in terms of model selection. Finally, Section 5 gives a brief a discussion, the key message being that we get provable posterior concentration results, optimal in a minimax sense in some cases, fast and easy computation, and strong finite-sample performance. Lengthy proofs and some auxiliary results are given in the Appendix.

2. The empirical Bayes model

2.1. The prior

Here, and in the theoretical analysis in Section 3, we take the error variance σ^2 to be known, as is often done (e.g., Castillo, Schmidt-Hieber and van der Vaart [8], Rigollet and Tsybakov [32]). Techniques for estimating σ^2 in the high-dimensional case are available; see Section 4. To specify a prior for β that incorporates sparsity, we decompose β as (S, β_S), where $S \subset \{1, \ldots, p\}$ denotes the "active set" of variables, $S = \{j : \beta_j \neq 0\}$, and β_S is the $|S|$-vector containing the particular non-zero values. Based on this decomposition, we can specify the prior for β in two steps: a prior for S and then a prior for β_S, given S.

First, the prior $\pi(S)$ for the model S decomposes as follows:

$$\pi(S) = \binom{p}{s}^{-1} f_n(s), \qquad s = 0, 1, \ldots, p, s = |S|, \tag{2}$$

where $f_n(s)$ is a probability mass function on the size $|S|$ of S. That is, we assign a prior distribution $f_n(s)$ on the model size and then, given the size, put a uniform prior on all models of the given size. Some conditions on $f_n(s)$ will be required for suitable posterior concentration. In particular, we assume that $f_n(s)$ is supported on $\{0, 1, \ldots, R\}$, not on $\{0, 1, \ldots, p\}$, where $R \leq n$ is the rank of the matrix X; see, also, Jiang [24], Abramovich and Grinshtein [1], Rigollet and Tsybakov [32], and Arias-Castro and Lounici [2]. That is,

$$f_n(s) = 0 \qquad \text{for all } s = R + 1, \ldots, p. \tag{3}$$

Our primary motivation for imposing this constraint is that in practical applications, the true value of s; that is, $s^\star = |S^\star|$, is typically much smaller than R. Even in the ideal case where $S^\star$ is known, if $|S^\star| > R$, then quality estimation of the corresponding parameters is not possible. Moreover, models containing a large number of variables can be difficult to interpret. Therefore, since having no more variables than samples in the fixed-model case is a reasonable assumption, we do not believe that restricting the support of our prior for the model size is a strong condition.

Second, for the conditional prior on β_S, given S that satisfies $|S| \leq R$, we propose to employ the available distribution theory for the least squares estimator $\widehat{\beta}_S$. Specifically, we take the prior for β_S, given S, as

$$\beta_S | S \sim \mathsf{N}_{|S|}\big(\widehat{\beta}_S, \gamma^{-1}\big(X_S^\top X_S\big)^{-1}\big). \tag{4}$$

Here, X_S is the matrix filled with columns of X corresponding to S, and $\gamma > 0$ is a tuning parameter, to be specified. This is reminiscent to Zellner's g-prior (e.g., Zellner [38]), except that it is centered at the least squares estimator; see Section 2.2 for more on this data-dependent prior centering. To summarize, our proposed prior Π for β is given by

$$\Pi(d\beta) = \sum_{S:|S| \leq R} \mathsf{N}_{|S|}\big(d\beta_S | \widehat{\beta}_S, \gamma^{-1}\big(X_S^\top X_S\big)^{-1}\big)\delta_0(d\beta_{S^c})\pi(S). \tag{5}$$

Following Martin and Walker [27], we refer to this data-dependent prior as an empirical prior; see Section 2.3. By restricting $|S| \leq R$, we can be sure that the least squares estimator $\widehat{\beta}_S$ is available, along with the usual distribution theory. In our implementation, γ^{-1} will be large, which means that the conditional prior for β_S is rather diffuse, so the dependence on the data, through $\widehat{\beta}_S$, is not overly strong.

Obviously, to properly define the conditional prior for β_S, we implicitly assume that $X_S^\top X_S$ is non-singular for all subsets S with $|S| \leq R$. This is only for simplicity, however, since the theory in Section 3 goes through without this assumption at the cost of making computations more difficult.

2.2. The likelihood function

For the likelihood function, write $L_n(\beta) = \mathsf{N}_n(Y|X\beta, \sigma^2 I)$ as the n-dimensional Gaussian density at Y, with mean $X\beta$, covariance matrix proportional to the identity matrix, and treated as a function of β. One unique feature of our approach so far is the centering of the (conditional) prior on the least squares estimator, which is greedy, in some sense. To prevent the posterior from tracking the data too closely, the second feature of our proposed approach is that we introduce a fractional power $\alpha \in (0, 1)$ on the likelihood. That is, instead of $L_n(\beta)$, our likelihood will be $L_n(\beta)^\alpha$; see Martin and Walker [27]. Other authors have advocated the use of a fractional likelihood, including Barron and Cover [4], Walker and Hjort [35], Zhang [40], Jiang and Tanner [25], Dalalyan and Tsybakov [12], and Grünwald and van Ommen [18], but these papers have different foci and none include a data-dependent (conditional) prior centering. In fact, we feel that this combination of centering and fractional likelihood regularization (see Section 2.3) is a powerful tool that can be used for a variety of high-dimensional problems.

Our analysis in what follows does not go through for the genuine Bayes case, corresponding to $\alpha = 1$, but α can be arbitrarily close to 1. Clearly, for finite-samples, the numerical differences between results for $\alpha \approx 1$ and for $\alpha = 1$ are negligible.

2.3. The posterior distribution

Given the prior Π for β and the fractional likelihood, we form an empirical Bayes posterior distribution, denoted by Π^n, for β using the standard Bayesian update. That is, for B a measurable subset of $\mathbb{R}^p$, we have

$$\Pi^n(B) = \frac{\int_B L_n(\beta)^\alpha \Pi(d\beta)}{\int_{\mathbb{R}^p} L_n(\beta)^\alpha \Pi(d\beta)}. \tag{6}$$

Computation of this empirical Bayes posterior will be discussed in Section 4.

We interpret "empirical Bayes" loosely – if the prior depends on data, then the corresponding posterior is empirical Bayes. The combination of a prior, data-dependent or not, with a fractional likelihood via Bayes formula can also be understood from this empirical Bayes point of view. Indeed,

$$L_n(\beta)^\alpha \Pi(d\beta) = L_n(\beta) \frac{\Pi(d\beta)}{L_n(\beta)^{1-\alpha}},$$

that is, the Bayes combination of a fractional likelihood with a prior is equivalent to a Bayes combination of the original likelihood function with a data-dependent prior. As Walker and Hjort [35] explain, rescaling the prior by a portion of the likelihood helps to protect from possible inconsistencies by penalizing those parameter values that "track the data too closely." Our proposal is obviously very different from the traditional empirical Bayes approach. As stated in Section 1, our goal is simply to construct a data-dependent distribution for β that is easy to compute and also has optimal concentration properties. As a guide, we have followed the familiar prior-to-posterior updating, but added a new twist, and we will demonstrate in Sections 3–4 that our proposed empirical Bayes posterior distribution (6) does, indeed, achieve the stated objectives.

3. Posterior concentration rates

3.1. Setup

Before getting into details about the concentration rates, we first want to clarify what is meant by asymptotics in this context. There is an implicit triangular array setup, that is, for each n, the response vector $Y^n = (Y_1^n, \ldots, Y_n^n)^\top$ is modeled according to (1) with the $n \times p$ design matrix $X^n = ((X_{ij}^n))$, of rank $R \leq n$, which we take to be deterministic but depending on n, and vector of coefficients $\beta^n = (\beta_1, \ldots, \beta_p)^\top$. When n is increased, more data is available so, even though there are more variables to contend with (since $p \gg n$), there is hope that something about the true β^n can be learnt, provided that it is sufficiently sparse. In what follows, we will use the standard notation in (1) which is less cumbersome but hides the triangular array formulation. It is important to keep in mind, however, that, throughout our analysis, p, R, and $s^\star$ depend implicitly on n.

We make some minimal standing assumptions. First, without loss of generality, we can assume that $s^\star \leq R \leq n \ll p$. No other assumptions concerning n, p, R, and $s^\star$ will be required. The results below also hold for all fixed tuning parameters $\alpha \in (0, 1)$ and $\gamma > 0$; see Section 4.1 for guidance on the practical choice of (α, γ). For the design matrix X, there is a standing simplifying assumption that we shall make. In particular, we assume that X_S is full-rank for each S satisfying $|S| \leq R$. This assumption holds, for example, if X satisfies the "sparse Riesz condition with rank n" discussed in Zhang and Huang [39] and Chen and Chen [10]. It is possible, however, to remove this condition, but it requires a modification of the empirical Bayes model. Indeed, if the prior π for S only puts positive mass on those S such that X_S is full-rank, and if $X_{S^\star}$ is full-rank, then the theoretical results presented below follow similarly. The drawback for adjusting the prior for S in this way is additional computational cost, that is, the less-than-full-rank models must be identified and removed by zeroing out the prior mass. We opt here to keep things simple by making the full-rank assumption.

3.2. A preliminary result

Let B be a generic event for $\beta \in \mathbb{R}^p$. Our empirical Bayes posterior probability of the event B in (6) can be rewritten as

$$\Pi^n(B) = \frac{\int_B R_n(\beta, \beta^\star)^\alpha \Pi(d\beta)}{\int R_n(\beta, \beta^\star)^\alpha \Pi(d\beta)}, \tag{7}$$

where $R_n(\beta, \beta^\star) = L_n(\beta)/L_n(\beta^\star)$ is the likelihood ratio. Let D_n denote the denominator in the above display, i.e., $D_n = \int R_n(\beta, \beta^\star)^\alpha \Pi(d\beta)$. The next result, which will be useful throughout our analysis, gives a sure lower bound on D_n.

Lemma 1. *There exists $c = c(\alpha, \gamma, \sigma^2) > 0$ such that $D_n \geq \pi(S^\star)e^{-c|S^\star|}$.*

Proof. D_n is an average of a non-negative S-dependent quantity with respect to $\pi(S)$. This average is clearly greater than the quantity for $S = S^\star$ times $\pi(S^\star)$. That is,

$$D_n > \pi(S^\star) \int R_n(\beta, \beta^\star)^\alpha \mathsf{N}(\beta_{S^\star} | \hat{\beta}_{S^\star}, \gamma^{-1}(X_{S^\star}^\top X_{S^\star})^{-1}) d\beta_{S^\star}$$

$$= \pi(S^\star) \int e^{-(\alpha/(2\sigma^2))\{\|Y - X_{S^\star}\beta_{S^\star}\|_2^2 - \|Y - X_{S^\star}\beta_{S^\star}^\star\|_2^2\}} \mathsf{N}(\beta_{S^\star} | \hat{\beta}_{S^\star}, \gamma^{-1}(X_{S^\star}^\top X_{S^\star})^{-1}) d\beta_{S^\star}.$$

Direct calculation shows that the lower bound above equals

$$\pi(S^\star)e^{(\alpha/(2\sigma^2))\|X_{S^\star}(\hat{\beta}_{S^\star} - \beta_{S^\star}^\star)\|_2^2}\left(1 + \frac{\alpha}{\gamma\sigma^2}\right)^{-|S^\star|/2}.$$

Using the trivial bound $\|\cdot\|_2 \geq 0$ on the norm in the exponent, the proof is complete if we let $c = \frac{1}{2}\log(1 + \frac{\alpha}{\gamma\sigma^2})$, which is clearly positive. $\square$

3.3. Prediction loss

We now present a result characterizing the concentration rate of the posterior distribution for the mean $X\beta$. Set

$$B_{\varepsilon_n} = \{\beta \in \mathbb{R}^p : \|X(\beta - \beta^\star)\|_2^2 > \varepsilon_n\}, \tag{8}$$

where ε_n is a positive sequence to be specified. Since this loss involves the X matrix, the notion of convergence we are considering here is related to prediction. Different loss functions will be considered in Section 3.5. As discussed in Bühlmann and van de Geer [6], for example, their equation (2.8), ε_n proportional to $s^\star \log p$ corresponds to the convergence rate for the lasso estimator. Intuitively, if $S^\star$ were known, then the best rate for the prediction error would be $s^\star$, so the logarithmic term acts as a penalty for having to also deal with the unknown model.

Let N_n be the numerator for the posterior probability of B_{ε_n}, as in (7), that is, $N_n = \int_{B_{\varepsilon_n}} R_n(\beta, \beta^\star)^\alpha \Pi(d\beta)$. We have the following bound on N_n.

Lemma 2. *There exists $d = d(\alpha, \sigma^2) > 0$ and $\varphi = \varphi(\alpha, \gamma, \sigma^2) > 1$ such that $\mathsf{E}_{\beta^\star}(N_n) \leq e^{-d\varepsilon_n} \sum_{S:|S|\leq R} \varphi^{|S|} \pi(S)$, uniformly in $\beta^\star$.*

Proof. See Appendix A.1.   □

To bound the posterior probability of B_{ε_n}, let $b_n = \pi(S^\star)e^{-cs^\star}$. Since $D_n \geq b_n$, surely, by Lemma 1, we have

$$\Pi^n(B_{\varepsilon_n}) = \frac{N_n}{D_n} \cdot 1(D_n \geq b_n) + \frac{N_n}{D_n} \cdot 1(D_n < b_n) \leq \frac{N_n}{b_n}.$$

Taking expectation and plugging in the bound in Lemma 2 gives

$$\mathsf{E}_{\beta^\star}\{\Pi^n(B_{\varepsilon_n})\} \leq e^{c|S^\star|-d\varepsilon_n} \frac{1}{\pi(S^\star)} \sum_S \varphi^{|S|}\pi(S)$$

$$= e^{cs^\star-d\varepsilon_n} \frac{\binom{p}{s^\star}}{f_n(s^\star)} \sum_{s=0}^R \varphi^s f_n(s);$$

which holds uniformly in $\beta^\star$ with $|S_{\beta^\star}| = s^\star$. Then the empirical Bayes concentration rate $\varepsilon_n = \varepsilon_n(p, R, s^\star)$ is such that the above upper bound vanishes. A first conclusion is that ε_n must satisfy $s^\star = o(\varepsilon_n)$. More precisely, if we set

$$\zeta_n = \zeta_n(p, R, s^\star) = \frac{\binom{p}{s^\star}}{f_n(s^\star)} \sum_{s=0}^R \varphi^s f_n(s),$$

then the rate ε_n satisfies

$$\log \zeta_n = O(\varepsilon_n) \qquad \text{as } n \to \infty. \tag{9}$$

This amounts to a condition on the prior f_n for $|S|$. Indeed, (9) requires that f_n should be sufficiently concentrated near $s^\star$, so that $f_n(s^\star)$ is not too small and the expectation of $\varphi^{|S|}$ with respect to f_n is not too big. Compare this to the prior support conditions in Ghosal, Ghosh and van der Vaart [17], Shen and Wasserman [33], and Walker, Lijoi and Prünster [36].

We are now ready to state and prove our first main concentration rate result. To keep the statement of the theorem concise, we give an asymptotic convergence result. However, Theorem 1 and Theorems 2–5 in the upcoming sections, are actually stronger than stated, since the proofs are based on getting explicit fixed-$(n, p, s^\star)$ bounds.

Theorem 1. *For any $s^\star \leq R$, if the prior f_n on $|S|$ admits ζ_n such that (9) holds with ε_n, then there exists a constant $M > 0$ such that $\mathsf{E}_{\beta^\star}\{\Pi^n(B_{M\varepsilon_n})\} \to 0$ as $n \to \infty$, uniformly in $\beta^\star$ with $|S_{\beta^\star}| = s^\star$.*

Proof. By Lemmas 1 and 2, and the growth condition (9), we have that, for large n,

$$\log \mathsf{E}_{\beta^\star}\{\Pi^n(B_{M\varepsilon_n})\} \leq \left(\frac{cs^\star}{\varepsilon_n} - Md + \frac{\log \zeta_n}{\varepsilon_n}\right)\varepsilon_n.$$

The first term inside the parentheses vanishes since $s^\star = o(\varepsilon_n)$. Next, under (9), there exists a $K > 0$ such that $(\log \zeta_n)/\varepsilon_n < K$. So, if we take M such that $Md > K$, then the upper bound above goes to $-\infty$ as $n \to \infty$. This implies the result. $\qquad\square$

Remark 1. What rates ε_n are desirable/attainable? The minimax rate for estimation under this prediction error loss is $\min\{R, s^\star \log(p/s^\star)\}$; see, for example, Rigollet and Tsybakov [32]. Note the phase transition between the ordinary $[s^\star \log(p/s^\star) < R]$ and the ultra high-dimensional $[s^\star \log(p/s^\star) > R]$ regimes. According to Remark 2, an empirical Bayes posterior concentration rate equal to $s^\star \log(p/s^\star)$ obtains for a class of priors on S, which is minimax optimal but only in the ordinary high-dimensional regime; this rate is slightly better than those obtained in Arias-Castro and Lounici [2] and Castillo, Schmidt-Hieber and van der Vaart [8], but see Gao, van der Vaart and Zhou [15], Corollary 5.3, for a result comparable to ours in Theorem 1. By picking a prior outside this class, in particular, one that puts a little mass on an overly-complex model, the minimax rate can be achieved in both the ordinary and ultra high-dimensional regimes. There is a price to be paid, however, for this complete minimax rate: the little piece of extra prior mass on the complex model is large enough to cause problems with the proofs of marginal posterior concentration properties for S. Justification of these claims can be found in Appendix B. Based on these observations, we conjecture that the priors on S that lead to minimax concentration rate under prediction error loss do not lead to desirable model selection properties. This is intuitively reasonable, since good prediction generally does not require a correctly specified model, but more work is needed to confirm this. Since we prefer to have a single prior that does well in all aspects, we will not concern ourselves here with attaining the optimal minimax rate in the ultra high-dimensional regime, though we do know how to obtain it.

Remark 2. The growth condition (9) holds with ε_n proportional to $s^\star \log(p/s^\star)$, the minimax rate in the ordinary high-dimensional case, if there exists constants a_1, a_2, c_1, c_2, C_1, and C_2 such that f_n satisfies

$$C_1 \left(\frac{1}{c_1 p^{a_1}} \right)^s \le f_n(s) \le C_2 \left(\frac{1}{c_2 p^{a_2}} \right)^s \qquad \text{for all } s = 0, 1, \ldots, R. \tag{10}$$

The proof of this claim follows from calculations similar to those in Example 1 below. Assumption 1 in Castillo, Schmidt-Hieber and van der Vaart [8] implies (10), but our restriction, $|S| \le R$, allows us to get rates for priors that may not satisfy (10).

Remark 3. Consider the expectation term $\sum_{s=0}^{R} \varphi^s f_n(s)$. The trivial bound φ^R could be used in the ultra high-dimensional case where $s^\star \log(p/s^\star) \gg R$. More generally, if f_n satisfies (10), then the formulas for partial sums of a geometric series reveal that this expectation term is bounded as $n \to \infty$. In fact, in the examples discussed below, it is easy to confirm that the expectation term is bounded. Therefore, the rate is determined completely by the prior concentration around $S^\star$.

Next, we identify the rate ε_n corresponding to several choices of prior f_n. The complexity prior in Example 1, which is simple and has good properties, will be our choice of prior in what follows; our proofs Sections 3.4–3.6 can be easily modified to cover any f_n that satisfies (10).

Example 1. The complexity prior for the model size $|S|$ in equation (2.3) of Castillo, Schmidt-Hieber and van der Vaart [8] is given by

$$f_n(s) \propto c^{-s} p^{-as}, \qquad s = 0, 1, \ldots, R, \tag{11}$$

where a and c are positive constants. This prior clearly satisfies the condition (10) in Remark 2. We claim that this complexity prior satisfies (9) with $\varepsilon_n = s^\star \log(p/s^\star)$. To see this, note that $\log f_n(s^\star)$ is lower bounded by

$$-s^\star \log(cs^{\star a}) - as^\star \log(p/s^\star) = -\left(a + \frac{\log c + a \log s^\star}{\log(p/s^\star)}\right) s^\star \log(p/s^\star).$$

The ratio inside the parentheses above vanishes since $s^\star \ll p$. Similarly, by Stirling's formula, we have that $\log \binom{p}{s^\star} \leq s^\star \log(p/s^\star)\{1 + o(1)\}$. Putting these two bounds together, and using the result in Remark 3, we can conclude that the complexity prior above yields a posterior concentration rate $s^\star \log(p/s^\star)$.

Example 2. Convergence rates can be obtained for other priors f_n. First, consider a beta–binomial prior for $|S|$, that is,

$$f_n(s) = \int_0^1 \binom{R}{s} w^{R-s}(1-w)^s a_n w^{a_n - 1} \, dw,$$

which corresponds to a $\mathsf{Beta}(a_n, 1)$ prior for W and a conditional $\mathsf{Bin}(R, 1 - w)$ prior for $|S|$, given $W = w$. For $a_n = aR$, for a constant $a > 0$, it can be shown that the corresponding rate ε_n is proportional to $s^\star \log(p/s^\star)$. If, on the other hand, f_n is a $\mathsf{Bin}(R, R^{-1})$ mass function, then similar calculations show that the concentration rate is $\varepsilon_n = s^\star \log p$, which agrees with the lasso rate in Bühlmann and van de Geer [6], but falls short of the rates discussed previously.

3.4. Effective dimension

Under our proposed prior, the empirical Bayes posterior distribution for β is concentrated on an R-dimensional subspace of the full p-dimensional parameter space. In the sparse case, where the true $\beta^\star$ has effective dimension $s^\star \leq R \ll p$, it is interesting to ask if the posterior distribution is actually concentrated on a space of dimension close to $s^\star$. Below we give an affirmative answer to this question under some conditions. Such considerations will also be useful in Sections 3.5 and 3.6.

For a given Δ, let $B_n(\Delta) = \{\beta \in \mathbb{R}^p : |S_\beta| \geq \Delta\}$ be those β vectors with no less than Δ nonzero entries. We say that the effective dimension of Π^n is bounded by $\Delta = \Delta_n$ if the expected posterior probability of $B_n(\Delta)$ vanishes as $n \to \infty$. Next, write

$$N_n(\Delta) = \int_{B_n(\Delta)} R_n(\beta, \beta^\star)^\alpha \Pi(d\beta),$$

for the numerator of the posterior probability of $B_n(\Delta)$.

Lemma 3. $\mathsf{E}_{\beta^\star}\{N_n(\Delta)\} \leq \sum_{s=\Delta}^{R} \varphi^s f_n(s)$ *for all* $\beta^\star$.

Proof. See Appendix A.2. □

We can combine Lemmas 3 and 1 to conclude that

$$\mathsf{E}_{\beta^\star}\left[\Pi^n\{B_n(\Delta)\}\right] \leq e^{cs^\star} \frac{\binom{p}{s^\star}}{f_n(s^\star)} \sum_{s=\Delta}^{R} \varphi^s f_n(s),$$

uniformly in $\beta^\star$ with $|S_{\beta^\star}| = s^\star$. Since $\varphi > 1$, we have $\sum_s \varphi^s f_n(s) > 1$ and, therefore,

$$\mathsf{E}_{\beta^\star}\left[\Pi^n\{B_n(\Delta)\}\right] \leq e^{cs^\star + \log \zeta_n} \sum_{s=\Delta}^{R} \varphi^s f_n(s). \tag{12}$$

So, if the tail of the prior f_n on the model size is sufficiently light, then the posterior probability assigned to models with complexity of order greater than $s^\star$ will be small. Under the conditions of Theorem 1, we know the magnitude of $\log \zeta_n$, but here we need additional control on the tails of f_n.

Theorem 2. *Let $s^\star \leq R$. If f_n is of the form* (11), *then $\mathsf{E}_{\beta^\star}[\Pi^n\{B_n(\Delta_n)\}] \to 0$, holds with $\Delta_n = Cs^\star$, uniformly in $\beta^\star$ with $|S_{\beta^\star}| = s^\star$, i.e., the effective dimension Π^n is bounded by $Cs^\star$.*

Proof. Recall that, for this f_n, $\log \zeta_n$ is of the order $s^\star \log(p/s^\star)$. Moreover, for a generic Δ, the summation $\sum_{s=\Delta}^{R} \varphi^s f_n(s)$ is bounded by a partial sum of a geometric series. In particular, the bound is $O(r^{\Delta+1})$, where $r = \varphi/cp^a$ and a, c are in (11). In that case,

$$r^{\Delta+1} = e^{-(\Delta+1)[a \log p + \log(c/\varphi)]}.$$

So, if Δ is a suitable multiple of $s^\star$, then clearly the $r^{\Delta+1}$ term dominates the $e^{cs^\star + \log \zeta_n}$ term. In particular, if $\Delta = Cs^\star$ with $C > a^{-1}$, then the product on the right-hand side of (12) vanishes, proving the claim. □

To summarize, our prior is such that the posterior distribution is supported on models of size no more than R. However, a good prior is one such that the posterior ought to be able to learn the size of the true model that generated the data, which is possibly much less than R. Theorem 2 shows that, indeed, if the prior f_n on the model size has sufficiently light tails, controlled by the prior exponent $a > 0$, then the posterior will concentrate on models of size proportional to $s^\star$, the true model size. We cannot take a $\tilde{s} < s^\star$ to replace $s^\star$ in $Cs^\star$, since we would need

$$\frac{\tilde{s}}{s^\star} \geq \frac{\log(p/s^\star)}{\log p} \to 1,$$

which confirms this particular point. Furthermore, we see exactly the effect that the prior exponent a has through the bound $C > a^{-1}$ on the proportionality constant. So, small a will have

the effect of spreading out the posterior to include some large (but not too large) models, while large a will keep the posterior concentrated on small models. Choosing small a is beneficial in finite-sample studies; see Section 4.

3.5. Other loss functions

Theorem 1 concerns the empirical Bayes posterior probability of sets of β which are near the true $\beta^\star$ relative to a distance depending on the design matrix X. A natural question is if the empirical Bayes posterior concentrates on neighborhoods of $\beta^\star$ with respect to more other metrics, such as ℓ_1- and ℓ_2-norms. An affirmative answer will require further conditions on X to separate β from $X\beta$.

In the low-dimensional case, with $p < n$, we have

$$\left\| X(\beta - \beta^\star) \right\|_2 \geq \lambda_{\min}(X^\top X)^{1/2} \left\| \beta - \beta^\star \right\|_2,$$

where $\lambda_{\min}(A)$ is the minimum eigenvalue of A, which is positive if A is non-singular. When $p \gg n$, X is not full rank and, therefore, the smallest eigenvalue of $X^\top X$ is zero, making the above inequality trivial and not useful. However, it is still possible to get something like the displayed inequality. Toward this, define the function

$$\kappa(s) = \kappa_X(s) = \inf_{\beta : 0 < |S_\beta| \leq s} \frac{\|X\beta\|_2}{\|\beta\|_2}, \qquad s = 1, \ldots, p. \tag{13}$$

The quantity $\kappa(s)$ is called the "smallest scaled sparse singular value of X of dimension s," similar to the quantity in equation (11) of Arias-Castro and Lounici [2] and that in Definition 2.3 of Castillo, Schmidt-Hieber and van der Vaart [8]. Its main purpose is to facilitate conversion of ℓ_2-norm concentration results for the mean vector $X\beta$ to ℓ_2-norm concentration results for β itself. Indeed, a result shown in Lemma 1, Arias-Castro and Lounici [2] is that a true $\beta^\star \in \mathbb{R}^p$ with $|S_{\beta^\star}| = s^\star$ is identifiable if and only if

$$\kappa(2s^\star) > 0. \tag{14}$$

Consequently, κ is an important quantity and will appear in Theorem 3 below. One can define quantities analogous to κ in order to get concentration results relative to the ℓ_1- or ℓ_∞-norm of β; see Castillo, Schmidt-Hieber and van der Vaart [8], Section 2.

The result presented below will follow almost immediately from Theorem 1 and the definition of κ. Indeed, for any β, we have

$$\left\| X(\beta - \beta^\star) \right\|_2 \geq \kappa(|S_{\beta - \beta^\star}|) \left\| \beta - \beta^\star \right\|_2. \tag{15}$$

For example, if $\|\beta - \beta^\star\|_2$ is lower-bounded, then so is $\|X(\beta - \beta^\star)\|_2$, for suitable κ, so a posterior concentration result for the ℓ_2-norm on β should follow from an analogous result for the ℓ_2 prediction error as in Theorem 1. The only obstacle is that the κ term on the right-hand depends on the particular β. The following result leads to the observation that $\kappa(|S_{\beta - \beta^\star}|)$ can be controlled by a term that depends only on $s^\star$.

Lemma 4. *For any β and $\beta^\star$, $\kappa(|S_{\beta-\beta^\star}|) \geq \kappa(|S_\beta| + |S_{\beta^\star}|)$.*

Proof. This follows since κ is non-increasing and $|S_{\beta-\beta^\star}| \leq |S_\beta| + |S_{\beta^\star}|$. $\qquad\square$

Under our prior formulation, we know that the posterior puts probability 1 on those β for which $|S_\beta| \leq R$. So, if $|S_{\beta^\star}| = s^\star$, then, trivially, $\kappa(|S_{\beta-\beta^\star}|) \geq \kappa(R + s^\star)$. For better control on the κ term in (15), recall that Theorem 2 says that the posterior probability of the event $\{|S_\beta| \geq Cs^\star\}$ vanishes as $n \to \infty$. Therefore, for $C' = C + 1$,

$$\kappa(|S_{\beta-\beta^\star}|) \geq \kappa(C's^\star) \tag{16}$$

holds for all β in a set with posterior probability approaching 1. Compare this to Theorem 1 of Castillo, Schmidt-Hieber and van der Vaart [8], and also to the corresponding model selection results for frequentist point estimators in, for example, Bühlmann and van de Geer [6], Chapter 7.

We are now ready for the concentration rate result with respect to the ℓ_2-norm loss on the parameter β itself. This time, set

$$B'_{\delta_n} = \left\{ \beta \in \mathbb{R}^p : \|\beta - \beta^\star\|_2^2 > \delta_n \right\},$$

where δ_n is a positive sequence to be specified.

Theorem 3. *For $s^\star \leq \min(n, R)$, suppose that the prior f_n satisfies (11) with exponent $a > 0$, so that Theorem 1 holds with ε_n equal to $s^\star \log(p/s^\star)$ and Theorem 2 holds with $\Delta_n = Cs^\star$, for $C > a^{-1}$. Then there exists a constant M such that $\mathsf{E}_{\beta^\star}\{\Pi^n(B'_{M\delta_n})\} \to 0$ as $n \to \infty$, uniformly in $\beta^\star$ with $|S_{\beta^\star}| = s^\star$, where*

$$\delta_n = \frac{s^\star \log(p/s^\star)}{\kappa(C's^\star)^2},$$

provided that $\kappa(C's^\star) > 0$, where $C' = 1 + C > 1 + a^{-1} > 1$.

Proof. It follows immediately from (15) that $\|\beta - \beta^\star\|_2^2 > M\delta_n$ implies

$$\|X(\beta - \beta^\star)\|_2^2 > M\kappa(|S_{\beta-\beta^\star}|)^2 \delta_n.$$

By definition of δ_n and the inequality (16), this last inequality implies

$$\|X(\beta - \beta^\star)\|_2^2 > Ms^\star \log(p/s^\star).$$

If we take M as in Theorem 1, then the event in the above display is exactly $B_{M\varepsilon_n}$. We have shown that $\Pi^n(B'_{M\delta_n}) \leq \Pi^n(B_{M\varepsilon})$. By Theorem 1, the expectation of the upper bound vanishes uniformly in $\beta^\star$ as $n \to \infty$, so the proof is almost complete. The remaining issue to deal with is an extra term in the upper bound for $\Pi^n(B'_{M\delta_n})$ coming from using $\kappa(C's^\star)$ in place of $\kappa(|S_{\beta-\beta^\star}|)$ above. However, this extra term is $o(1)$ by Theorem 2, and, therefore, does not actually impact the proof. $\qquad\square$

Compare this result to the third in Theorem 2 of Castillo, Schmidt-Hieber and van der Vaart [8]. First, our rate is slightly better, $s^\star \log(p/s^\star)$ compared to the lasso rate $s^\star \log p$. Second, our bound does not depend on a "compatibility number" (e.g., Castillo, Schmidt-Hieber and van der Vaart [8], Definition 2.1), which also improves the rate and makes interpretation of our result easier. A referee has indicated that the improved results are as a direct consequence of the $(X_S^\top X_S)^{-1}$ term that appears in the prior for β_S. Also, the condition $\kappa(C's^\star) > 0$, with $C' = 1 + a^{-1}$ and $a < 1$, agrees with the condition, roughly, $\kappa((2 + \varepsilon)s^\star) > 0$ for some $\varepsilon > 0$, in Arias-Castro and Lounici [2]; that is, just a little more than identifiability, as in (14) is needed.

3.6. Model selection

Interest here is on the model S and not directly on the regression coefficients. In this case, it is convenient to work with the marginal posterior distribution for S which, thanks to the simple conjugate structure in the conditional prior, we can write explicitly as

$$\pi^n(S) \propto \pi(S) e^{-(\alpha/(2\sigma^2))\|Y - \hat{Y}_S\|^2} \nu^{-|S|}, \tag{17}$$

where $\nu = (\gamma + \alpha/\sigma^2)^{1/2}$. Then

$$\pi^n(S) \leq \frac{\pi^n(S)}{\pi^n(S^\star)} = \frac{\pi(S)}{\pi(S^\star)} \nu^{|S^\star| - |S|} e^{(\alpha/(2\sigma^2))\{\|Y - \hat{Y}_{S^\star}\|^2 - \|Y - \hat{Y}_S\|^2\}}. \tag{18}$$

From this bound, we can show that the posterior concentrates on models contained in $S^\star$, that is, asymptotically, it will not charge any models with unnecessary variables. Furthermore, this conclusion requires no conditions on the X matrix or true $\beta^\star$. For simplicity, we will focus on the particular complexity prior f_n in (11) shown previously to yield desirable posterior concentration properties.

Theorem 4. *Let the constant $a > 0$ in the complexity prior* (11) *be such that $p^a \gg R$. Then* $\mathsf{E}_{\beta^\star}\{\Pi^n(\beta : S_\beta \supset S_{\beta^\star})\} \to 0$, *uniformly over $\beta^\star$.*

Proof. See Appendix A.3.   □

Theorem 4 says that, asymptotically, our empirical Bayes posterior will not include any unnecessary variables. It remains to say what it takes for the posterior to asymptotically identify all the important variables. The first condition is one on the X matrix, specifically, if $s^\star$ is the true model size, then we require $\kappa(s^\star) > 0$; this is implied by monotonicity of κ and the identifiability condition (14) in Section 3.5. For our second assumption, we consider the magnitudes of the non-zero entries in a $s^\star$-sparse $\beta^\star$. Intuitively, we cannot hope to be able to distinguish between an actual zero and a very small non-zero, but defining what is "very small" requires some care. Here, we define this cutoff by

$$\rho_n = \frac{\sigma}{\kappa(s^\star)} \left\{ \frac{2M(1 + \alpha)}{\alpha} \log p \right\}^{1/2}, \tag{19}$$

where $M > 0$ is a constant to be determined. In particular, coefficients of magnitude greater than ρ_n are large enough to be detected. The so-called *beta-min* condition assumes that all the non-zero coefficients are sufficiently far from zero. The cutoff ρ_n in (19) is better than that appearing in equation (2.18) in Bühlmann and van de Geer [6] for the lasso model selector but comparable to that in Theorem 1 of Arias-Castro and Lounici [2] and in the third part of Theorem 5 in Castillo, Schmidt-Hieber and van der Vaart [8], where the latter requires additional assumptions on X.

Theorem 5. *For any $s^\star \leq R$, let $\beta^\star$ be such that $|S_{\beta^\star}| = s^\star$ and*

$$\min_{j \in S^\star} |\beta_j^\star| \geq \rho_n,$$

with $M > a + 1$, where $a > 0$ is in the complexity prior, with $p^a \gg R$. Assuming the condition of Theorem 2 holds, if $\kappa(s^\star) > 0$, then $\mathsf{E}_{\beta^\star}\{\Pi^n(\beta : S_\beta = S_{\beta^\star})\} \to 1$.

Proof. See Appendix A.4. □

4. Numerical results

4.1. Implementation

To compute our empirical Bayes posterior distribution, we employ a Markov chain Monte Carlo method. To start, recall from (17) that we can write the marginal posterior mass function, $\pi^n(S)$, for the model S can be written down explicitly, that is,

$$\pi^n(S) \propto \pi(S) e^{-(\alpha/(2\sigma^2))\|Y - \widehat{Y}_S\|^2} \left(\gamma + \frac{\alpha}{\sigma^2}\right)^{-|S|/2},$$

where $\widehat{Y}_S = X_S \widehat{\beta}_S$ is the least-squares prediction for model S. Intuitively, there are three contributing factors to the posterior distribution for S, namely, the prior probability of the model, a measure of how well the model fits the data, and an additional penalty on the complexity of the model. So, clearly, the posterior distribution will favor models with smaller number of variables that provide adequate fit to the observed Y. This provides further intuition about theorems presented in Section 3.

Besides this intuition, the formula $\pi^n(S)$ provides a convenient way to run a Rao–Blackwellized Metropolis–Hastings method to sample from the posterior distribution of S. Indeed, if $q(S'|S)$ is a proposal function, then a single iteration of our proposed Metropolis–Hastings sampler goes as follows:

1. Given a current state S, sample $S' \sim q(\cdot|S)$.
2. Move to the new state S' with probability

$$\min\left\{1, \frac{\pi^n(S')}{\pi^n(S)} \frac{q(S|S')}{q(S'|S)}\right\};$$

otherwise, stay at state S.

Repeating this process M times, we obtain a sample of models $S_1, \ldots, S_M$ from the posterior $\pi^n(S)$. Monte Carlo approximations of, say, the inclusion probabilities (Section 4.2) of individual variables can then easily be computed based on this sample. In our case, we use a symmetric proposal distribution $q(S'|S)$, that is, one that samples S' uniformly from those models that differ from S in exactly one position, which simplifies the acceptance probability above since the q-ratio is identically 1. Also, we initialize our Markov chain Monte Carlo search at the model selected by lasso.

To implement this procedure, some additional tuning parameters need to be specified. First, recall that $(\alpha, \gamma) = (1, 0)$ corresponds to the genuine Bayes model with a flat prior for β_S. Our theory does not cover this case, but we can mimic it by picking something close. Here we consider $\alpha = 0.999$ and $\gamma = 0.001$; in our experience, the performance is not sensitive to the choice of (α, γ) in a neighborhood of $(0.999, 0.001)$. Second, for the prior on the model size, we employ the complexity prior (11) with $c = 1$ and $a = 0.05$, i.e., $f_n(s) \propto p^{-0.05s}$. The choice of small a makes the prior sufficiently spread out, allowing the posterior to move across the model space and, in particular, helping the Markov chain for S to mix reasonably well. Third, in practice, the error variance σ^2 is seldom known, so some procedure to handle unknown σ^2 is needed. We proposed to modify our empirical Bayes posterior by plugging in an estimate of σ^2. In particular, we use a residual mean square error based on a lasso fit (Reid, Tibshirani and Friedman [30]).

Finally, if samples from the β posterior are desired, then these can easily be obtained, via conjugacy, after a sample of S is available. In particular, the conditional posterior distribution for β_S, given S, is normal with mean $\hat{\beta}_S$ and variance $(\gamma + \frac{\alpha}{\sigma^2})^{-1}(X_S^\top X_S)^{-1}$. R code to implement our procedure is available at www.math.uic.edu/~rgmartin.

4.2. Simulations

In this section, we reconsider some of the simulation experiments performed by Narisetty and He [28], which are related to experiments presented in Johnson and Rossell [26]. In each setting, the error variance is $\sigma^2 = 1$; the covariate matrix is obtained by sampling from a multivariate normal distribution with zero mean, unit variance, and constant pairwise correlation $\rho = 0.25$; and the true model $S^\star$ has $s^\star = 5$. The particular correlation structure among the covariates is given practical justification in Johnson and Rossell [26]. Under this setup, we consider three different settings:

> *Setting* 1. $n = 100$, $p = 500$, and $\beta_{S^\star} = (0.6, 1.2, 1.8, 2.4, 3.0)^\top$;
> *Setting* 2. $n = 200$, $p = 1000$, and $\beta_{S^\star}$ the same as in Setting 1;
> *Setting* 3. $n = 100$, $p = 500$, and $\beta_{S^\star} = (0.6, 0.6, 0.6, 0.6, 0.6)^\top$.

Our Settings 1–2 correspond to the two (n, p) configurations in Case 2 of Narisetty and He [28] and our Setting 3 is the same as their Case 3.

We carry out model selection by retaining those variables whose inclusion probability $p_j = \Pi^n(\beta_j \neq 0)$, $j = 1, \ldots, p$, exceeds 0.5; this is the so-called median probability model, shown to be optimal, in a certain sense, by Barbieri and Berger [3]. Alternatively, one could select the model with the largest posterior probability, but this is more expensive computationally compared

Table 1. Simulation results for Setting 1. First seven rows taken from Table 2 (top) in Narisetty and He [28]; the *EB* row corresponds to our empirical Bayes procedure

Method	$\bar{p}_0$	$\bar{p}_1$	$\mathsf{P}(\hat{S} = S^\star)$	$\mathsf{P}(\hat{S} \supseteq S^\star)$	FDR
BASAD	0.001	0.948	0.730	0.775	0.011
BASAD.BIC	0.001	0.948	0.190	0.915	0.146
BCR.Joint			0.070	0.305	0.268
SpikeSlab			0.000	0.040	0.626
Lasso.BIC			0.005	0.845	0.466
EN.BIC			0.135	0.835	0.283
SCAD.BIC			0.045	0.980	0.328
EB	0.002	0.959	0.680	0.795	0.051

to the median probability model – only p inclusion probabilities instead of up to 2^p model probabilities. In all cases, the posterior almost immediately concentrates on the true model. Our Markov chain Monte Carlo required only 5000 iterations to reach convergence, which took only a few seconds on an ordinary laptop computer: about 10 seconds for Setting 1 and about 25 seconds for Setting 2.

To summarize the performance, we consider five different measures. First, we consider the mean inclusion probability for those variables in and out of the active set $S^\star$, respectively, that is,

$$\bar{p}_1 = \frac{1}{s^\star} \sum_{j \in S^\star} p_j \quad \text{and} \quad \bar{p}_0 = \frac{1}{p - s^\star} \sum_{j \notin S^\star} p_j.$$

We expect the former to be close to 1 and the later to be close to 0. Next, we consider the probability that the model selected by our empirical Bayes method, denoted by $\hat{S}$ is equal to or contains the true model $S^\star$. Finally, we also compute the false discovery rate of our selection procedure. A summary of these quantities for our empirical Bayes method, denoted by *EB*, across the three settings is given in Tables 1–3.

For comparison, we consider those methods discussed in Narisetty and He [28], including their two Bayesian methods, denoted by BASAD and BASAD.BIC. Two other Bayesian methods considered are the credible region approach of Bondell and Reich [5], denoted by BCR.Joint, and the spike-and-slab method of Ishwaran and Rao [20,21], denoted by SpikeSlab. We also consider three penalized likelihood methods, all tuned with BIC, namely, the lasso (Tibshirani [34]), the elastic net (Zou and Hastie [42]), and the smoothly clipped absolute deviation (Fan and Li [13]), denoted by Lasso.BIC, EN.BIC, and SCAD.BIC, respectively. The results for these methods are taken from Tables 2–3 in Narisetty and He [28], which were obtained based on 200 samples taken from the models described in Settings 1–3 described above.

Our selection method based on our empirical Bayes posterior is the overall the best among those being compared in terms of selecting the true model and false discovery rate. In addition to the strong finite-sample performance of our model selection procedure, our theory is arguably stronger than that available for the other methods in this comparison. Take, for example, the BASAD method of Narisetty and He [28], the next-best-performer in the simulation study. Their

Table 2. Simulation results for Setting 2. First seven rows taken from Table 2 (bottom) in Narisetty and He [28]; the *EB* row corresponds to our empirical Bayes procedure

Method	$\bar{p}_0$	$\bar{p}_1$	$P(\hat{S} = S^\star)$	$P(\hat{S} \supseteq S^\star)$	FDR
BASAD	0.000	0.986	0.930	0.950	0.000
BASAD.BIC	0.000	0.986	0.720	0.990	0.046
BCR.Joint			0.090	0.250	0.176
SpikeSlab			0.000	0.050	0.574
Lasso.BIC			0.020	1.000	0.430
EN.BIC			0.325	1.000	0.177
SCAD.BIC			0.650	1.000	0.091
EB	0.000	0.998	0.945	0.990	0.015

method produces a posterior distribution for β but since their prior has no point mass, this posterior cannot concentrate on a lower-dimensional subspace of $\mathbb{R}^p$. So, it is not clear if their posterior distribution for β can attain the minimax concentration rate without tuning the prior using knowledge about the underlying sparsity level.

5. Discussion

We have presented an empirical Bayes model for the sparse high-dimensional regression problem. Though the proposed approach has some unusual features, such as a data-dependent prior, we characterize the posterior concentration rate, which agrees with the optimal minimax rate in some cases. To our knowledge, this is the only available minimax concentration rate result for a full posterior distribution in the sparse high-dimensional linear model. Moreover, our formulation allows for relatively simple posterior computation, via Markov chain Monte Carlo, and simulation studies show that model selection by thresholding the posterior inclusion probabilities outperforms a variety of existing methods.

Table 3. Simulation results for Setting 3. First seven rows taken from Table 3 in Narisetty and He [28]; the *EB* row corresponds to our empirical Bayes procedure

Method	$\bar{p}_0$	$\bar{p}_1$	$P(\hat{S} = S^\star)$	$P(\hat{S} \supseteq S^\star)$	FDR
BASAD	0.002	0.622	0.185	0.195	0.066
BASAD.BIC	0.002	0.622	0.160	0.375	0.193
BCR.Joint			0.030	0.315	0.447
SpikeSlab			0.000	0.000	0.857
Lasso.BIC			0.000	0.520	0.561
EN.BIC			0.040	0.345	0.478
SCAD.BIC			0.045	0.340	0.464
EB	0.003	0.811	0.305	0.350	0.092

The general strategy proposed here goes as follows. Suppose we have a high-dimensional parameter, and different models S identify a set of parameters θ_S. Suppose further that θ is sparse in the sense that only a few of its entries are non-null. Then an empirical Bayes model is obtained by specifying a prior for (S, θ_S) as $\pi(S)\pi(d\theta_S|S)$, where $\pi(d\theta_S|S)$ would be allowed to depend on data through, say, the maximum likelihood estimator $\hat{\theta}_S$ of θ_S. Intuitively, the idea is to center the conditional prior on a data-dependent point, say $\hat{\theta}_S$, and then use the fractional likelihood to prevent the posterior to track the data too closely. We believe this is a general tool that can be used in high-dimensional problems, and one possible application of this approach, which we plan to explore, is a mixture model where S represents the number of mixture components, and θ_S is the set of parameters associated with a mixture model with S mixture components.

Appendix A: Proofs

A.1. Proof of Lemma 2

Write $B_n = B_{\varepsilon_n}$. Rewrite the numerator N_n of the posterior (7) as

$$N_n = \int_{B_n} \sum_S \pi(S) \left\{ \frac{\mathsf{N}(Y|X\beta_{S+}, \sigma^2 I)}{\mathsf{N}(Y|X\beta^\star, \sigma^2 I)} \right\}^\alpha \mathsf{N}\big(\beta_S|\hat{\beta}_S, \gamma^{-1}(X_S^\top X_S)^{-1}\big) d\beta_S$$

$$= \sum_S \pi(S) \int_{B_n(S)} \left\{ \frac{\mathsf{N}(Y|X\beta_{S+}, \sigma^2 I)}{\mathsf{N}(Y|X\beta^\star, \sigma^2 I)} \right\}^\alpha \mathsf{N}\big(\beta_S|\hat{\beta}_S, \gamma^{-1}(X_S^\top X_S)^{-1}\big) d\beta_S,$$

where the sum is over all S with $|S| \leq n$, β_{S+} is a p-vector made by augmenting β_S with $\beta_j = 0$ for all $j \in S^c$, and $B_n(S)$ is the set of all β_S such that $\beta_{S+} \in B_n$. Focus on a single S. Taking expectation of the inner integral with respect to $Y \sim \mathsf{N}(X\beta^\star, \sigma^2 I)$ gives

$$\int_{B_n(S)} \mathsf{E}\left[\left\{ \frac{\mathsf{N}(Y|X\beta_{S+}, \sigma^2 I)}{\mathsf{N}(Y|X\beta^\star, \sigma^2 I)} \right\}^\alpha \mathsf{N}\big(\beta_S|\hat{\beta}_S, \gamma^{-1}(X_S^\top X_S)^{-1}\big) \right] d\beta_S.$$

Apply Hölder's inequality to the inside expectation, that is, for $h > 1$ and $q = (h-1)/h$,

$$\mathsf{E}\left[\left\{ \frac{\mathsf{N}(Y|X\beta_{S+}, \sigma^2 I)}{\mathsf{N}(Y|X\beta^\star, \sigma^2 I)} \right\}^\alpha \mathsf{N}\big(\beta_S|\hat{\beta}_S, \gamma^{-1}(X_S^\top X_S)^{-1}\big) \right]$$

$$\leq \mathsf{E}^{1/h}\left[\left\{ \frac{\mathsf{N}(Y|X\beta_{S+}, \sigma^2 I)}{\mathsf{N}(Y|X\beta^\star, \sigma^2 I)} \right\}^{h\alpha} \right] \mathsf{E}^{1/q}\left[\mathsf{N}^q\big(\beta_S|\hat{\beta}_S, \gamma^{-1}(X_S^\top X_S)^{-1}\big) \right]. \tag{20}$$

If $h\alpha < 1$, then a Renyi divergence formula is available for the first term, giving

$$\mathsf{E}^{1/h}\left[\left\{ \frac{\mathsf{N}(Y|X\beta_{S+}, \sigma^2 I)}{\mathsf{N}(Y|X\beta^\star, \sigma^2 I)} \right\}^{h\alpha} \right] = e^{-(\alpha(1-h\alpha)/(2\sigma^2))\|X(\beta_{S+} - \beta^\star)\|^2}. \tag{21}$$

For the second term in the product above, recall that $\hat{\beta}_S = (X_S^\top X_S)^{-1} X_S^\top Y$. Then

$$X_S \beta_S - X_S \hat{\beta}_S = X_S (X_S^\top X_S)^{-1} X_S^\top (X_S \beta_S - Y),$$

and, therefore, since $X_S (X_S^\top X_S)^{-1} X_S^\top$ is idempotent of rank $|S|$, we get that

$$Z := \frac{1}{\sigma^2} \| X_S \beta_S - X_S \hat{\beta}_S \|^2 = \frac{1}{\sigma^2} \left\| X_S (X_S^\top X_S)^{-1} X_S^\top (X_S \beta_S - Y) \right\|^2$$

is distributed as a non-central chi-square with $|S|$ degrees of freedom and non-centrality parameter $\lambda = \frac{1}{\sigma^2} \| X_S (\beta_S - (X_S^\top X_S)^{-1} X_S^\top X \beta^\star) \|^2$. Then

$$
\begin{aligned}
&\mathsf{E}^{1/q} \left[\mathsf{N}^q \left(\beta_S | \hat{\beta}_S, \gamma^{-1} (X_S^\top X_S)^{-1} \right) \right] \\
&= \frac{\gamma^{|S|/2} |X_S^\top X_S|^{1/2}}{(2\pi)^{|S|/2}} \mathsf{E}^{1/q} \left(e^{-(q\gamma/2)Z} \right) \\
&= \frac{\gamma^{|S|/2} |X_S^\top X_S|^{1/2}}{(2\pi)^{|S|/2}} (1 + q\gamma)^{-|S|/(2q)} e^{-(\gamma/(2(1+q\gamma)))\lambda} \\
&= \frac{\gamma^{|S|/2} |X_S^\top X_S|^{1/2}}{(2\pi)^{|S|/2}} (1 + q\gamma)^{-|S|/(2q)} e^{-(\gamma/(2\sigma^2(1+q\gamma))) \| X_S (\beta_S - (X_S^\top X_S)^{-1} X_S^\top X \beta^\star) \|^2},
\end{aligned}
\tag{22}
$$

where the second equality is from the standard formula for the moment generating function of a non-central chi-square random variable. Now we must integrate the upper bound (20) over $A_n(S)$ with respect to β_S. It is clear from the definition of $B_n(S)$ that the quantity in (21) is bounded on $B_n(S)$, that is,

$$e^{-(\alpha(1-h\alpha)/(2\sigma^2)) \| X(\beta_{S+} - \beta^\star) \|^2} \leq e^{-(\alpha(1-h\alpha)/(2\sigma^2))\varepsilon_n}, \qquad \beta_S \in B_n(S).$$

It is also clear that the expression (22) resembles a normal density in β_S, modulo some multiplicative factors. The algebra is tedious, but the integral of (22) with respect to β_S is bounded above by

$$\varphi^{|S|} \qquad \text{where } \varphi = \left\{ \frac{(1 + q\gamma\sigma^2)^{1-1/q}}{\sigma^2} \right\}^{1/2}.$$

Putting everything together, we have that

$$\mathsf{E}(N_n) \leq e^{-(\alpha(1-h\alpha)/(2\sigma^2))\varepsilon_n} \sum_S \varphi^{|S|} \pi(S).$$

Taking $d = \alpha(1 - h\alpha)/2\sigma^2$ completes the proof.

A.2. Proof of Lemma 3

The proof is an application of ideas used in the proof of Lemma 2. In particular, $N_n(\Delta)$ equals

$$\sum_{S:\Delta \leq |S| \leq n} \pi(S) \int \left\{ \frac{\mathsf{N}(Y|X\beta_{S+}, \sigma^2 I)}{\mathsf{N}(Y|X\beta^\star, \sigma^2 I)} \right\}^\alpha \mathsf{N}\big(\beta_S|\hat{\beta}_S, \gamma^{-1}\big(X_S^\top X_S\big)^{-1}\big)\, d\beta_S.$$

Take expectation with respect to $Y \sim \mathsf{N}(X\beta^\star, \sigma^2 I)$ as in the proof of Lemma 1 and move expectation to the inside of the integral. Working with each S term separately, apply Hölder's inequality to bound the expectation of the product. This upper bound consists of a product of three terms just like in the previous proof. The first is bounded by 1; the second is $\varphi^{|S|}$; and the third is a probability density function in β_S. Then the integral over β_S is bounded by $\varphi^{|S|}$ and the claim follows.

A.3. Proof of Theorem 4

Fix $\beta^\star$ and write $S^\star = S_{\beta^\star}$ as usual. Write P_S for the $n \times n$ matrix projecting onto the column space of X_S. If $S \supset S^\star$, then

$$\|Y - \hat{Y}_{S^\star}\|^2 - \|Y - \hat{Y}_S\|^2 = Y^\top (P_S - P_{S^\star})Y,$$

and, since $P_S - P_{S^\star}$ is idempotent of rank $|S| - |S^\star|$, this quantity is distributed as a non-central chi-square with $|S| - |S^\star|$ degrees of freedom and non-centrality parameter

$$\big(X\beta^\star\big)^\top (P_S - P_{S^\star})\big(X\beta^\star\big).$$

By definition of P_S, it turns out that the non-centrality parameter in the above display is zero, so it is actually an ordinary/central chi-square. From the chi-square moment generating function, we immediately get

$$\mathsf{E}_{\beta^\star}\big\{\pi^n(S)\big\} \leq \frac{\pi(S)}{\pi(S^\star)} z^{|S^\star| - |S|},$$

where z is a constant that depends only on $(\alpha, \gamma, \sigma^2)$. Then

$$\mathsf{E}_{\beta^\star}\big\{\Pi^n\big(\beta : S_\beta \supset S^\star\big)\big\} = \sum_{S:S \supset S^\star} \mathsf{E}_{\beta^\star}\big\{\pi^n(S)\big\} \leq \sum_{S:S \supset S^\star} \frac{\pi(S)}{\pi(S^\star)} z^{|S^\star| - |S|}.$$

Plug in our complexity prior and simplify the upper bound:

$$\sum_{s > s^\star} \frac{\binom{p-s^\star}{p-s}\binom{p}{s^\star}}{\binom{p}{s}} \left(\frac{z}{cp^a}\right)^{s - s^\star}.$$

From

$$\frac{\binom{p-s^\star}{p-s}\binom{p}{s^\star}}{\binom{p}{s}} = \binom{s}{s^\star} = \binom{s}{s-s^\star} \le s^{s-s^\star},$$

the upper bound becomes

$$\sum_{s=s^\star}^{R} \left(\frac{zs}{cp^a}\right)^{s-s^\star} \le \frac{zR}{cp^a} \times O(1).$$

So, if a is such that $p^a \gg R$, the upper bound vanishes, completing the proof.

A.4. Proof of Theorem 5

In light of Theorem 4, it suffices to show that $\mathsf{E}_{\beta^\star}\{\Pi^n(\beta : S_\beta \subset S^\star)\} \to 0$. To start, take a generic $S \subset S^\star$. Then, from (18), we have

$$\frac{\pi^n(S)}{\pi^n(S^\star)} = \frac{\pi(S)}{\pi(S^\star)} \nu^{s^\star - |S|} e^{-(\alpha/(2\sigma^2))\{\|Y-\hat{Y}_S\|^2 - \|Y-\hat{Y}_{S^\star}\|^2\}}.$$

The exponent $Z_S := \frac{1}{\sigma^2}\{\|Y - \hat{Y}_S\|^2 - \|Y - \hat{Y}_{S^\star}\|^2\}$ is a chi-square random variable with $s^\star - |S|$ degrees of freedom and non-centrality parameter

$$\lambda_S := \frac{1}{\sigma^2}(X\beta^\star)^\top (P_{S^\star} - P_S)(X\beta^\star).$$

The algebra is a bit tedious, but we can simplify λ_S as

$$\lambda_S = \frac{1}{\sigma^2}\left\|(I - P_S)X_{S^\star \cap S^c}\beta^\star_{S^\star \cap S^c}\right\|^2.$$

From the non-central chi-square moment generating function we have

$$\mathsf{E}_{\beta^\star}\{\pi^n(S)\} \le \frac{\pi(S)}{\pi(S^\star)} z^{s^\star - |S|} e^{-(\alpha/(2(1+\alpha)))\lambda_S},$$

where $z = \nu/(1 + \alpha)$. The irrepresentability result in Lemma 5 of Arias-Castro and Lounici [2] gives a lower bound on λ_S:

$$\lambda_S \ge \frac{1}{\sigma^2}\kappa\left(|S^\star \cap S^c|\right)^2\left\|\beta^\star_{S^\star \cap S^c}\right\|^2.$$

Monotonicity of κ implies that

$$\kappa\left(|S^\star \cap S^c|\right) \ge \kappa\left(s^\star\right) > 0$$

and, furthermore, by the beta-min condition,

$$\left\|\beta^\star_{S^\star \cap S^c}\right\|^2 \ge \rho_n^2\left(s^\star - |S|\right).$$

Putting everything together, including the definition of ρ_n, we get

$$\mathsf{E}_{\beta^\star}\big\{\pi^n(S)\big\} \leq \frac{\pi(S)}{\pi(S^\star)} z^{s^\star-|S|} e^{-(\alpha/(2(1+\alpha)))(1/\sigma^2)\kappa(s^\star)^2\rho_n^2(s^\star-|S|)}$$

$$= \frac{\pi(S)}{\pi(S^\star)}\big(zp^{-M}\big)^{s^\star-|S|}.$$

If we can show that the sum of our upper bound above, over all $S \subset S^\star$, vanishes, then we are done. Plugging in our complexity prior, we need to bound

$$\sum_{s<s^\star} \frac{\binom{s^\star}{s}\binom{p}{s^\star}}{\binom{p}{s}}\big(zcp^{a-M}\big)^{s^\star-s}.$$

where r is a constant that depends only on $(\alpha, \gamma, \sigma^2)$. Note that

$$\frac{\binom{s^\star}{s}\binom{p}{s^\star}}{\binom{p}{s}} = \binom{p-s}{p-s^\star} = \binom{p-s}{s^\star-s} \leq p^{s^\star-s}.$$

Then the summation can be bounded above by

$$\sum_{s<s^\star}\big(zcp^{a+1-M}\big)^{s^\star-s} \leq p^{a+1-M} \times O(1),$$

where the inequality follows from the formula for partial sums of a geometric series. Since $M > a+1$, the upper bound vanishes, completing the proof.

Appendix B: Justification of claims in Remark 1

Consider a prior $\tilde{\pi}$ for the model S of the form

$$\tilde{\pi}(S) = (1-w_n)\pi(S) + w_n\delta_{S_0}(S),$$

where π is a prior on models of size $|S| \leq R$, for $R = \min\{n, \mathrm{rank}(X)\}$, S_0 is a fixed model with $|S_0| = R$ and $\mathrm{span}(X_{S_0}) = \mathrm{span}(X)$, and $w_n = e^{-rR}$ for $r > 0$ to be determined; a similar setup is taken in Gao, van der Vaart and Zhou [15], Section 5.10, and the choice $w_n \equiv 1/2$ is considered in Rigollet and Tsybakov [31], equation (3.4). With the prior $\tilde{\pi}$, it is easy to see that the denominator D_n of the posterior satisfies

$$D_n \geq \tfrac{1}{2}\max\big\{\pi\big(S^\star\big)g\big(S^\star\big), w_n g(S_0)\big\}$$

for sufficiently large n (so that $1 - w_n > \frac{1}{2}$, say), where

$$g(S) = e^{(\alpha/(2\sigma^2))\{\|X_{S^\star}(\hat{\beta}_{S^\star}-\beta_{S^\star}^\star)\|^2+\|Y-X_{S^\star}\hat{\beta}_{S^\star}\|^2-\|Y-X_S\hat{\beta}_S\|^2\}}\left(1+\frac{\alpha}{\gamma\sigma^2}\right)^{-|S|/2}.$$

For the case $S = S^\star$, in the first term of the maximum, we have

$$g\big(S^\star\big) = e^{(\alpha/(2\sigma^2))\|X_{S^\star}(\hat{\beta}_{S^\star} - \beta^\star_{S^\star})\|^2}\left(1 + \frac{\alpha}{\gamma\sigma^2}\right)^{-|S^\star|/2},$$

just like in the proof of Lemma 1. Hence,

$$g\big(S^\star\big) \geq \left(1 + \frac{\alpha}{\gamma\sigma^2}\right)^{-|S^\star|/2}.$$

For the second term, since $|S| \leq R$, we have

$$g(S_0) \geq e^{(\alpha/(2\sigma^2))\{\|X_{S^\star}(\hat{\beta}_{S^\star} - \beta^\star_{S^\star})\|^2 + \|Y - X_{S^\star}\hat{\beta}_{S^\star}\|^2 - \|Y - X_{S_0}\hat{\beta}_{S_0}\|^2\}}\left(1 + \frac{\alpha}{\gamma\sigma^2}\right)^{-R/2}.$$

Since the span of X_{S_0} contains that of $X_{S^\star}$, by assumption, we have that

$$\|Y - X_{S^\star}\hat{\beta}_{S^\star}\|^2 \geq \|Y - X_{S_0}\hat{\beta}_{S_0}\|^2$$

and, consequently, the term in the exponent above is bigger than $\frac{\alpha}{2\sigma^2}\|X_{S^\star}(\hat{\beta}_{S^\star} - \beta^\star_{S^\star})\|^2$, which is obviously positive. Therefore, the second term in the maximum is

$$\geq w_n e^{-cR} e^{(\alpha/(2\sigma^2))\|X_{S^\star}(\hat{\beta}_{S^\star} - \beta^\star_{S^\star})\|^2} > w_n e^{-cR} = e^{-(r+c)R},$$

where $c = \frac{1}{2}\log(1 + \frac{\alpha}{\gamma\sigma^2})$ is as in Lemma 1. Finally, for $A = r + c$, we have

$$D_n \geq \tfrac{1}{2}\max\big\{\pi\big(S^\star\big)e^{-c|S^\star|}, e^{-AR}\big\}$$

with probability 1, for large n, as desired.

We claim that, with this new prior $\tilde{\pi}$, the posterior can achieve the minimax rate for the prediction loss under both the ordinary and ultra high-dimensional regimes. That is, we get the optimal rate

$$\varepsilon_n = \min\big\{R, s^\star \log\big(p/s^\star\big)\big\}.$$

This follows easily from the denominator bound discussed above, so long as our current numerator bound from Lemma 2 also holds for the new prior. The majority of the proof of Lemma 2 has nothing to do with the model prior, so we can immediately jump to the following conclusion:

$$\mathsf{E}_{\beta^\star}(N_n) \leq e^{-k\varepsilon_n}\sum_{S:|S|\leq R}\varphi^{|S|}\tilde{\pi}(S),$$

where d and φ are as in the proof of Lemma 2. Now, for the weighted average part, we have

$$\sum_S \varphi^{|S|}\tilde{\pi}(S) \leq \sum_S \varphi^{|S|}\pi(S) + w_n\varphi^R.$$

The first term in this upper bound is just like that in the proof of Lemma 2, so we have a handle on this. We need to choose w_n in such a way that the second term is also controlled. Since $w_n = e^{-rR}$ for some $r > 0$, it follows that we need $r \geq \log \varphi$. With this choice, the optimal minimax rate can be achieved in both ordinary and ultra high-dimensional regimes.

We claimed in Remark 1 that there is a price to be paid, in terms of model selection performance, if one uses the prior $\tilde{\pi}$ discussed above. The problem is that the weight w_n assigned to the large model S_0, with $|S_0| = R > s^\star$, is considerably larger than the weight $(1 - w_n)\pi(S^\star)$ assigned to the true model $S^\star$. Then the corresponding posterior mass assigned to S_0 is too large, large enough to pull the posterior away from the true model, leading to inconsistency.

Acknowledgements

References

[1] Abramovich, F. and Grinshtein, V. (2010). MAP model selection in Gaussian regression. *Electron. J. Stat.* **4** 932–949. MR2721039

[2] Arias-Castro, E. and Lounici, K. (2014). Estimation and variable selection with exponential weights. *Electron. J. Stat.* **8** 328–354. MR3195119

[3] Barbieri, M.M. and Berger, J.O. (2004). Optimal predictive model selection. *Ann. Statist.* **32** 870–897. MR2065192

[4] Barron, A.R. and Cover, T.M. (1991). Minimum complexity density estimation. *IEEE Trans. Inform. Theory* **37** 1034–1054. MR1111806

[5] Bondell, H.D. and Reich, B.J. (2012). Consistent high-dimensional Bayesian variable selection via penalized credible regions. *J. Amer. Statist. Assoc.* **107** 1610–1624. MR3036420

[6] Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data*: *Methods*, *Theory and Applications*. *Springer Series in Statistics*. Heidelberg: Springer. MR2807761

[7] Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351. MR2382644

[8] Castillo, I., Schmidt-Hieber, J. and van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *Ann. Statist.* **43** 1986–2018. MR3375874

[9] 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

[10] Chen, J. and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95** 759–771. MR2443189

[11] Clyde, M. and George, E.I. (2004). Model uncertainty. *Statist. Sci.* **19** 81–94. MR2082148

[12] Dalalyan, A.S. and Tsybakov, A.B. (2008). Aggregation by exponential weighting, sharp PAC-Bayesian bounds, and sparsity. *Mach. Learn.* **72** 39–61.

[13] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

[14] Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statist. Sinica* **20** 101–148. MR2640659

[15] Gao, C., van der Vaart, A.W. and Zhou, H.H. (2015). A general framework for Bayes structured linear models. Unpublished manuscript. Available at arXiv:1506.02174.

[16] George, E.I. and McCullogh, R.E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88** 881–889.

[17] Ghosal, S., Ghosh, J.K. and van der Vaart, A.W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531. MR1790007

[18] Grünwald, P. and van Ommen, T. (2014). Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. Unpublished manuscript. Available at arXiv:1412.3730.

[19] Heaton, M.J. and Scott, J.G. (2010). Bayesian Computation and the Linear Model. In *Frontiers of Statistical Decision Making and Bayesian Analysis* (M.-H. Cheh, D. Dey, P. Müller, D. Sun and K. Ye, eds.) 527–545. New York: Springer.

[20] Ishwaran, H. and Rao, J.S. (2005). Spike and slab gene selection for multigroup microarray data. *J. Amer. Statist. Assoc.* **100** 764–780. MR2201009

[21] Ishwaran, H. and Rao, J.S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *Ann. Statist.* **33** 730–773. MR2163158

[22] James, G.M. and Radchenko, P. (2009). A generalized Dantzig selector with shrinkage tuning. *Biometrika* **96** 323–337. MR2507146

[23] James, G.M., Radchenko, P. and Lv, J. (2009). DASSO: Connections between the Dantzig selector and lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 127–142. MR2655526

[24] Jiang, W. (2007). Bayesian variable selection for high dimensional generalized linear models: Convergence rates of the fitted densities. *Ann. Statist.* **35** 1487–1511. MR2351094

[25] 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

[26] Johnson, V.E. and Rossell, D. (2012). Bayesian model selection in high-dimensional settings. *J. Amer. Statist. Assoc.* **107** 649–660. MR2980074

[27] Martin, R. and Walker, S.G. (2014). Asymptotically minimax empirical Bayes estimation of a sparse normal mean vector. *Electron. J. Stat.* **8** 2188–2206. MR3273623

[28] Narisetty, N.N. and He, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *Ann. Statist.* **42** 789–817. MR3210987

[29] Park, T. and Casella, G. (2008). The Bayesian lasso. *J. Amer. Statist. Assoc.* **103** 681–686. MR2524001

[30] Reid, S., Tibshirani, R. and Friedman, J. (2014). A study of error variance estimation in lasso regression. Unpublished manuscript. Available at arXiv:1311.5274.

[31] Rigollet, P. and Tsybakov, A. (2011). Exponential screening and optimal rates of sparse estimation. *Ann. Statist.* **39** 731–771. MR2816337

[32] Rigollet, P. and Tsybakov, A.B. (2012). Sparse estimation by exponential weighting. *Statist. Sci.* **27** 558–575. MR3025134

[33] Shen, X. and Wasserman, L. (2001). Rates of convergence of posterior distributions. *Ann. Statist.* **29** 687–714. MR1865337

[34] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **58** 267–288. MR1379242

[35] Walker, S. and Hjort, N.L. (2001). On Bayesian consistency. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 811–821. MR1872068

[36] Walker, S.G., Lijoi, A. and Prünster, I. (2007). On rates of convergence for posterior distributions in infinite-dimensional models. *Ann. Statist.* **35** 738–746. MR2336866

[37] Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3000 shared controls. *Nature* **447** 661–678.

[38] Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques. Stud. Bayesian Econometrics Statist.* **6** 233–243. North-Holland, Amsterdam. MR0881437

[39] Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist.* **36** 1567–1594. MR2435448

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

[41] Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469

[42] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 301–320. MR2137327