Empirical Bayes posterior concentration in sparse high-dimensional linear models

Ryan Martin    Raymond Mess
Department of Mathematics, Statistics, and Computer Science
University of Illinois at Chicago
(rgmartin, rmess1)@uic.edu

Stephen G. Walker
Department of Mathematics
University of Texas at Austin
s.g.walker@math.utexas.edu

July 1, 2014

Abstract

We propose an empirical Bayes approach for inference in the $p \gg n$ normal linear model. Assuming the regression coefficients are sparse, in the sense that no more than $n$ of them are non-zero, our empirical Bayes posterior distribution for the regression coefficients concentrates at the frequentist minimax rate for suitable priors on the model size. Model selection consistency is established, and simulation studies show the strong finite-sample performance of our method.

Keywords and phrases: Data-dependent prior; fractional likelihood; minimax; regression; variable selection.

1 Introduction

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

$$Y = X\beta + \varepsilon,$$

where $Y$ is a $n \times 1$ vector of response variables, $X$ is a $n \times p$ matrix of predictor variables, $\beta$ is a $p \times 1$ vector of slope coefficients, and $\varepsilon$ is a $n \times 1$ vector of iid $\mathcal{N}(0, \sigma^2)$ random errors. We assume a fixed design, so $X$ is non-stochastic. 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 phenotype, or observable trait, the number of subjects, $n$, is 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, the number which have a genuine association with the trait
is generally quite small. For example, the Wellcome Trust (2007) has confirmed that only seven genes have a non-negligible association with Type I diabetes. Therefore, it is reasonable to assume that $\beta$ is sparse, i.e., 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 1996), the smoothly clipped absolute deviation (Fan and Li 2001), the adaptive lasso (Zou 2006), and the Dantzig selector (Candes and Tao 2007; James and Radchenko 2009; James et al 2009). Fan and Li (2010) 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 McCulloch 1993) and the methods based on spike-and-slab priors (Ishwaran and Rao 2005a,b). These methods and others are reviewed in Clyde and George (2004) and Heaton and Scott (2010). More recently, Bondell and Reich (2012), Johnson and Rossell (2012), and Narisetty and He (2014) 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 $\beta$. It is natural to ask under what conditions will the $\beta$ posterior distribution concentrate around the true value at an appropriate or optimal rate. Castillo and van der Vaart (2012) considered this question in the special case where $n = p$ and $X$ is the identity matrix. They showed that, for suitable priors, the posterior distribution concentrates around the truth at the frequentist minimax rate. Recently, their analysis was extended in Castillo et al (2014) to the general $p \gg n$ regression context discussed above. In particular, they propose a model with a certain sparsity prior on the model (see Section 2.1) and a Laplace prior on the $\beta$ values included by the model, making connections to the lasso (e.g. Park and Casella 2008). For example, they show that, for this prior, under conditions on the lasso tuning parameter and on the compatibility number of the design matrix $X$, the posterior distribution concentrates at the lasso’s frequentist oracle rate. To our knowledge, these are the first Bayesian posterior concentration rate results in the high-dimensional linear model setting. However, their rates apparently fall short of the minimax rates in Verzelen (2012)

Recently, for the special case of (1) where $p = n$ and $X$ is the identity matrix, Martin and Walker (2013) propose an empirical Bayes model and prove that the corresponding empirical Bayes posterior concentrates at the frequentist minimax rate. In this paper, we generalize their empirical Bayes approach to the case of the sparse high-dimensional linear model. Section 2 describes the prior and the corresponding empirical Bayes model, and in Section 3 we prove various concentration rate results for our empirical Bayes posterior about the sparse truth. For certain priors, the rates are optimal in the sense they match the minimax rates. We also give model selection consistency results, identifying the parameters our posterior asymptotically flags as being non-zero. 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 demonstrate the strong finite-sample performance of our method compared to others in terms of model selection. Finally Section 5 gives a brief a discussion, the key message being that we get provably optimal concentration results, fast and easy computation,
and strong finite-sample performance. Proofs of some auxiliary results are given in the Appendix.

2 The empirical Bayes model

2.1 The prior

Here, and in our theoretical analysis in Section 3, we take the error variance $\sigma^2$ to be known, as do Castillo et al. (2014). Techniques for estimating $\sigma^2$ in the high-dimensional case are available (Giraud et al. 2012) and, in Section 4, we consider a particular plug-in estimate provided by Reid et al. (2014). To specify a prior for $\beta$ that incorporates sparsity, we decompose $\beta$ as $(S, \beta_S)$, where $S \subset \{1, \ldots, p\}$ denotes the “active set” of variables, $S = \{j : \beta_j \neq 0\}$, and $\beta_S$ is the $|S|$-vector containing the particular non-zero values. Based on this decomposition, we can specify the prior for $\beta$ in two steps: a prior for $S$ and then a prior for $\beta_S$, given $S$. We discuss both of these priors in turn, but we should mention at this point that the prior on $S$ is of primary importance, encouraging the posterior to focus on models of the appropriate size; the conditional prior on $\beta_S$ apparently has less of an effect (Martin and Walker 2013, Remark 1).

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

$$
\pi(S) = \frac{1}{(\frac{p}{2})^s} \cdot f_n(s), \quad s = 0, 1, \ldots, p, \quad s = |S|, \quad (2)
$$

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. For now, we focus on the key novelty of our approach – a prior size restriction on the model. That is, we assume that $f_n(s)$ is supported on $\{0, 1, \ldots, n\}$, not on $\{0, 1, \ldots, p\}$. In other words, we set

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

Our primary motivation for imposing this constraint is that in practical applications, the true value of $s$; i.e. $s^* = |S^*|$, is typically smaller than $n$; and often $s$ is small, such as the seven in Wellcome Trust (2007). For example, suppose the true model $S^*$ were known. Even in this ideal case, if $|S^*| > n$, then good estimation of the corresponding parameters would not be possible since the design matrix would be of insufficient rank to admit a unique solution to the normal equations. 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.

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

$$
\beta_S \mid S \sim N_{|S|}(\hat{\beta}_S, \gamma^{-1}(X_S^TX_S)^{-1});
$$
here, $X_S$ is the columns of $X$ corresponding to $S$, and $\gamma > 0$ is a tuning parameter, to be specified. To summarize, our proposed prior $\Pi$ for $\beta$ is given by

$$
\Pi(d\beta) = \sum_{S:|S|\leq n} N_{|S|}(d\beta_S \mid \hat{\beta}_S, \gamma^{-1}(X_S^\top X_S)^{-1}) \delta_0(d\beta_{S^c}) \pi(S). \tag{4}
$$

Note that the prior for $\beta_S$ depends on the data. Following Martin and Walker (2013), we refer to this as an “empirical Bayes” prior; see Section 2.3. The restriction $|S| \leq n$ allows the least squares estimator $\hat{\beta}_S$ to be available and its associated distribution theory holding. As will be made clear in what follows, the tuning parameter $\gamma$ will be quite small, making $\gamma^{-1}$ quite large. This means that the conditional prior for $\beta_S$ is diffuse, so the dependence on the data, through $\hat{\beta}_S$, is not overly strong.

Obviously, in the above construction of the conditional prior for $\beta_S$, we are assuming that $X_S^\top X_S$ is non-singular for all subsets $S$ of size no more than $n$. This would hold, for example, if $X$ satisfies the “sparse Riesz condition with rank $n$” discussed in Zhang and Huang (2008) and also in Chen and Chen (2008). If there exists some $S$ with $|S| \leq n$ such that $X_S^\top X_S$ is singular, then we can replace the regular matrix inverse in (4) with the unique Moore–Penrose generalized inverse (e.g., Seber and Lee 2003, Appendix A.10), and all of the theory presented below carries through, provided that each $X_S$, with $|S| \leq n$ has rank proportional to $|S|$. We will not consider this case any further here.

2.2 The likelihood function

For the likelihood function, write $L_n(\beta) = N_n(Y \mid X\beta, \sigma^2 I)$ 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 $\beta$. Another interesting feature of the proposed approach is that we shall consider 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 Walker and Hjort (2001) and Martin and Walker (2013). Our theory does not allow for the case $\alpha = 1$, but $\alpha$ can be arbitrarily close to 1. For finite-samples, the difference between numerical results for $\alpha \approx 1$ and $\alpha = 1$ are negligible. Also, as will be made clear in what follows, if $\alpha \to 1$, then $\gamma \to 0$, making the conditional prior for $\beta_S$ more diffuse. Therefore, the case of a genuine Bayes model with non-informative flat conditional prior for $\beta_S$ is a limiting case of our model.

2.3 The posterior distribution

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

$$
\Pi^n(B) = \frac{\int_B L_n(\beta)^\alpha \Pi(d\beta)}{\int_{\mathbb{R}^p} L_n(\beta)^\alpha \Pi(d\beta)}. \tag{5}
$$

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

We interpret “empirical Bayes” loosely—if the prior depends on data, then it is empirical Bayes. The combination of a prior, data-dependent or not, with a fractional
likelihood via Bayes formula can also be reinterpreted as genuinely empirical Bayes. The idea is that
\[ L_n(\beta)^\alpha \Pi(d\beta) = L_n(\beta) \frac{\Pi(d\beta)}{L_n(\beta)^{1-\alpha}}, \]
i.e., the Bayes combination of a fractional likelihood with a prior is equivalent to a Bayes combination of the correct likelihood function with a data-dependent prior. As Walker and Hjort (2001) explain, rescaling the prior by a portion of the likelihood helps to protect from possible inconsistencies, penalizing those parameter values with too high a likelihood, namely those that “track the data too closely.” In the next section we show that our empirical Bayes posterior also has desirable large-sample properties, though our results are not consequences of those in Walker and Hjort (2001).

3 Posterior concentration rates

3.1 Setup

In this section, we explore the concentration properties of the empirical Bayes posterior constructed in the previous section. Before getting into these details, we first want to clarify what is meant by asymptotics in this context. There is an implicit triangular array setup, i.e., for each \( n \), the response vector \( Y^n = (Y^n_1, \ldots, Y^n_n)^\top \) is modeled by the Gaussian linear regression \((\Pi)\) with the \( n \times p \) matrix of predictors \( X^n = ((X^n_{ij})) \) 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 \( \beta^n \) can be learned, provided that it is sufficiently sparse. In what follows, we will use the standard notation in \((\Pi)\) which is less cumbersome but hides the triangular array formulation.

There is an extreme version of the “\( p \gg n \)” problem that will be of interest here. The “ultra high-dimensional” problem is one such that \( s^* \log(p/s^*) \gg n \), where \( s^* \) denotes the size of the true model (Verzelen 2012, p. 40). Often, this ultra high-dimensional situation arises when \( p \) is exponentially large, i.e., \( \log p \propto n \), but Verzelen also mentions cases with \( p \) polynomial in \( n \) and \( s^* \) is close to \( n \). Recall we assume \( s^* \leq n \).

3.2 Preliminary result

As is standard, we will assume that the observed data \( Y \) is generated by the linear model \((\Pi)\) with true \( p \)-vector of regression coefficients, \( \beta^* \), having active set \( S^* = S_{\beta^*} \) of cardinality \( s^* = |S^*| \); all these quantities depend on \( n \), but this dependence is suppressed in the notation. Let \( B \) be a generic event for \( \beta \in \mathbb{R}^p \). Our empirical Bayes posterior probability of the event \( B \) in \((\Pi)\) can be rewritten as
\[ \Pi^n(B) = \frac{\int_B R_n(\beta, \beta^*)^\alpha \Pi(d\beta)}{\int R_n(\beta, \beta^*)^\alpha \Pi(d\beta)}, \]
where \( R_n(\beta, \beta^*) = L_n(\beta)/L_n(\beta^*) \) is the likelihood ratio. Let \( D_n \) denote the denominator in the above display, i.e., \( D_n = \int R_n(\beta, \beta^*)^\alpha \Pi(d\beta) \). The next result, which will be useful throughout our analysis, gives a lower bound on \( D_n \).
Lemma 1. Fix $\alpha \in (0, 1)$ and choose $\gamma$ such that $\log(1 + \frac{\alpha}{\gamma}) \geq \frac{\alpha}{\gamma}$. If $s^* \leq n$, there exists $c = c(\alpha, \gamma) > 0$ such that $D_n \geq \pi(S^*)e^{-cs^*}$ for all large $n$, with probability 1.

Proof. See Appendix A.1

3.3 Prediction loss

We now present a result characterizing the concentration rate of the posterior distribution for the mean $X\beta$. Our result is comparable to Theorem 11 and the first part of Theorem 2 in Castillo et al. (2014), though our setup is simpler. Set

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

(7)

where $\varepsilon_n$ 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 (2011), e.g., their equation (2.8), $\varepsilon_n$ proportional to $s^*\log p$ corresponds the oracle convergence rate for the lasso estimator. Verzelen (2012) shows that the minimax risk for this case corresponds to $\varepsilon_n$ proportional to $s^*\log(p/s^*)$. Intuitively, if $S^*$ were known, then the best rate for the prediction error would be $s^*$, 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 $B_{\varepsilon_n}$, as in (3), i.e.,

$$N_n = \int_{B_{\varepsilon_n}} R_n(\beta, \beta^*)^\alpha \Pi(d\beta).$$

We have the following bound on $N_n$.

Lemma 2. Fix $\alpha \in (0, 1)$. Then there exists $d = d(\alpha) > 0$ and $\varphi = \varphi(\alpha, \gamma) > 1$ such that $E_{\beta^*}(N_n) \leq e^{-dn} \sum_{S:|S|\leq n} \varphi^{|S|}\pi(S)$, uniformly in $\beta^*$ with $|S_{\beta^*}| \leq n$.

Proof. See Appendix A.2

Following the intuition in Section 2.3, we take $\alpha$ close to 1. In Lemma 2, this makes the constant $d$ small. Also, if $\alpha$ is close to 1, then $\gamma$ can be chosen small enough so that the constant $\varphi$ is not too big, i.e., close to $2^{1/2}$. Justification for these statements can be found in the proof of Lemma 2 and the choice of $\alpha$ and $\gamma$ is considered in Section 4.

To combine the results in Lemmas 1 and 2, we use the following argument. Let $b_n = \pi(S^*)e^{-cs^*}$ be the bound in Lemma 1. Using the indicator function $1(\cdot)$, write

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

Taking expectation, we get

$$E_{\beta^*}\{\Pi^\alpha(B_{\varepsilon_n})\} \leq \frac{E_{\beta^*}(N_n)}{b_n} + P_{\beta^*}(D_n < b_n).$$

The latter probability is $o(1)$, as $n \to \infty$, from Lemma 1. So, plugging in the specific bounds from Lemmas 1 and 2 we get

$$E_{\beta^*}\{\Pi^\alpha(B_{\varepsilon_n})\} \leq e^{|S^*| - dn} \frac{1}{\pi(S^*)} \sum_{S} \varphi^{|S|}\pi(S) + o(1)$$

$$= e^{cs^* - de_n} \frac{p_n}{f_n(s^*)} \sum_{s=0}^{n} \varphi^sf_n(s) + o(1);$$

$$= e^{cs^* - de_n} \frac{p_n}{f_n(s^*)} \sum_{s=0}^{n} \varphi^sf_n(s) + o(1);$$

6
this holds uniformly in $\beta^*$ such that $|S^*| \leq n$. The empirical Bayes concentration rate $\varepsilon_n$ is such that the first term in the above upper bound vanishes. If we set

$$\zeta_n = \left(\frac{p}{s^*}\right) f_n(s^*) \sum_{s=0}^{n} \varphi^s f_n(s),$$

then the empirical Bayes posterior concentration rate $\varepsilon_n$ satisfies

$$\log \zeta_n = O(\varepsilon_n), \quad \text{as } n \to \infty. \tag{8}$$

This amounts to a condition on the prior $f_n$ for $|S|$. Indeed, (8) requires that $f_n$ should be sufficiently concentrated near $s^*$, so that $f_n(s^*)$ is not too small and the expectation of $\varphi^{|S|}$ with respect to $f_n$ is not too big. Compare this to the standard prior support conditions needed for posterior concentration (Ghosal et al. 2000; Shen and Wasserman 2001; Walker et al. 2007). We summarize this discussion in the following theorem.

**Theorem 1.** Choose $\alpha \in (0, 1)$ and $\gamma > 0$ as in Lemmas 1 and 2. If the prior $f_n$ on $|S|$ admits $\zeta_n$ such that (8) holds with $\varepsilon_n$, then there exists a constant $M > 0$ such that

$$E_{\beta^*}\{\Pi^n(B_{M\varepsilon_n})\} \to 0 \quad \text{as } n \to \infty,$$

uniformly over all $\beta^*$ with $|S_{\beta^*}| \leq n$.

**Proof.** By Lemmas 1 and 2, and the growth condition (8), we have that, for large $n$,

$$\log E_{\beta^*}\{\Pi^n(B_{M\varepsilon_n})\} \leq \left(\frac{c_{S^*}}{M\varepsilon_n} - d + \frac{\log \zeta_n}{M\varepsilon_n}\right) M\varepsilon_n.$$ 

The first term inside the parentheses is bounded, i.e., $\limsup_n (s^*/\varepsilon_n) \leq J$. Assuming $s^* \leq \varepsilon_n$ then $J$ can be taken as zero. Next, under (8), there exists a $K > 0$ such that $(\log \zeta_n)/\varepsilon_n < K$. So, if we take $M$ such that $dM > \max\{K, cJ\}$, then the upper bound above goes to $-\infty$ as $n \to \infty$. This implies the result. \hfill \Box

**Remark 1.** The growth condition (8) holds with $\varepsilon_n$ proportional to the minimax rate if there exists constants $a_1$, $a_2$, $c_1$, $c_2$, $C_1$, and $C_2$ such that $f_n$ satisfies

$$C_1 \left(\frac{1}{c_1 p^a_1}\right)^s \leq f_n(s) \leq C_2 \left(\frac{1}{c_2 p^a_2}\right)^s \quad \text{for all } s = 1, \ldots, n \tag{9}$$

The proof of this claim follows from calculations similar to those in Example 1 below. Assumption 1 in Castillo et al. (2014) implies (9), but our restriction to models with $|S| \leq n$ allows us to get the minimax rate for priors that may not satisfy (9).

**Remark 2.** In the ultra high-dimensional case, the rate $s^* \log(p/s^*)$ would be much larger than $n$. In such cases, it is sufficient to apply the trivial bound $\sum_{s=0}^{n} \varphi^s f_n(s) \leq \varphi^n = e^n \log \varphi$, since a term of order $n$ in the exponent is negligible.

The next two examples give priors $f_n$ for which the corresponding empirical Bayes posterior distribution concentrates at the minimax rate. Our focus here will be the ultra high-dimensional setting, so that the bound in Remark 2 can be applied.

**Example 1.** The complexity prior for the model size $|S|$ in equation (2.3) of Castillo et al. (2014) is given by

$$f_n(s) \propto c^{-s} p^{-as}, \quad s = 0, 1, \ldots, n, \tag{10}$$

$$\zeta_n = \left(\frac{p}{s^*}\right) f_n(s^*) \sum_{s=0}^{n} \varphi^s f_n(s),$$

then the empirical Bayes posterior concentration rate $\varepsilon_n$ satisfies

$$\log \zeta_n = O(\varepsilon_n), \quad \text{as } n \to \infty. \tag{8}$$

This amounts to a condition on the prior $f_n$ for $|S|$. Indeed, (8) requires that $f_n$ should be sufficiently concentrated near $s^*$, so that $f_n(s^*)$ is not too small and the expectation of $\varphi^{|S|}$ with respect to $f_n$ is not too big. Compare this to the standard prior support conditions needed for posterior concentration (Ghosal et al. 2000; Shen and Wasserman 2001; Walker et al. 2007). We summarize this discussion in the following theorem.

**Theorem 1.** Choose $\alpha \in (0, 1)$ and $\gamma > 0$ as in Lemmas 1 and 2. If the prior $f_n$ on $|S|$ admits $\zeta_n$ such that (8) holds with $\varepsilon_n$, then there exists a constant $M > 0$ such that

$$E_{\beta^*}\{\Pi^n(B_{M\varepsilon_n})\} \to 0 \quad \text{as } n \to \infty,$$

uniformly over all $\beta^*$ with $|S_{\beta^*}| \leq n$.

**Proof.** By Lemmas 1 and 2, and the growth condition (8), we have that, for large $n$,

$$\log E_{\beta^*}\{\Pi^n(B_{M\varepsilon_n})\} \leq \left(\frac{c_{S^*}}{M\varepsilon_n} - d + \frac{\log \zeta_n}{M\varepsilon_n}\right) M\varepsilon_n.$$ 

The first term inside the parentheses is bounded, i.e., $\limsup_n (s^*/\varepsilon_n) \leq J$. Assuming $s^* \leq \varepsilon_n$ then $J$ can be taken as zero. Next, under (8), there exists a $K > 0$ such that $(\log \zeta_n)/\varepsilon_n < K$. So, if we take $M$ such that $dM > \max\{K, cJ\}$, then the upper bound above goes to $-\infty$ as $n \to \infty$. This implies the result. \hfill \Box

**Remark 1.** The growth condition (8) holds with $\varepsilon_n$ proportional to the minimax rate if there exists constants $a_1$, $a_2$, $c_1$, $c_2$, $C_1$, and $C_2$ such that $f_n$ satisfies

$$C_1 \left(\frac{1}{c_1 p^a_1}\right)^s \leq f_n(s) \leq C_2 \left(\frac{1}{c_2 p^a_2}\right)^s \quad \text{for all } s = 1, \ldots, n \tag{9}$$

The proof of this claim follows from calculations similar to those in Example 1 below. Assumption 1 in Castillo et al. (2014) implies (9), but our restriction to models with $|S| \leq n$ allows us to get the minimax rate for priors that may not satisfy (9).

**Remark 2.** In the ultra high-dimensional case, the rate $s^* \log(p/s^*)$ would be much larger than $n$. In such cases, it is sufficient to apply the trivial bound $\sum_{s=0}^{n} \varphi^s f_n(s) \leq \varphi^n = e^n \log \varphi$, since a term of order $n$ in the exponent is negligible.

The next two examples give priors $f_n$ for which the corresponding empirical Bayes posterior distribution concentrates at the minimax rate. Our focus here will be the ultra high-dimensional setting, so that the bound in Remark 2 can be applied.

**Example 1.** The complexity prior for the model size $|S|$ in equation (2.3) of Castillo et al. (2014) is given by

$$f_n(s) \propto c^{-s} p^{-as}, \quad s = 0, 1, \ldots, n, \tag{10}$$
where \(a\) and \(c\) are positive constants. This prior clearly satisfies the condition (3) in Remark 1. We claim that this complexity prior satisfies (5) with \(\varepsilon_n = s^* \log(p/s^*)\). To see this, note that
\[
\log f_n(s^*) \leq -s^* \log(cs^*a) - as^* \log(p/s^*) = -\left(a + \frac{\log c + a \log s^*}{\log(p/s^*)}\right)s^* \log(p/s^*).
\]
The ratio inside the parentheses above vanishes since \(s^* \ll p\). Similarly, by Stirling’s formula, we have that
\[
\log \binom{p}{s^*} \leq s^* \log(p/s^*) \{1 + o(1)\}.
\]
Putting these two bounds together, and using the result in Remark 2, we can conclude that the complexity prior above yields a posterior concentration rate proportional to the minimax rate \(s^* \log(p/s^*)\).

**Example 2.** Next consider a beta–binomial prior for \(|S|\), i.e.,
\[
f_n(s) = \int_0^1 \binom{n}{s} w^{n-s}(1-w)^s a_n w^{a_n-1} dw,
\]
which corresponds to a Beta\((a_n, 1)\) prior for \(W\) and a conditional Bin\((n, 1-w)\) prior for \(|S|\), given \(W = w\). Here we take \(a_n = an\) for a constant \(a > 0\). Following the calculations in Martin and Walker (2013), we have
\[
\frac{f_n(s^*)}{\binom{n}{s^*}} = an \int_0^{1-s^*/n} w^{n-s^*+an-1}(1-w)^{s^*} \, dw
\]
\[
> an \int_0^{1-s^*/n} w^{n-s^*+an-1}(1-w)^{s^*} \, dw
\]
\[
> \frac{an}{n-s^*+an} \binom{s^*}{n} \left(\frac{1-s^*}{n}\right)^{n-s^*+an}
\]
\[
> \frac{a}{a+1} \left(\frac{s^*}{n}\right)^{2s^*} \left(1 - \frac{s^*}{n}\right)^{an}.
\]
When \(n\) is large, the right-most term in the above inequality is of the order \(e^{-as^*}\). Next, recall the general inequalities satisfied by the binomial coefficients:
\[
\binom{n}{s} \leq \left(\frac{n}{s}\right)^s \leq \left(\frac{n e}{s}\right)^s, \quad s = 1, \ldots, n.
\]
From these inequalities, we clearly have
\[
\frac{\binom{n}{s}}{\binom{p}{s^*}} \geq \left(\frac{n}{p}\right)^{s^*} e^{-s^*}.
\]
Therefore, for large \(n\),
\[
\frac{f_n(s^*)}{\binom{p}{s^*}} = \frac{f_n(s^*)}{\binom{n}{s^*}} \cdot \left(\frac{n}{s^*}\right)
\]
\[
\geq \frac{a}{a+1} \left(\frac{s^*}{n}\right)^{2s^*} e^{-as^*} \cdot \left(\frac{n}{p}\right)^{s^*} e^{-s^*}
\]
\[
= \frac{a}{a+1} e^{-(a+1)s^*} \left(\frac{s^*}{p}\right)^{s^*} \left(\frac{s^*}{n}\right)^{s^*}
\]
\[
\geq \frac{a}{a+1} e^{-(a+1)s^*} \left(\frac{s^*}{p}\right)^{2s^*},
\]
where the last inequality holds since \( p \gg n \). Ignoring the expectation of \( \varphi^{[S]} \) in \( \zeta_n \), as it is a lower-order term in the ultra high-dimensional case (Remark 2), it is now easy to confirm that the beta–binomial prior \( f_n \) is such that (8) holds with \( \varepsilon_n \) proportional to the minimax rate \( s^* \log(p/s^*) \).

There are apparently reasonable priors whose corresponding posterior concentration rate is slower than the minimax rate.

**Example 3.** Let \( f_n \) be a Bin\((n, n^{-1})\) mass function. Using the same lower and upper bounds on the binomial coefficients as in the previous example, we have

\[
\frac{f_n(s^*)}{\binom{p}{s^*}} = \left( \frac{n}{s^*} \right) \left( \frac{n-1}{n} \right)^{n-s^*} \geq \left( \frac{n}{p} \right)^{s^*} e^{-s^*} \left( \frac{n-1}{n} \right)^{n-s^*} \geq p^{-s^*} e^{-s^*-1} \{1 + o(1)\}.
\]

For this prior, the expectation of \( \varphi^{[S]} \) is actually bounded as \( n \to \infty \), so, for high-dimensional or ultra high-dimensional cases, one finds that \( \log \zeta_n = O(s^* \log p) \). This matches the oracle inequality for the lasso quoted in Equation (2.8) of B"uhlmann and van de Geer (2011), but falls short of the minimax rate in the previous examples.

### 3.4 Effective dimension

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

For a given \( \Delta \), let \( B_n(\Delta) = \{ \beta \in \mathbb{R}^p : |S_\beta| \geq \Delta \} \) be those \( \beta \) vectors with no less than \( \Delta \) non-zero entries. We say that the effective dimension of \( \Pi_n \) 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.** Under the conditions of Lemma 2, \( \mathbb{E}_{\beta^\star} \{ N_n(\Delta) \} \leq \sum_{s=\Delta}^{n} \varphi^s f_n(s) \) for all \( \beta^\star \).

**Proof.** See Appendix A.3.

We can combine Lemma 3 and Lemma 1 to conclude that, for sufficiently large \( n \),

\[
\mathbb{E}_{\beta^\star} [\Pi^n \{ B_n(\Delta) \}] \leq e^{cs^\star} \left( \frac{p}{s^*} \right) \sum_{s=\Delta}^{n} \varphi^s f_n(s) + o(1),
\]
uniformly in $\beta^*$ with $|S_{\beta^*}| = s^*$. Then an asymptotic bound on the effective dimension is a $\Delta = \Delta_n$ such that the upper bound above vanishes as $n \to \infty$. Since $\varphi > 1$, we have $\sum_s \varphi^s f_n(s) > 1$ and, therefore,

$$E_{\beta^*} [\prod^n \{B_n(\Delta)\}] \leq e^{cs^* + \log \zeta_n} \sum_{s=\Delta}^n \varphi^s f_n(s) + o(1). \quad (11)$$

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 $s^*$ will be sufficiently 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.** Assume the conditions of Theorem 1. If $\Delta = \Delta_n$ is such that the upper bound in (11) vanishes as $n \to \infty$, then $\Delta_n$ bounds the effective dimension. Specifically, if $f_n$ satisfies condition (9), then the effective dimension is proportional to $s^*$.

**Proof.** The proof of the first claim is trivial, in light of the bound (11). For the second claim, first recall that, under (9), $\log \zeta_n$ is of the order $s^* \log (p/s^*)$, the minimax rate. Next, it is easy to check that, under (9), the summation $\sum_{s=\Delta}^n \varphi^s f_n(s)$ is upper-bounded by a partial sum of a geometric series. In particular, that summation is $O(r^{\Delta+1})$, where $r = \varphi/c_2 p^{a_2}$. In that case,

$$r^{\Delta+1} = e^{-(\Delta+1)[a_2 \log p + \log(c_2/\varphi)]}.$$ 

So, if $\Delta$ is a large multiple of $s^*$, then clearly the $r^{\Delta+1}$ term dominates the $e^{cs^* + \log \zeta_n}$ term. Therefore, the product vanishes, proving the claim. \hfill $\square$

**Remark 3.** If $\log \zeta_n$ is much larger than $n$, e.g., in the ultra high-dimensional case, then the trivial bound as in Remark 2 gives

$$E_{\beta^*} [\prod^n \{B_n(\Delta)\}] \leq e^{cs^* + \log \zeta_n + n \log \varphi} \sum_{s=\Delta}^n f_n(s) + o(1).$$

This is intuitive: if the prior probability of the event $B_n(\Delta)$ is sufficiently small, then we can expect that the corresponding posterior probability to be vanishingly small.

To summarize, our prior is such that the posterior distribution is supported on models of size no more than $n$. However, a good prior is one such that the posterior ought to be able to learn what is the size of the true model that generated the data, provided the latter is no more than $n$. Theorem 2 shows that, indeed, if the prior $f_n$ on the model size has sufficiently light tails, then the posterior will concentrate on models of size proportional to $s^*$, the true model size.

### 3.5 Other loss functions

The result in Theorem 1 concerns the empirical Bayes posterior probability of sets of $\beta$ which are near the true $\beta^*$ relative to a distance depending on the design matrix $X$. It is natural to ask whether the empirical Bayes posterior concentrates on neighborhoods of $\beta^*$ with respect to more natural metrics, ones that do not directly depend on the design
matrix, such as \( \ell_1 \)- and \( \ell_2 \)-norms. An affirmative answer will require further conditions on \( X \) to separate \( \beta \) from \( X\beta \).

In the low-dimensional case, with \( p < n \), if the matrix \( X \) is full rank, then

\[
\|X(\beta - \beta^*)\|_2 \geq \lambda_{\min}(X^\top X)\|\beta - \beta^*\|_2,
\]

where \( \lambda_{\min}(A) \) returns 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, in general, making the above inequality trivial and not useful. However, it is still possible to get something like the displayed inequality. Towards this, we need the function

\[
\kappa(s) = \kappa_X(s) = \inf_{\beta:0<|S_\beta|\leq s} \frac{\|X\beta\|_2}{\|X\|_2 \|\beta\|_2}, \quad s = 1, \ldots, p,
\]

where the \( \ell_2 \)-norm of the matrix \( X \) is the maximal diagonal element of \( X^\top X \), i.e.,

\[
\|X\|_2 = \max_{j=1,\ldots,p} (X^\top X)_{jj}^{1/2}.
\]

The quantity \( \kappa(s) \) is called the “smallest scaled sparse singular value of dimension \( s \)” in Definition 2.3 of Castillo et al. (2014). Its main purpose is to facilitate conversion of \( \ell_2 \)-norm concentration results for the mean vector \( X\beta \) to \( \ell_2 \)-norm concentration results for \( \beta \) itself. One can define quantities analogous to \( \kappa \) in order to get concentration results relative to the \( \ell_1 \) or \( \ell_\infty \)-norm of \( \beta \), but this is simple and we will not consider this here; see Castillo et al. (2014), Section 2.

The result presented below will follow immediately from Theorem 1 and the definition of \( \kappa \). Indeed, for any \( \beta \), we have

\[
\|X(\beta - \beta^*)\|_2 \geq \kappa(|S_{\beta - \beta^*}|) \|X\|_2 \|\beta - \beta^*\|_2.
\]

For example, if \( \|\beta - \beta^*\|_2 \) is lower-bounded, then so is \( \|X(\beta - \beta^*)\|_2 \), so a posterior concentration result for the \( \ell_2 \)-norm on \( \beta \) should follow from an analogous result for the \( \ell_2 \) prediction error as in Theorem 1. The only obstacle is that the \( \kappa \) term on the right-hand depends on the particular \( \beta \). The following result leads to the observation that \( \kappa(|S_{\beta - \beta^*}|) \) can be controlled by a term that depends only on \( s^* \).

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

**Proof.** Follows since \( \kappa \) is non-increasing and \( |S_{\beta - \beta^*}| \leq |S_\beta| + |S_{\beta^*}| \). \( \square \)

Under our prior formulation, we know that the posterior puts probability 1 on those \( \beta \) for which \( |S_\beta| \leq n \). So, if \( |S_{\beta^*}| = s^* \), then, trivially, \( \kappa(|S_{\beta - \beta^*}|) \geq \kappa(n + s^*) \). However, for better control on the \( \kappa \) term, we prefer to have a sharper bound on \( |S_\beta| \); Theorem 2 provides such a bound. Indeed, Theorem 2 states that, for large \( n \), the posterior probability of the event \( \{|S_\beta| \geq C s^*\} \) tends to be small. Therefore, we expect that, for some constant \( C' \),

\[
\kappa(|S_{\beta - \beta^*}|) \geq \kappa(C' s^*)
\]

holds for all \( \beta \) in a set with high posterior probability. Compare this to the result in Theorem 1 of Castillo et al. (2014), and also to the corresponding model selection results for frequentist point estimators in, e.g., Bühlmann and van de Geer (2011, Chap. 7).
We are now ready for the concentration rate result with respect to the \(\ell_2\)-norm loss on the parameter \(\beta\) itself. This time, set

\[
B'_{\delta_n} = \{\beta \in \mathbb{R}^p : \|\beta - \beta^*\|^2_2 > \delta_n\},
\]

where \(\delta_n\) is a positive sequence to be specified.

**Theorem 3.** Assume the conditions of Theorem [1]. Suppose that the prior \(f_n\) satisfies [9], so that Theorem [1] holds with \(\varepsilon_n\) equal to \(s^* \log(p/s^*)\) and Theorem [2] holds with \(\Delta_n\) proportional to \(s^*\). Then there exists constants \(M\) and \(C'\) such that

\[
\mathbb{E}_{\beta} \{\Pi_n(B'_{M\delta_n})\} \to 0 \quad \text{as} \quad n \to \infty,
\]

uniformly in \(\beta^*\) with \(|S_{\beta^*}| = s^*\), where

\[
\delta_n = \frac{s^* \log(p/s^*)}{\kappa(C's^*)^2\|X\|^2_2}.
\]

**Proof.** It follows immediately from (13) that \(\|\beta - \beta^*\|^2_2 > M\delta_n\) implies

\[
\|X(\beta - \beta^*)\|^2_2 > M\kappa(|S_{\beta^*}|^2\|X\|^2_2\delta_n).
\]

By definition of \(\delta_n\) and the inequality (14), this last inequality implies

\[
\|X(\beta - \beta^*)\|^2_2 > Ms^* \log(p/s^*)\]

If we take \(M\) as in Theorem [2] then the event in the above display is exactly \(B_{M\varepsilon_n}\). We have shown that \(\Pi_n(B'_{M\delta_n}) \leq \Pi_n(B_{M\varepsilon_n})\). By Theorem [2] the expectation of the upper bound vanishes uniformly in \(\beta^*\) 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 \(\kappa(C's^*)\) in place of \(\kappa(|S_{\beta^*}|^2\|X\|^2_2\delta_n)\) above. However, this extra term is \(o(1)\) by Theorem [2] and, therefore, does not actually affect the proof.

Compare this result to the third in Theorem 2 of [Castillo et al. 2014]. First, our result involves the minimax rate discussed in [Verzelen 2012] compared to the slightly weaker lasso oracle rate \(s^* \log p\) in place of \(s^* \log(p/s^*)\). Second, our rate does not depend on a “compatibility number” (e.g., Castillo et al. 2014, Definition 2.1), which makes interpretation of our result much easier.

### 3.6 Model selection

Under the high-dimensional linear model (1), an important question is if data can be used to identify the true model from which it was generated. In our case, the question of interest is if the posterior distribution for \(S\) will concentrate, in a specific sense to be explained, around the true model \(S^*\).

Let \(\mathcal{M}\) denote the set of models \(S\), with positive prior probability, that properly contains the true model \(S^*\), i.e., \(\mathcal{M} = \{S : S \supset S^*, S \neq S^*, |S| \leq n\}\). Also, let \(\mathcal{M}_s \subset \mathcal{M}\) be the collection of all models \(S\) in \(\mathcal{M}\) with \(|S| = s\), for \(s = s^* + 1, \ldots, n\). Then the numerator of \(\Pi_n(\mathcal{M})\), the posterior probability of \(\mathcal{M}\), is given by

\[
N_n(\mathcal{M}) = \sum_{S \in \mathcal{M}} \pi(S) \int_{\mathbb{R}^p} R_n(\beta, \beta^*)^a \Pi(d\beta).
\]

Then we have the following bound on the expectation of \(N_n(\mathcal{M})\).
Lemma 5. Under the conditions of Lemma 2, $E_{\beta^*}\{N_n(\mathcal{M})\} \leq \sum_{s=s^*+1}^{n} \varphi^s (p-s^*) (p)^{-1} f_n(s)$ for all $\beta^*$.

Proof. Using an argument similar to that in the proof of Lemma 3, we get that

$$E_{\beta^*}\{N_n(\mathcal{M})\} \leq \sum_{S \in \mathcal{M}} \varphi^{|S|} \pi(S) = \sum_{s=s^*+1}^{n} \varphi^s f_n(s) \frac{\pi_s}{(p)^s}.$$ 

The claim follows since $|\mathcal{M}_s| = (\frac{p-s^*}{s-s^*})$.

The first result we consider, relevant to model selection, is the following, which says that the posterior distribution of $S$ will, on average, assign vanishing mass to models that contain any unimportant variables. In other words, any model selection strategy based on our empirical Bayes posterior will have vanishing false discovery rate.

Theorem 4. Assume the conditions of Lemmas 1 and 2, and that $\log p \gg s^*$. If $f_n$ is the complexity prior (10) with $a > 1$, then $E_{\beta^*}\{\Pi^n(\mathcal{M})\} \to 0$.

Proof. By an argument similar to that following Lemma 2, we have

$$E_{\beta^*}\{\Pi^n(\mathcal{M})\} \leq \frac{E_{\beta^*}\{N_n(\mathcal{M})\}}{\pi(S^*)e^{-cs^*}} + o(1)$$

for suitable constant $c$. Inserting the bound in Lemma 5 and simplifying gives

$$E_{\beta^*}\{\Pi^n(\mathcal{M})\} \leq e^{-cs^*} \sum_{s=s^*+1}^{n} \varphi^s f_n(s) \frac{(p)}{(p^s)} \left(\frac{p-s^*}{s-s^*}\right) + o(1).$$

We have the two trivial bounds:

$$\frac{(p)}{(p^s)} < 1 \quad \text{for} \quad s^* < s \ll p, \quad \text{and} \quad \left(\frac{p-s^*}{s-s^*}\right) \leq p^{s-s^*}.$$

Plug these in, applying the formula (10) for $f_n$, and the summation becomes

$$e^{-cs^*} \sum_{s=s^*+1}^{n} \varphi^s p^{s-s^*} \left(\frac{1}{cp^a}\right)^{s-s^*} = (\varphi e^c)^{s^*} \sum_{s=s^*+1}^{n} \left(\frac{\varphi}{cp^a}\right)^{s-s^*}.$$ 

Let $r = \varphi/cp^{a-1}$. Since $a > 1$, for large $n$, the latter term in the above display is a partial sum of a convergent geometric series, so the upper bound is of the order $(\varphi e^c)^{s^*} \cdot r$. Since $\log p \gg s^*$, this upper bound vanishes, proving the claim.

In addition to the result in Theorem 4, one might ask if the posterior will concentrate on models that include all important variables. Our second result gives an affirmative answer to this question, provided that the non-zero $\beta_j^*$ are not too close to zero.

Theorem 5. Suppose the conditions of Theorems 1 and 3 hold. Let $M\delta_n$ be as defined in Theorem 3. Then $E_{\beta^*}\{\Pi^n(S \supset \{j : |\beta_j^*|^2 \geq M\delta_n\})\} \to 1$. 

13
Proof. If \( S \not\supset \{ j : |\beta_j^*| \geq M\delta_n \} \), then \( \| \beta - \beta^* \|_2^2 \geq M\delta_n \). The expected posterior probability of the latter event vanishes, according to Theorem 3 and, therefore, the same holds for the former. Consequently, \( E_{\beta^*} \{ \pi^n(S \supset \{ j : |\beta_j| \geq M\delta_n \}) \} \to 1. \]

From the previous two theorems, we get a model selection consistency result provided that all of the non-zero \( \beta_j^* \) are not too close to zero. Let

\[
\hat{B}_{M\delta_n} = \left\{ \beta : \min_{j \in S_0} |\beta_j| \geq M\delta_n \right\}.
\]

That is, if \( \beta \in \hat{B}_{M\delta_n} \), then all of its non-zero entries have squared magnitude at least \( M\delta_n \). Compare this to the “beta-min” conditions in, e.g., B"uhlmann and van de Geer (2011), and to the analogous result in Corollary 1 of Castillo et al. (2014).

**Corollary 1.** Assume the conditions of Theorems 4 and 5. Then for any \( \beta^* \) in \( \hat{B}_{M\delta_n} \), with \( |S_{\beta^*}| \leq n \), we have \( E_{\beta^*} \{ \pi^n(S = S_{\beta^*}) \} \to 1. \)

**Proof.** Theorem 4 says that, asymptotically, the posterior cannot give positive mass to models with unnecessary variables. Furthermore, by definition of \( \hat{B}_{M\delta_n} \), Theorem 5 says that, asymptotically, the posterior must give unit mass to models that contain \( S_{\beta^*} \). Putting these two conclusions together says that \( \pi^n(S = S_{\beta^*}) \) converges to 1.

---

**4 Numerical results**

**4.1 Implementation**

To compute our empirical Bayes posterior distribution, we employ a Markov chain Monte Carlo method. To start, recall that the joint posterior for \( (S, \beta_S) \) is proportional to

\[
\pi^n(S, \beta_S) \propto \pi(S) N(Y \mid X_S \beta_S, \sigma^2 I_n)^n N(\beta_S \mid \hat{\beta}_S, \gamma^{-1}(X_S^\top X_S)^{-1}).
\]

The effect of the fractional power on the likelihood is, up to proportionality, to replace the variance \( \sigma^2 \) with \( \sigma^2/\alpha \). So, the joint posterior for \( (S, \beta_S) \) is proportional to

\[
\pi(S) N(Y \mid X_S \beta_S, \sigma^2 \alpha^{-1} I_n) N(\beta_S \mid \hat{\beta}_S, \gamma^{-1}(X_S^\top X_S)^{-1}).
\]

Due to conjugacy, it is straightforward to integrate out \( \beta_S \) from the above expression, leaving the marginal posterior distribution for \( S \) with mass function \( \pi^n(S) \) given by

\[
\pi^n(S) \propto \pi(S) e^{-\frac{\alpha}{2\sigma^2} \| Y - \hat{Y}_S \|^2 (\gamma + \alpha/\sigma^2)^{-|S|/2}},
\]

where \( \hat{Y}_S = X_S \hat{\beta}_S \) 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 insight on Theorems 2 and 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 \( q(S' \mid S) \) is a proposal function, then a single iteration of our proposed Metropolis–Hastings sampler goes as follows:
1. Given a current state $S$, sample $S' \sim q(\cdot \mid S)$.

2. Move to the new state $S'$ with probability

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

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 of individual variables can then easily be computed based on this sample.

If samples of the corresponding $\beta_S$ are desired, then these can easily be obtained, via conjugacy, after a sample of $S$ is available. In particular, the posterior distribution for $\beta_S$, given $S$ and data, is a normal distribution with mean $\hat{\beta}_S$ and variance

$$(\gamma + \alpha/\sigma^2)^{-1}(X_S^\top X_S)^{-1}.$$

R code is available at [www.math.uic.edu/~rgmartin](http://www.math.uic.edu/~rgmartin). In our case, we use a symmetric proposal distribution $q(S' \mid S)$, i.e., 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.

To implement this procedure, the tuning parameters $\alpha$ and $\gamma$ need to be specified. Recall that $\alpha$ equal to 1 corresponds to the genuine Bayesian case. If we take $\alpha$ close to 1, then the theory holds if $\gamma$ is sufficiently close to 0. For a default implementation, we recommend to work with the limiting case $\alpha = 1$ and $\gamma = 0$, although our convergence theory only holds for $(\alpha, \gamma)$ settings arbitrarily close to this one.

Finally, in practice, the error variance $\sigma^2$ is seldom known, so some procedure to handle unknown $\sigma^2$ is needed. We proposed to modify our empirical Bayes posterior by plugging in an estimate of $\sigma^2$. A reasonable strategy is to use a sort of residual mean square error based on a lasso fit with tuning parameter selected via cross validation (Reid et al. 2014). This model selected by lasso driven by cross validation is also our choice for initializing our Metropolis–Hastings algorithm described above.

### 4.2 Simulations

In this section, we reconsider some of the simulation experiments performed by Narisetty and He (2014), which are related to experiments presented in Johnson and Rossell (2012). 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 $S^*$ has $s^* = 5$. The particular correlation structure among the covariates is given practical justification in Johnson and Rossell (2012).

Under this setup, we consider three different settings:

**Setting 1.** $n = 100$, $p = 500$, and $\beta_{S^*} = (0.6, 1.2, 1.8, 2.4, 3.0)^\top$;

**Setting 2.** $n = 200$, $p = 1000$, and $\beta_{S^*}$ same as in Setting 1.

**Setting 3.** $n = 100$, $p = 500$, and $\beta_{S^*} = (0.6, 0.6, 0.6, 0.6, 0.6)^\top$. 

15
Our Settings 1–2 correspond to the two \((n, p)\) configurations in Case 2 of Narisetty and He (2014) and our Setting 3 is the same as their Case 3.

For our empirical Bayes method, we employ the complexity prior (10) with \(c = 1\) and \(a = 0.05\), i.e., \(f_n(s) \propto p^{-0.05s}\). The choice of small \(a\) is to allow the prior to be sufficiently spread out to allow the posterior to move across the model space and, in particular, for the Markov chain for \(S\) to mix reasonably well. 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 (2004).

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 \(S^*\), respectively, i.e.,

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

We expect the former to be close to 1 and the latter 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 \(S^*\). 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 (2014), 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 (2012), denoted by BCR.Joint, and the spike-and-slab method of Ishwaran and Rao (2005a,b), denoted by SpikeSlab. We also consider three penalized likelihood methods, all tuned with BIC, namely, the lasso (Tibshirani 1996), the elastic net (Zou and Hastie 2005), and the smoothly clipped absolute deviation (Fan and Li 2001), 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 (2014), which were obtained based on 200 samples taken from the models described in Settings 1–3 described above.

In all three settings, our selection method based on our empirical Bayes posterior is 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 (2014), the next-best-performer in the simulation study. Their method produces a posterior distribution for \(\beta\) 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 \(\beta\) 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 demonstrate that, for certain priors, the posterior concentrates...
| Method   | $\tilde{p}_0$ | $\tilde{p}_1$ | $P(S = S^*)$ | $P(S \supseteq S^*)$ | FDR  |
|----------|---------------|---------------|---------------|------------------------|------|
| BASAD    | 0.001         | 0.948         | 0.730         | 0.775                  | 0.011|
| BASAD.BIC| 0.000         | 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         | 0.966         | 0.745         | 0.835                  | 0.049|

Table 1: Simulation results for Setting 1. First seven rows taken from Table 2 (top) in Narisetty and He (2014); the $EB$ row corresponds to our empirical Bayes procedure.

| Method   | $\tilde{p}_0$ | $\tilde{p}_1$ | $P(S = S^*)$ | $P(S \supseteq S^*)$ | 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         | 0.999         | 0.950         | 0.995                  | 0.011|

Table 2: Simulation results for Setting 2. First seven rows taken from Table 2 (bottom) in Narisetty and He (2014); the $EB$ row corresponds to our empirical Bayes procedure.

| Method   | $\tilde{p}_0$ | $\tilde{p}_1$ | $P(S = S^*)$ | $P(S \supseteq S^*)$ | 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         |                        |      |
| $EB$     | 0.004         | 0.786         | 0.230         | 0.291                  | 0.101|

Table 3: Simulation results for Setting 3. First seven rows taken from Table 3 in Narisetty and He (2014); the $EB$ row corresponds to our empirical Bayes procedure.
around the sparse truth and the optimal minimax rate. To our knowledge, this is the first proof of minimax rate concentration 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.

Our key assumption throughout is that the true model is sufficiently sparse, in particular, \( \| \beta^* \|_0 = |\{ j : \beta^*_j \neq 0 \}| \leq n \). We argue that this is a very reasonable assumption. If we do not assume \( \| \beta^* \|_0 \leq n \) then a proposal to extend the method presented in the paper would be to use an alternative estimator which works for \(|S| > n\), such as the lasso. Extensions will not be trivial since the distributional properties of such estimators are quite different to those of the least squares estimator \( \hat{\beta}_S \) with \(|S| < n\).

The general strategy proposed here, and also in Martin and Walker (2013), goes as follows. Suppose we have a high-dimensional parameter, and different models \( S \) identify a subset of “non-null” parameters \( \theta_S \). Suppose further that \( \theta \) 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 \((S, \theta_S)\) 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 \( \theta_S \). One possible application of this approach, which we plan to explore, is a mixture model where \( S \) represents the number of mixture components, and \( \theta_S \) is the set of parameters associated with a mixture model with \( S \) mixture components.

A Proofs

A.1 Proof of Lemma 1

Write \( \| \cdot \| \) for the \( \ell_2 \)-norm \( \| \cdot \|_2 \). The denominator \( D_n \) in (6) involves an average over all suitable models \( S \) with respect to \( \pi(S) \). This average is greater than the quantity for \( S = S^* \) times \( \pi(S^*) \). That is, \( D_n \) is bigger than

\[
\pi(S^*) \int e^{-\frac{n}{2\sigma^2} \{ \| Y - X_{S^*} \beta_{S^*} \|^2 - \| Y - X_{S^*} \beta_{S^*}^\star \|^2 \}} \mathcal{N}(\beta_{S^*} | \hat{\beta}_{S^*}, \gamma^{-1}(X_{S^*}^\top X_{S^*})^{-1}) \, d\beta_{S^*}.
\]

Direct calculation shows that the lower bound equals

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

According to the least-squares distribution theory, the quantity in the exponent is proportional to a \( \text{ChiSq}(|S^*|) \) random variable so, with probability 1 for all sufficiently large \( n \), the exponential term equals \( e^{\frac{n}{2\sigma^2} |S^*|} \). Then the lower bound can be written as

\[
\pi(S^*) e^{-\left( \frac{1}{2} \log(1 + \frac{\alpha}{\gamma \sigma^2}) - \frac{1}{2} \frac{\alpha}{\sigma^2} \right) |S^*|}.
\]

Under the stated conditions, the bracketed term in the exponent is positive. Set \( c \) to be this positive quantity to complete the proof.

18
A.2 Proof of Lemma 2

Write $B_n = B_{ε_n}$. Rewrite the numerator $N_n$ of the posterior (6) as

$$N_n = \int_{B_n} \sum_S \pi(S) \left\{ \frac{N(Y \mid X_{S+}, \sigma^2 I)}{N(Y \mid X_{β^*}, \sigma^2 I)} \right\} ^α N(β_S \mid \hat{β}_S, γ^{-1}(X_S^TX_S)^{-1}) dβ_S$$

$$= \sum_S \pi(S) \int_{B_n(S)} \left\{ \frac{N(Y \mid X_{β^*+}, \sigma^2 I)}{N(Y \mid X_{β^*}, \sigma^2 I)} \right\} ^α N(β_S \mid \hat{β}_S, γ^{-1}(X_S^TX_S)^{-1}) dβ_S,$$

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

$$\int_{B_n(S)} E \left[ \left\{ \frac{N(Y \mid X_{β^*+}, \sigma^2 I)}{N(Y \mid X_{β^*}, \sigma^2 I)} \right\} ^α N(β_S \mid \hat{β}_S, γ^{-1}(X_S^TX_S)^{-1}) \right] dβ_S.$$

Apply Hölder’s inequality to the inside expectation, i.e., for $h > 1$ and $q = (h - 1)/h$,

$$E \left[ \left\{ \frac{N(Y \mid X_{β^*+}, \sigma^2 I)}{N(Y \mid X_{β^*}, \sigma^2 I)} \right\} ^α N(β_S \mid \hat{β}_S, γ^{-1}(X_S^TX_S)^{-1}) \right] \leq E^{1/h} \left[ \left\{ \frac{N(Y \mid X_{β^*+}, \sigma^2 I)}{N(Y \mid X_{β^*}, \sigma^2 I)} \right\} ^{oha} \right] E^{1/q} \left[ N^q(β_S \mid \hat{β}_S, γ^{-1}(X_S^TX_S)^{-1}) \right]. \quad (15)$$

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

$$E^{1/h} \left[ \left\{ \frac{N(Y \mid X_{β^*+}, \sigma^2 I)}{N(Y \mid X_{β^*}, \sigma^2 I)} \right\} ^{oha} \right] = e^{- \frac{a(1-ha)}{2\sigma^2} \|X(β_{S+} - β^*)\|^2}. \quad (16)$$

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

$$X_Sβ_S - X_S\hat{β}_S = X_S(X_S^TX_S)^{-1}X_S^T(X_Sβ_S - Y),$$

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

$$Z := \|X_Sβ_S - X_S\hat{β}_S\|^2 = \|X_S(X_S^TX_S)^{-1}X_S^T(X_Sβ_S - Y)\|^2$$

is distributed as a non-central chi-square with $|S|$ degrees of freedom and non-centrality parameter $λ = \|X_S(β_S - (X_S^TX_S)^{-1}X_S^TX_β^*)\|^2$. Then

$$E^{1/q} \left[ N^q(β_S \mid \hat{β}_S, γ^{-1}(X_S^TX_S)^{-1}) \right] = \gamma^{\frac{|S|}{2}} \frac{X_S^TX_S^{1/2}}{(2π)^{|S|/2}} E^{1/q} (e^{-\frac{Z}{2}})$$

$$= \gamma^{\frac{|S|}{2}} \frac{X_S^TX_S^{1/2}}{(2π)^{|S|/2}} (1 + qγ) \frac{|S|^q}{2^q} e^{-\frac{Z}{2(1+qγ)}}$$

$$= \gamma^{\frac{|S|}{2}} \frac{X_S^TX_S^{1/2}}{(2π)^{|S|/2}} (1 + qγ) \frac{|S|^q}{2^q} e^{-\frac{Z}{2(1+qγ)}} \|X_S(β_S - (X_S^TX_S)^{-1}X_S^TX_β^*)\|^2, \quad (17)$$

19
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 (15) over \( A_n(S) \) with respect to \( \beta_S \). It is clear from the definition of \( B_n(S) \) that the quantity in (16) is bounded on \( B_n(S) \), i.e.,

\[
e^{-\frac{\alpha(1-h\alpha)}{2\sigma^2}\|X(\beta_S-\beta^*)\|^2} \leq e^{-\frac{\alpha(1-h\alpha)}{2\sigma^2}\epsilon_n}, \quad \beta_S \in B_n(S).
\]

It is also clear that the expression (17) resembles a normal density in \( \beta_S \), modulo some multiplicative factors. The algebra is tedious, but the integral of (17) with respect to \( \beta_S \) is bounded above by

\[
\varphi^{|S|} \quad \text{where} \quad \varphi = \varphi(\gamma, q) = [(1 + \gamma q)^{1-\frac{1}{q}}]^\frac{1}{2}.
\]

Putting everything together, we have that

\[
E(N_n) \leq e^{-\frac{\alpha(1-h\alpha)}{2\sigma^2}\epsilon_n} \sum_S \varphi^{|S|} \pi(S).
\]

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

### A.3 Proof of Lemma 3

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

\[
\sum_{S: \Delta \leq |S| \leq n} \pi(S) \int \{ N(Y | X\hat{\beta}_S^+, \sigma^2 I) \}^\alpha N(\beta_S | \hat{\beta}_S, \gamma^{-1}(X_S^\top X_S)^{-1}) \ d\beta_S,
\]

Take expectation with respect to \( Y \sim N(X\beta^*, \sigma^2 I) \) 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 \( \beta_S \). Then the integral over \( \beta_S \) is bounded by \( \varphi^{|S|} \) and the claim follows.

### References

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

Bondell, H. D. and Reich, B. J. (2012). Consistent high-dimensional Bayesian variable selection via penalized credible regions. *J. Amer. Statist. Assoc.*, 107(500):1610–1624.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data*. Springer Series in Statistics. Springer, Heidelberg.

Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when \( p \) is much larger than \( n \). *Ann. Statist.*, 35(6):2313–2351.
Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. (2014). Bayesian linear regression with sparse priors. Unpublished manuscript, arXiv:1403.0735.

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.

Chen, J. and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771.

Clyde, M. and George, E. I. (2004). Model uncertainty. *Statist. Sci.*, 19(1):81–94.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96(456):1348–1360.

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statist. Sinica*, 20(1):101–148.

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

Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531.

Giraud, C., Huet, S., and Verzelen, N. (2012). High-dimensional regression with unknown variance. *Statist. Sci.*, 27(4):500–518.

Heaton, M. J. and Scott, J. G. (2010). Bayesian computation and the linear model. In Cheh, M.-H., Dey, D., Müller, P., Sun, D., and Ye, K., editors, *Frontiers of Statistical Decision Making and Bayesian Analysis*, pages 527–545. Springer.

Ishwaran, H. and Rao, J. S. (2005a). Spike and slab gene selection for multigroup microarray data. *J. Amer. Statist. Assoc.*, 100(471):764–780.

Ishwaran, H. and Rao, J. S. (2005b). Spike and slab variable selection: frequentist and Bayesian strategies. *Ann. Statist.*, 33(2):730–773.

James, G. M. and Radchenko, P. (2009). A generalized Dantzig selector with shrinkage tuning. *Biometrika*, 96(2):323–337.

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(1):127–142.

Johnson, V. E. and Rossell, D. (2012). Bayesian model selection in high-dimensional settings. *J. Amer. Statist. Assoc.*, 107(498):649–660.

Martin, R. and Walker, S. G. (2013). Asymptotically minimax empirical Bayes estimation of a sparse normal mean. Unpublished manuscript, arXiv:1304.7366.

Narisetty, N. N. and He, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *Ann. Statist.*, 42(2):789–817.
Park, T. and Casella, G. (2008). The Bayesian lasso. *J. Amer. Statist. Assoc.*, 103(482):681–686.

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

Seber, G. A. F. and Lee, A. J. (2003). *Linear Regression Analysis*. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition.

Shen, X. and Wasserman, L. (2001). Rates of convergence of posterior distributions. *Ann. Statist.*, 29(3):687–714.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288.

Verzelen, N. (2012). Minimax risks for sparse regressions: ultra-high dimensional phenomena. *Electron. J. Stat.*, 6:38–90.

Walker, S. and Hjort, N. L. (2001). On Bayesian consistency. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 63(4):811–821.

Walker, S. G., Lijoi, A., and Prünster, I. (2007). On rates of convergence for posterior distributions in infinite-dimensional models. *Ann. Statist.*, 35(2):738–746.

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

Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist.*, 36(4):1567–1594.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(2):301–320.