The Annals of Statistics
2015, Vol. 43, No. 5, 1986–2018
DOI: 10.1214/15-AOS1334

BAYESIAN LINEAR REGRESSION WITH SPARSE PRIORS

BY ISMAËL CASTILLO[1,*], JOHANNES SCHMIDT-HIEBER[2,†]
AND AAD VAN DER VAART[2,†]

CNRS Paris and Leiden University*[†]

We study full Bayesian procedures for high-dimensional linear regression under sparsity constraints. The prior is a mixture of point masses at zero and continuous distributions. Under compatibility conditions on the design matrix, the posterior distribution is shown to contract at the optimal rate for recovery of the unknown sparse vector, and to give optimal prediction of the response vector. It is also shown to select the correct sparse model, or at least the coefficients that are significantly different from zero. The asymptotic shape of the posterior distribution is characterized and employed to the construction and study of credible sets for uncertainty quantification.

1. Introduction. Consider estimation of a parameter $\beta \in \mathbb{R}^p$ in the linear regression model

$$Y = X\beta + \varepsilon, \tag{1.1}$$

where X is a given, deterministic $(n \times p)$ matrix, and ε is an n-variate standard normal vector. The model is standard, but we are interested in the *sparse* setup, where $n \le p$, and possibly $n \ll p$, and "many" or "most" of the coefficients β_i of the parameter vector are zero, or close to zero. We study a Bayesian approach based on priors that set a selection of coefficients β_i a priori to zero; equivalently, priors that distribute their mass over models that use only a (small) selection of the columns of X. Bayes's formula gives a posterior distribution as usual. We study this under the "frequentist" assumption that the data Y has in reality been generated according to a given (sparse) parameter β^0. The expectation under the previous distribution is denoted $\mathbb{E}_{\beta^0}$.

Specifically, we consider a prior Π on β that first selects a *dimension* s from a prior π_p on the set $\{0, \dots, p\}$, next a random subset $S \subset \{1, 2, \dots, p\}$ of cardinality $|S| = s$ and finally a set of nonzero values $\beta_S := \{\beta_i : i \in S\}$ from a prior density g_S on $\mathbb{R}^S$. Formally, the prior on (S, β) can be expressed as

$$(S, \beta) \mapsto \pi_p(|S|) \frac{1}{\binom{p}{|S|}} g_S(\beta_S) \delta_0(\beta_{S^c}), \tag{1.2}$$

Received March 2014; revised March 2015.

[1] Supported in part by ANR Grants "Banhdits" ANR-2010-BLAN-0113-03 and "Calibration" ANR-2011-BS01-010-01.

[2] Supported by the European Research Council under ERC Grant Agreement 320637.

MSC2010 subject classifications. Primary 62G20; secondary 62G05.

Key words and phrases. Bayesian inference, sparsity.

where the term $\delta_0(\beta_{S^c})$ refers to the coordinates $\beta_{S^c} := (\beta_i : i \in S^c)$ being zero. We focus on the situation where g_S is a product $\otimes g$ of densities over the coordinates in S, for g a fixed continuous density on $\mathbb{R}$, with the Laplace density as an important special case. The prior π_p is crucial for expressing the "sparsity" of the parameter. One of the main findings of this paper is that weights $\pi_p(s)$ that decrease slightly faster than exponential in the dimension s give good performance.

Priors of the type of (1.2) were considered by many authors, including [8, 20, 21, 26, 34, 41, 48]. Other related contributions include [2, 3, 33]. The paper [16] contains a theoretical analysis similar to the present paper, but restricted to the special case that the regression matrix X is the identity and $p = n$; see Example 1. The general model (1.1) shares some features with this special case, but is different in that it must take account of the noninvertibility of X and its interplay with the sparsity assumption, especially for the case of recovering the parameter β, as opposed to estimating the mean $X\beta$. While the proofs in [16] use a factorization of the model along the coordinate axes, exponential tests and entropy bounds, in the present paper we employ a direct and refined analysis of the posterior ratio (1.4), exploiting the specific form of the prior Laplace density g. Furthermore, even for the case that X is the identity matrix, the present paper provides several new results of interest: distributional approximations to the posterior distribution, insight in the scaling of the prior on the nonzero coordinates and oracle formulations of the contraction rates.

Algorithms for the computation of the posterior distribution corresponding to (1.2), especially for the "spike and slab" prior described in Example 5 below, are routine for small dimensions p and n (e.g., [18, 21, 25–27, 34, 41, 48]). For large dimensions the resulting computations are intensive, due to the large number of possible submodels S. Many authors are currently developing algorithms that can cope with larger numbers of covariates, in the sparse setup considered in the present paper. In Section 5 we review recent progress on various methods, of which some are feasible for values of p up to hundreds or thousands [7, 8, 16, 31, 35, 38–40, 42, 43, 47]. Although this upper bound will increase in the coming years, clearly it falls far short of the dimensions attainable by (point) estimation methods based on convex programming, such as the LASSO. Other Bayesian approaches to sparse regression that do not explicitly include model selection (e.g., [14, 23, 24]) can cope with somewhat higher dimensions, but truly high-dimensional models are out of reach of fully Bayesian methods at the present time.

Not surprisingly to overcome the nonidentifiability of the full parameter vector β in the overspecified model (1.1), we borrow from the work on sparse regression within the non-Bayesian framework; see [1, 6, 9, 10, 13, 19, 29, 36, 49, 50]. Good performance of the posterior distribution is shown under *compatibility* and *smallest sparse eigenvalue* conditions; see Section 2. Although the constants in these results are not as sharp as results for the LASSO, the posterior contraction rates obtained are broadly comparable to convergence rates of the LASSO.

The LASSO and its variants are important frequentist methods for sparse signal recovery. As the LASSO is a posterior mode (for an i.i.d. Laplace prior on the β_i), it may seem to give an immediate link between Bayesian and non-Bayesian methods. However, we show in Section 3 that the LASSO is essentially non-Bayesian, in the sense that the corresponding *full* posterior distribution is a useless object.

In contrast, the posterior distribution resulting from the prior (1.2) gives both reasonable reconstruction of the parameter β and a quantification of uncertainty through the spread in the posterior distribution. We infer this from combining results on the contraction rate of the full posterior distribution with distributional approximations. The latter show that the posterior distribution behaves asymptotically as a mixture of Bernstein–von Mises type approximations to submodels, where the location of the mixture components depends on the setting. The latter approximations are new, also for the special case that X is the identity matrix.

It is crucial for these results that the prior (1.2) models sparsity through the *model selection* prior π_p, and separates this from modeling the nonzero coordinates through the prior densities g_S. For instance, in the case that g_S is a product of Laplace densities, this allows the scale parameter to be constant or even to tend to zero, thus making this prior uninformative. This is in stark contrast to the choice of the smoothing parameter in the (Bayesian) LASSO, which must tend to infinity in order to shrink parameters to zero, where it cannot differentiate between truly small and nonzero parameters. Technically this has the consequence that the essential part of the proofs is to show that the posterior distribution concentrates on sets of small dimension. This sets it apart from the frequentist literature on sparse regression, although, as mentioned, many essential ideas reappear here in a Bayesian framework.

The paper is organized as follows. In Section 2 we present the main results of the paper. We specialize to Laplace priors on the nonzero coefficients and investigate the ability of the posterior distribution to recover the parameter vector β, the predictive vector $X\beta$ and the set of nonzero coordinates. Furthermore, we derive a distributional approximation to the posterior distribution, and apply this to construct and study credible sets. In Section 3 we present the negative result on the Bayesian interpretation of the LASSO. Next in Section 4 we show that for recovery of only the predictive vector $X\beta$, significantly milder conditions than in Section 2 suffice. Proofs are deferred to Section 6 and the supplementary material [15].

1.1. *Notation.* For a vector $\beta \in \mathbb{R}^p$ and a set $S \subset \{1, 2, \dots, p\}$ of indices, β_S is the vector $(\beta_i)_{i \in S} \in \mathbb{R}^S$, and $|S|$ is the cardinality of S. The *support* of the parameter β is the set $S_\beta = \{i : \beta_i \neq 0\}$. The support of the true parameter β^0 is denoted S_0, with cardinality $s_0 := |S_0|$. Similarly, for a generic vector β^*, we write $S_* = S_{\beta^*}$ and $s_* = |S_*|$. We write $s = |S|$ if there is no ambiguity to which set S is referred to. For $1 \leq q < \infty$ and $\beta \in \mathbb{R}^p$, let $\|\beta\|_q := (\sum_{i=1}^p |\beta_i|^q)^{1/q}$.

We let $X_{\cdot,i}$ be the ith column of X, and

$$\|X\| = \max_{i=1,\dots,p} \|X_{\cdot,i}\|_2 = \max_{i=1,\dots,p} \left(X^t X\right)_{i,i}^{1/2}. \tag{1.3}$$

For the prior Π defined above, Bayes's formula gives the following expression for the posterior distribution $\Pi[\cdot|Y]$. For any Borel set B of $\mathbb{R}^p$,

$$\Pi[B|Y] = \int_B e^{-\|Y-X\beta\|_2^2/2}\,d\Pi(\beta) \Big/ \int e^{-\|Y-X\beta\|_2^2/2}\,d\Pi(\beta). \tag{1.4}$$

2. Main results.

2.1. *Prior.* In this section we consider the prior (1.2), with g_S the product of $|S|$ Laplace densities $\beta \mapsto 2^{-1}\lambda \exp(-\lambda|\beta|)$. We allow the (inverse) scale parameter λ to change with p, within the range, with $\|X\|$ defined in (1.3),

$$\frac{\|X\|}{p} \le \lambda \le 2\overline{\lambda}, \qquad \overline{\lambda} = 2\|X\|\sqrt{\log p}. \tag{2.1}$$

The quantity $\overline{\lambda}$ in the upper bound is the usual value of the regularization parameter λ of the LASSO [as in (3.1) below]. Its large value causes the LASSO to shrink many coordinates β_i to zero, as is desired in the sparse situation. However, in our Bayesian setup, sparsity should be induced by model selection, through the prior π_p on the model dimension, and the Laplace prior densities model only the nonzero coordinates. Large values of λ would shrink the nonzero coordinates to zero, which is clearly undesirable and unnatural. Thus it is natural to assume $\lambda \ll \overline{\lambda}$, and fixed values of λ, and even values decreasing to zero, may well be natural, depending on the regression setting. We shall see that small values of λ permit a distributional approximation to the posterior distribution centered at unbiased estimators. The results below hold for all λ in the range (2.1), but they are meant to be read for a specific sequence of λ and are not suitable for optimization over λ.

The precise interpretation of the size of λ is confounded with the regression setting, the error variance (which we have set to unity for simplicity of notation) and the scaling of the regression matrix. The following three special cases shed some light on this.

EXAMPLE 1 (Sequence model). In the *sequence model* considered in [28] and [16], the observation is a vector $(Y_1, \dots, Y_n)$ of independent coordinates $Y_i \sim N(\beta_i, 1)$. This corresponds to $X = I$ and $n = p$ in the present setting (1.1), whence $\|X\| = 1$. Condition (2.1) then reduces to $p^{-1} \le \lambda \le 4\sqrt{\log p}$. Fixed values of λ, as considered in [16], are easily included. As there is only one observation per parameter, it may not be unreasonable to consider $\lambda \to 0$, in order to create noninformative priors for the nonzero coefficients. This is allowed easily also.

EXAMPLE 2 (Sequence model, multiple observations). In an extension of the sequence model of the preceding example, the n observations are from normal distributions $N(\beta_i, \sigma_n^2)$ with variances $\sigma_n^2 \to 0$. By defining the Y_i as σ_n^{-1}

times the original observations, we can fit this into model (1.1), which has unit error variances. If we keep the original definition of the β_i, then the regression matrix is $X = \sigma_n^{-1} I$, and hence $\|X\| = \sigma_n^{-1}$. Condition (2.1) then reduces to $\sigma_n^{-1}/n \le \lambda \le 4\sigma_n^{-1}\sqrt{\log n}$. Fixed values of λ are included if $n\sigma_n \gtrsim 1$, and values tending to zero if $n\sigma_n \to \infty$. By sufficiency of the sample mean in the normal location problem this corresponds to a sufficient number of replicate measurements on every parameter β_i in the original problem.

EXAMPLE 3 (Response model). If every row of the regression equation $Y = X\beta + \varepsilon$ refers to a measurement of an instance of a fixed relationship between an input vector $X_{i,\cdot} \in \mathbb{R}^p$ and the corresponding output Y_i, then the entry $X_{i,j}$ of X is the value of individual i on the jth covariable. It is then reasonable to think of these entries as being sampled from some fixed distribution, independent of n and p, in which case $\|X\|$ will (typically) be of the order $\sqrt{n}$. A fundamental example is the case where the entries of X are independent standard Gaussian $\mathcal{N}(0, 1)$. Condition (2.1) then reduces to $\sqrt{n}/p \le \lambda \le 4\sqrt{n}\sqrt{\log p}$. Fixed values of λ, as considered in [16], are included, provided $p \gtrsim \sqrt{n}$.

Although condition (2.1) does not exclude shrinkage through large values of λ, as for the LASSO, the most interesting situation is that sparsity is induced through model selection. The prior π_p on model dimension is crucial; it must downweight big models, but at the same time give sufficient mass to the true model. Exponential decrease turns out to work.

ASSUMPTION 1 (Prior dimension). There are constants $A_1, A_2, A_3, A_4 > 0$ with

$$A_1 p^{-A_3}\pi_p(s-1) \le \pi_p(s) \le A_2 p^{-A_4}\pi_p(s-1), \qquad s = 1, \dots, p. \tag{2.2}$$

EXAMPLE 4. Assumption (2.2) is met by the priors of the form, for constants $a, c > 0$,

$$\pi_p(s) \propto c^{-s} p^{-as}, \qquad s = 0, 1, \dots, p. \tag{2.3}$$

We refer to these priors as *complexity priors*, as their rate of decrease reflects the number of models $\binom{p}{s}$ of given size s for $s \ll p$; cf. [16].

EXAMPLE 5 (Slab and spike). Modeling the coordinates $\beta_1, \dots, \beta_p$ as i.i.d. variables from a mixture $(1-r)\delta_0 + rG$, of a Dirac measure δ_0 at zero and a Laplace distribution G, is included in (1.2) with π_p the binomial distribution with parameter p and r. The size r of the point mass at zero controls the model selection. The overall prior obtained by choosing r from a Beta $(1, p^u)$ hyper prior with $u > 1$ satisfies (2.3); cf. Example 2.2 in [16]. This prior is universal in that it is free of unknown smoothing parameters.

To conclude the discussion on the prior, we briefly comment on the case that the noise vector has unknown variance $\sigma^2 > 0$; that is, we observe $Y = X\beta + \sigma\varepsilon$. In this case one may use an empirical Bayesian approach, which replaces the unknown parameter by an estimator, or a hierarchical Bayesian approach, which puts a prior on σ^2, a common choice being an inverse Gamma prior. Since $Y/\sigma = X(\beta/\sigma) + \varepsilon$, it is natural to apply the prior, as in this paper, to the parameter β/σ. Thus given σ^2 and a model S, we choose the prior density on the nonzero values $\beta_S = \{\beta_i : i \in S\}$ as the product of $|S|$ Laplace densities $\beta \mapsto \lambda/(2\sigma)\exp(-\lambda|\beta|/\sigma)$, conditional on the estimated or prior value of σ.

2.2. *Design matrix.* The parameter β in model (1.1) is not estimable without conditions on the regression matrix. For the interesting case $p > n$, it is even necessarily unidentifiable. If β is known to be sparse, then "local invertibility" of the Gram matrix X^tX is sufficient for estimability, even in the case $p > n$. This is made precise in the following definitions, which are based on the literature, but with simplified notation suited to our Bayesian setup. For accessibility we include short discussions on the relations between the various concepts.

DEFINITION 2.1 (Compatibility). The compatibility number of model $S \subset \{1, \dots, p\}$ is given by

$$\phi(S) := \inf\left\{\frac{\|X\beta\|_2|S|^{1/2}}{\|X\|\|\beta_S\|_1} : \|\beta_{S^c}\|_1 \le 7\|\beta_S\|_1, \beta_S \neq 0\right\}.$$

The compatibility number (which is $\phi_{\mathrm{comp}}(7, S)/\|X\|$ in the notation of [10], page 157) compares the ℓ^2-norm of the predictive vector $X\beta$ to the ℓ^1-norm of the parameter β_S. A model S is considered "compatible" if $\phi(S) > 0$. It then satisfies the nontrivial inequality $\|X\beta\|_2|S|^{1/2} \ge \phi(S)\|X\|\|\beta_S\|_1$. We shall see that true vectors β^0 with compatible support S_{β^0} can be recovered from the data, uniformly in a lower bound on the size of their compatibility numbers.

The number 7 has no particular interest, but for simplicity we use a numerical value instead of an unspecified constant. Since the vectors β in the infimum satisfy $\|\beta_S\|_1 \le \|\beta\|_1 \le 8\|\beta_S\|_1$, it would not be a great loss of generality to replace β_S in the denominator of the quotient by β. However, the factor $|S|^{1/2}$ in the numerator may be seen as resulting from the comparison of the ℓ^1- and ℓ^2-norms of β_S through the Cauchy–Schwarz inequality: $\|\beta_S\|_1 \le |S|^{1/2}\|\beta_S\|_2$. Replacing $\|\beta_S\|_1/|S|^{1/2}$ by $\|\beta_S\|_2$ would make the compatibility number smaller, and hence give a more restrictive condition.

The compatibility number involves the full vectors β (also their coordinates outside of S) and allows to reduce the recovery problem to sparse vectors. The next two definitions concern sparse vectors only, but unlike the compatibility number, they are uniform in vectors up to a given dimension. In the notation of [10] (pages 156–157) the numbers in the definitions are the minima over $|S| \le s$ of the numbers $\Lambda_{\min,1}(\Sigma_{1,1}(S))/\|X\|$ and $\Lambda_{\min}(\Sigma_{1,1}(S))/\|X\|$, respectively.

DEFINITION 2.2 (Uniform compatibility in sparse vectors). The compatibility number in vectors of dimension s is defined as

$$\overline{\phi}(s) := \inf\left\{\frac{\|X\beta\|_2 |S_\beta|^{1/2}}{\|X\| \|\beta\|_1} : 0 \neq |S_\beta| \leq s\right\}.$$

DEFINITION 2.3 (Smallest scaled sparse singular value). The smallest scaled singular value of dimension s is defined as

$$\widetilde{\phi}(s) := \inf\left\{\frac{\|X\beta\|_2}{\|X\| \|\beta\|_2} : 0 \neq |S_\beta| \leq s\right\}. \tag{2.4}$$

For recovery we shall impose that these numbers for s equal to (a multiple of) the dimension of the true parameter vector are bounded away from zero. Since $\|\beta\|_1 \leq |S_\beta|^{1/2}\|\beta\|_2$ by the Cauchy–Schwarz inequality, it follows that $\widetilde{\phi}(s) \leq \overline{\phi}(s)$, for any $s > 0$. The stronger assumptions on the design matrix imposed through $\widetilde{\phi}(s)$ will be used for recovery with respect to the ℓ^2-norm, whereas the numbers $\overline{\phi}(s)$ suffice for ℓ^1-reconstruction. In Definition 2.3, "scaled" refers to the scaling of the matrix X by division by the maximum column length $\|X\|$; if the latter is unity, then $\widetilde{\phi}(s)$ is just the smallest scaled singular value of a submatrix of X of dimension s.

The final and strongest invertibility condition is in terms of "mutual coherence" of the regression matrix, which is the maximum correlation between its columns.

DEFINITION 2.4 (Mutual coherence). The *mutual coherence number* is

$$\mathrm{mc}(X) = \max_{1 \leq i \neq j \leq p} \frac{|\langle X_{\cdot,i}, X_{\cdot,j}\rangle|}{\|X_{\cdot,i}\|_2 \|X_{\cdot,j}\|_2}.$$

The "(K, s) *mutual coherence condition*" is that this number is bounded above by $(Ks)^{-1}$, in which case reconstruction is typically possible for true vectors β of dimension up to s. As correlations are easy to interpret, conditions of this type, which go back to [19], have been used by many authors. (Notably, Bunea, Tsybakov and Wegkamp [11] show that for reconstructions using the ℓ^1- and ℓ^2-norms, taking the maximum over all correlations can be relaxed to a maximum over pairs that involve at least one "active" coordinate.) The following lemma shows that they are typically stronger than conditions in terms of compatibility numbers or sparse singular values. The lemma is embodied in Lemma 2 in [32], and is closely related to the inequalities obtained in [46]. For ease of reference we provide a proof in the supplementary material [15].

LEMMA 1. $\phi(S)^2 \geq \overline{\phi}(1)^2 - 15|S|\,\mathrm{mc}(X)$; $\overline{\phi}(s)^2 \geq \widetilde{\phi}(s)^2 \geq \overline{\phi}(1)^2 - s\,\mathrm{mc}(X)$.

By evaluating the infimum in Definition 2.2 with β equal to unit vectors, we see that $\widetilde{\phi}(1) = \overline{\phi}(1) = \min_i \|X_{\cdot,i}\|_2/\|X\|$, which will typically be bounded away from zero. Thus the lemma implies that compatibility numbers and sparse singular values are certainly bounded away from zero for models up to size a multiple of $1/\operatorname{mc}(X)$, that is, models of size satisfying the "mutual coherence condition." This makes the mutual coherence the strongest of the three "sparse invertibility" indices introduced previously. We note that the reverse inequalities do not hold in general, and indeed the compatibility constant can easily be bounded away from zero, even if the mutual coherence number is much larger than $1/s$.

For many other possible indices (including "restricted isometry" and "irrepresentability"), and extensive discussion of their relationships, we refer to Sections 6.13 and 7.5 of [10]. In particular, the diagram on page 177 exhibits compatibility as the weakest condition that still allows oracle bounds for prediction and reconstruction by the LASSO for the ℓ^2- and ℓ^1-norms. The results on posterior contraction and model selection presented below are in the same spirit. In addition we consider contraction with respect to the ℓ^∞-norm, and for (only) the latter we employ the more restrictive mutual coherence number, analogously to the study of [32] of the LASSO and the Dantzig estimator under the supremum norm. Thus mutual coherence is useful in two ways: it may provide a simple (albeit crude) way to bound the other indices, and it may allow to use stronger norms. Direct verification of compatibility may be preferable, as this applies to a much broader set of regression matrices.

The following well-studied examples may help appreciate the discussion:

EXAMPLE 6 (Sequence model). In the sequence model of Example 1 the regression matrix X is the identity, and hence the compatibility numbers are 1, and the mutual coherence number is zero. This is the optimal situation, under which all results below are valid. (The compatibility numbers are maximally 1, as follows by evaluating them with a unit vector.)

Regression with orthogonal design can be transformed to this situation.

EXAMPLE 7 (Response model). In the response setting of Example 3 it is reasonable to assume that the entries of X are i.i.d. random variables. Under exponential moment conditions, it is shown in [12] that in this situation and for not extremely large p the mutual coherence number is with high probability bounded by a multiple of $(n/\log p)^{-1/2}$. [Specifically, this is true for $\log p = o(n)$ or $\log p = o(n^{\alpha/(4+\alpha)})$ if the entries are bounded or possess an exponential moment of order α, resp.] In view of Lemma 1 the compatibility and sparse singular value indices of models up to dimension a multiple of $\sqrt{n/\log p}$ are then bounded away from zero. This implies that the results on model selection and ℓ^1- and ℓ^2-contraction rates in the following certainly apply if the number of nonzero regression coefficients is smaller than this order. For a survey on more recent results on lower bounds of the compatibility number and the smallest sparse eigenvalue, see Section 6.2 of [45].

EXAMPLE 8. By scaling the columns of the design matrix it can be ensured that the $(p \times p)$-matrix $C := X^t X/n$ has unit diagonal. Then $\|X\| = \sqrt{n}$, and the off-diagonal elements $C_{i,j}$ are the correlations between the columns.

It is shown in [51] that if $C_{i,j}$ is equal to a constant r with $0 < r < (1 + cs)^{-1}$, or $|C_{i,j}| \leq c/(2s - 1)$, for every $i \neq j$, then models up to dimension s satisfy the "strong irrepresentability condition" and hence are consistently estimable. Since these examples satisfy the mutual coherence condition, up to a constant, these examples are also covered in the present paper, for every norm and aspect considered.

As another example, Zhao and Yu [51] consider correlations satisfying $C_{i,j} = \rho^{|i-j|}$, for $0 < \rho < 1$ and $p = n$. In this case *all* eigenvalues of C are bounded away from zero by a margin that depends on ρ only, whence the numbers $\tilde{\phi}(s)$ are bounded away from zero, for every dimension s. This implies that the results on dimensionality, model selection and ℓ^1- and ℓ^2-rates obtained below are valid. On the other hand, the mutual coherence number is equal to ρ, which excludes the ℓ^∞-results.

As a final example, the authors of [51] consider matrices C that vanish except in small blocks along the diagonal. Such matrices can also not be handled in general through the mutual coherence number, but do cooperate with the other sparse invertibility indices.

2.3. *Dimensionality*, *contraction*, *model selection*.

For simplicity the main results are stated in limit form, for $p, n \to \infty$. More precise assertions, including precise values of "large" constants, can easily be deduced from the proofs.

The results are obtained under the assumption of Gaussian noise in model (1.1). In fact, as indicated in Remark 1 in the supplementary material [15], many of the assertions are robust under misspecification of the error distribution.

The first theorem shows that the posterior distribution does not overshoot the true dimension of the parameter by more than a factor. In the interesting case that $\lambda \ll \overline{\lambda}$, this factor can be simplified to $1 + M/A_4$ for any constant $M > 2$ if the true parameter is compatible. The constant A_4 comes from condition (2.2). As a consequence, $1 + M/A_4$ can be made arbitrarily close to one by choosing a suitable prior on the dimension. (Although the convergence to zero in this and the following theorems is uniform, it can be read off from the proofs that the speed of convergence deteriorates for very small λ. Also only the dominating terms in the dependence of the dimension or contraction rate are shown. Thus the theorems as stated are not suitable for optimization over λ. In particular, it should not be concluded that the smallest possible λ is optimal.)

THEOREM 1 (Dimension). *If* λ *satisfies* (2.1) *and* π_p *satisfies* (2.2) *then*, *with* $s_0 = |S_{\beta^0}|$ *and for any* $M > 2$,

$$\sup_{\beta^0} \mathbb{E}_{\beta^0} \Pi\left(\beta : |S_\beta| > s_0 + \frac{M}{A_4}\left(1 + \frac{16}{\phi(S_0)^2} \frac{\lambda}{\overline{\lambda}}\right) s_0 \Big| Y\right) \to 0.$$

The theorem is a special case of Theorem 10 in Section 6. As all our results, the theorem concerns the full posterior distribution, not only a measure of its center. However, it may be compared to similar results for point estimators, as in Chapter 7 of [10].

The second theorem concerns the ability of the posterior distribution to recover the true model from the data. It gives rates of contraction of the posterior distribution both regarding *prediction error* $\|X\beta - X\beta^0\|_2$ and regarding the parameter β relative to the ℓ^1- and ℓ^2- and ℓ^∞-distances. Besides on the dimensionality, the rate depends on compatibility. Set

$$\begin{aligned} \overline{\psi}(S) &= \overline{\phi}\biggl(\biggl(2 + \frac{3}{A_4} + \frac{33}{\phi(S)^2}\frac{\lambda}{\overline{\lambda}}\biggr)|S|\biggr), \\ \widetilde{\psi}(S) &= \widetilde{\phi}\biggl(\biggl(2 + \frac{3}{A_4} + \frac{33}{\phi(S)^2}\frac{\lambda}{\overline{\lambda}}\biggr)|S|\biggr). \end{aligned} \tag{2.5}$$

In the interesting case that $\lambda \ll \overline{\lambda}$, these numbers are asymptotically bounded below by $\overline{\phi}((2 + \frac{4}{A_4})|S_\beta|)$ and $\widetilde{\phi}((2 + \frac{4}{A_4})|S_\beta|)$ if $\phi(S_\beta)$ is bounded away from zero. Thus the following theorem gives rates of recovery that are uniform in true vectors β such that $\phi(S_\beta)$ and $\overline{\phi}((2 + \frac{4}{A_4})|S_\beta|)$ or $\widetilde{\phi}((2 + \frac{4}{A_4})|S_\beta|)$ are bounded away from zero. [Again the theorem, even though uniform in λ satisfying (2.1), is meant to be read for a given sequence of λ.]

THEOREM 2 (Recovery). *If λ satisfies* (2.1), *and π_p satisfies* (2.2), *then for sufficiently large M, with $S_0 = S_{\beta^0}$,*

$$\sup_{\beta^0} \mathbb{E}_{\beta^0}\Pi\biggl(\beta : \|X(\beta - \beta^0)\|_2 > \frac{M}{\overline{\psi}(S_0)}\frac{\sqrt{|S_0|\log p}}{\phi(S_0)}\Big| Y\biggr) \to 0,$$

$$\sup_{\beta^0} \mathbb{E}_{\beta^0}\Pi\biggl(\beta : \|\beta - \beta^0\|_1 > \frac{M}{\overline{\psi}(S_0)^2}\frac{|S_0|\sqrt{\log p}}{\|X\|\phi(S_0)^2}\Big| Y\biggr) \to 0,$$

$$\sup_{\beta^0} \mathbb{E}_{\beta^0}\Pi\biggl(\beta : \|\beta - \beta^0\|_2 > \frac{M}{\widetilde{\psi}(S_0)^2}\frac{\sqrt{|S_0|\log p}}{\|X\|\phi(S_0)}\Big| Y\biggr) \to 0.$$

Furthermore, for every $c_0 > 0$, any $d_0 < c_0^2(1 + 2/A_4)^{-1}/8$, and s_n with $\lambda s_n\sqrt{\log p}/\|X\| \to 0$, for sufficiently large M,

$$\sup_{\substack{\beta^0:\phi(S_0)\ge c_0, \widetilde{\psi}(S_0)\ge c_0 \\ |S_0|\le s_n, |S_0|\le d_0\,\mathrm{mc}(X)^{-1}}} \mathbb{E}_{\beta^0}\Pi\biggl(\beta : \|\beta - \beta^0\|_\infty > M\frac{\sqrt{\log p}}{\|X\|}\Big| Y\biggr) \to 0.$$

The first three assertions of the theorem are consequences of the following theorem of oracle type, upon choosing $\beta^* = \beta^0$ in this theorem. The fourth assertion is proved in Section 6 under the conditions of Theorem 6 below. In the framework of Example 3, for instance say for i.i.d. Gaussian design and $\lambda = 1$, the fourth assertion is true with large probability uniformly over sparse vectors such that $|S_0| \le s_n = o(\sqrt{n/\log p})$.

An *oracle inequality* for the prediction error of a point estimator $\widehat{\beta}$ is an assertion that with large probability, and some penalty function $\mathrm{pen}(\beta)$,

$$\|X(\widehat{\beta} - \beta^0)\|_2^2 \lesssim \inf_{\beta^*} \|X(\beta^* - \beta^0)\|_2^2 + \mathrm{pen}(\beta^*);$$

see, for example, [10], Theorem 6.2, or [6] for the LASSO or the Dantzig selector. Few oracle-type results for *posterior measures* have been developed. (The results of [4], for projection estimators in white noise, are close relatives.) The following theorem is an example of such a statement. Given compatibility it shows that the bulk of the vectors β in the support of the posterior distribution satisfy an oracle inequality with penalty $\mathrm{pen}(\beta) = |S_\beta|$.

THEOREM 3 (Recovery, oracle). *If λ satisfies (2.1), and π_p satisfies (2.2), then, for $\overline{\psi}$ and $\widetilde{\psi}$ given in (2.5), there exists a constant M such that uniformly over β^0 and β^* with $|S_*| \le |S_0|$, where $S_* = S_{\beta^*}$,*

$$\mathbb{E}_{\beta^0}\Pi\biggl(\beta : \|X(\beta - \beta^0)\|_2 > \frac{M}{\overline{\psi}(S_0)}\biggl[\|X(\beta^* - \beta^0)\|_2 + \frac{\sqrt{|S_*|\log p}}{\phi(S_*)}\biggr]\Big| Y\biggr) \to 0,$$

$$\begin{aligned}\mathbb{E}_{\beta^0}\Pi\biggl(\beta : \|\beta - \beta^0\|_1 &> \|\beta^* - \beta^0\|_1 \\ &+ \frac{M}{\overline{\psi}(S_0)^2}\biggl[\frac{\|X(\beta^* - \beta^0)\|_2^2}{\|X\|\sqrt{\log p}} + \frac{|S_*|\sqrt{\log p}}{\|X\|\phi(S_*)^2}\biggr]\Big| Y\biggr) \to 0,\end{aligned}$$

$$\mathbb{E}_{\beta^0}\Pi\biggl(\beta : \|\beta - \beta^0\|_2 > \frac{M}{\|X\|\widetilde{\psi}(S_0)^2}\biggl[\|X(\beta^* - \beta^0)\|_2 + \frac{\sqrt{|S_*|\log p}}{\phi(S_*)}\biggr]\Big| Y\biggr) \to 0.$$

Besides the choice $\beta^* = \beta^0$, which yields the first three assertions of Theorem 2, other choices of β^* also give interesting results. For instance, in the sequence model of Example 1, the choice $\beta^* = 0$ gives that

$$\sup_{\beta^0} \mathbb{E}_{\beta^0}\Pi(\beta : \|\beta - \beta^0\|_2 > M\|\beta^0\|_2 | Y) \to 0.$$

For $\|\beta^0\|_2^2$ smaller than $|S_{\beta^0}| \log p$, this improves on Theorem 2, by quantifying the rate in the sizes and not only the number of nonzero coordinates in β^0.

The posterior distribution induces a distribution on the set of models $S \subset \{1, 2, \dots, p\}$, which updates the prior masses given to these models by (1.2). It is desirable that this puts most of its mass on the true model S_{β^0}. As the support of a vector β^0 is defined only in a qualitative manner by its coordinates β_i^0 being zero or not, this will not be true in general. However, the following theorem shows, under (only strong) compatibility, that the posterior distribution will not charge models that are strict supersets of the true model, no matter the magnitudes of the nonzero coordinates in β^0. This may be considered the effect of model selection through the prior π_p, which under our assumptions prefers smaller models, enough so that it will not add unnecessary coordinates when all truly nonzero coordinates are present.

THEOREM 4 (Selection: no supersets). *If λ satisfies* (2.1), *and π_p satisfies* (2.2) *with $A_4 > 1$, then for every $c_0 > 0$ and any $s_n \leq p^a$ with $s_n \lambda \sqrt{\log p}/\|X\| \to 0$ and $a < A_4 - 1$,*

$$\sup_{\substack{\beta^0:\phi(S_0)\geq c_0 \\ |S_0|\leq s_n, \widetilde{\psi}(S_0)\geq c_0}} \mathbb{E}_{\beta^0}\Pi(\beta : S_\beta \supset S_{\beta^0}, S_\beta \neq S_{\beta^0}|Y) \to 0.$$

A nonzero coordinate of β^0 that is too close to zero cannot be detected as being nonzero by any method. Consequently, the posterior distribution may well charge models S that contain only a subset of the true model S_{β^0} and possibly other coordinates, which is not excluded by the preceding theorem. The following theorem gives thresholds for detection, which become smaller as the compatibility conditions become stronger. The theorem may be compared to results in terms of *beta-min conditions* for point estimators; see, for example, [10], Corollary 7.6.

THEOREM 5 (Selection). *If λ satisfies* (2.1), *and π_p satisfies* (2.2), *then, for sufficiently large M,*

$$\inf_{\beta^0} \mathbb{E}_{\beta^0}\Pi\left(\beta : S_\beta \supset \left\{i : |\beta_i^0| \geq \frac{M}{\widetilde{\psi}(S_0)^2} \frac{|S_0|\sqrt{\log p}}{\|X\|\phi(S_0)^2}\right\}\middle|Y\right) \to 1,$$

$$\inf_{\beta^0} \mathbb{E}_{\beta^0}\Pi\left(\beta : S_\beta \supset \left\{i : |\beta_i^0| \geq \frac{M}{\widetilde{\psi}(S_0)^2} \frac{\sqrt{|S_0|\log p}}{\|X\|\phi(S_0)}\right\}\middle|Y\right) \to 1.$$

Furthermore, for every $c_0 > 0$, any $d_0 \leq c_0^2(1 + 2/A_4)^{-1}/8$, and any s_n with $\lambda s_n \sqrt{\log p}/\|X\| \to 0$,

$$\inf_{\substack{\beta^0:\phi(S_0)\geq c_0, \widetilde{\psi}(S_0)\geq c_0 \\ |S_0|\leq s_n, |S_0|\leq d_0 \,\mathrm{mc}(X)^{-1}}} \mathbb{E}_{\beta^0}\Pi\left(\beta : S_\beta \supset \left\{i : |\beta_i^0| \geq \frac{M\sqrt{\log p}}{\|X\|}\right\}\middle|Y\right) \to 1.$$

By combining Theorems 4 and 5 we see that under the assumptions of the theorems the posterior distribution *consistently selects* the correct model if *all* nonzero coordinates of β^0 are bounded away from 0 by the thresholds given in Theorem 5. For M as in the preceding theorem, let

$$\widetilde{\mathrm{B}} = \left\{\beta : \min_{i\in S_\beta} |\beta_i| \geq \frac{M}{\widetilde{\psi}(S)^2} \frac{\sqrt{|S_\beta|\log p}}{\|X\|\phi(S_\beta)}\right\}.$$

Define $\overline{\mathrm{B}}$ similarly with $\sqrt{|S_\beta|\log p}$ in the threshold replaced by $|S_\beta|\sqrt{\log p}$ and with $\overline{\psi}$ instead of $\widetilde{\psi}$.

COROLLARY 1 (Consistent model selection). *If λ satisfies (2.1), and π_p satisfies (2.2) with $A_4 > 1$, and $s_n \leq p^a$ such that $a < A_4 - 1$ and $s_n\lambda\sqrt{\log p}/\|X\| \to 0$, then, for every $c_0 > 0$,*

$$\inf_{\substack{\beta^0 \in \widetilde{\mathrm{B}}: \phi(S_0) \geq c_0 \\ |S_0| \leq s_n, \widetilde{\psi}(S_0) \geq c_0}} \mathbb{E}_{\beta^0}\Pi(\beta : S_\beta = S_{\beta^0}|Y) \to 1.$$

The same is true with $\widetilde{\mathrm{B}}$ and $\widetilde{\phi}$ replaced by $\overline{\mathrm{B}}$ and $\overline{\phi}$.

Consistent posterior model selection implies in particular, that the model with the largest posterior mass is model selection consistent in the frequentist sense. This can be established as in the proof of Theorem 2.5 in [22].

2.4. *Distributional approximation.* In this section we show that the posterior distribution can be approximated by a mixture of normal distributions. Moreover, given consistent selection of the true model, this mixture collapses to a single normal distribution. We restrict to what we shall refer to as the *small lambda regime*,

$$\frac{\lambda}{\|X\|}|S_{\beta^0}|\sqrt{\log p} \to 0. \tag{2.6}$$

In this case the centering of the normal distributions does not depend on the size of scaling parameters λ. In contrast, in the "large lambda regime," which includes the usual order of magnitude of the smoothing parameter in the LASSO, the posterior distribution mimics the LASSO, and gives a biased reconstruction of the true parameter; see Theorem 1 in the supplementary material [15].

The small lambda regime includes a variety of possible choices within our general assumption (2.1). A smaller value of λ corresponds to a noninformative prior on the nonzero coordinates of the parameter vector. Here "small" is relative, depending on the model and the number of observations.

EXAMPLE 9 (Small lambda regime). For the minimal choice $\lambda = \|X\|/p$ in (2.1) the small lambda regime (2.6) simplifies to $|S_{\beta^0}| \ll p/\sqrt{\log p}$. Thus the regime applies to a wide range of true parameters.

In the sequence model with multiple observations given in Example 2 and the response model of Example 3, we have $\|X\| = \sigma_n^{-1}$ and $\|X\| \sim n^{1/2}$, respectively, and λ is in the small lambda regime if $\lambda|S_{\beta^0}|$ is much smaller than $1/(\sigma_n\sqrt{\log p})$ and $\sqrt{n/\log p}$, respectively. The second allows $\lambda = O(1)$ if $|S_{\beta^0}|\sqrt{\log p/n} \to 0$.

For a given model $S \subset \{1, \dots, p\}$ let X_S be the $n \times |S|$-submatrix of the regression matrix X consisting of the columns $X_{\cdot,i}$ with $i \in S$, and let $\widehat{\beta}_{(S)}$ be a least square estimator in the restricted model $Y = X_S\beta_S + \varepsilon$, that is,

$$\widehat{\beta}_{(S)} \in \operatorname*{arg\,min}_{\beta_S \in \mathbb{R}^S} \|Y - X_S\beta_S\|_2^2.$$

In case the restricted model would be correctly specified, the least squares estimator would possess a $\mathcal{N}(\beta_S^0, (X_S^t X_S)^{-1})$-distribution, and the posterior distribution (in a setting where the data washes out the prior) would be asymptotically equivalent to a $\mathcal{N}(\widehat{\beta}_{(S)}, (X_S^t X_S)^{-1})$-distribution, by the Bernstein–von Mises theorem. In our present situation, the posterior distribution is approximated by a random mixture of these normal distributions, of the form

$$\Pi^\infty(\cdot|Y) = \sum_{S \in \mathcal{S}_0} \widehat{w}_S \mathcal{N}(\widehat{\beta}_{(S)}, (X_S^t X_S)^{-1}) \otimes \delta_{S^c},$$

where δ_{S^c} denotes the Dirac measure at $0 \in \mathbb{R}^{S^c}$, the weights $(\widehat{w}_S)_S$ satisfy

$$\widehat{w}_S \propto \frac{\pi_p(s)}{\binom{p}{s}} \left(\frac{\lambda}{2}\right)^s (2\pi)^{s/2} |X_S^t X_S|^{-1/2} e^{(1/2)\|X_S\widehat{\beta}_{(S)}\|_2^2} 1_{S \in \mathcal{S}_0} \tag{2.7}$$

and, for a sufficiently large M

$$\mathcal{S}_0 = \left\{S : |S| \le \left(2 + \frac{4}{A_4}\right)|S_{\beta^0}|, \|\beta_{S^c}^0\|_1 \le M|S_{\beta^0}|\sqrt{\log p}/\|X\|\right\}.$$

The weights $(\widehat{w}_S)$ are a data-dependent probability distribution on the collection of models $\mathcal{S}_0$. The latter collection can be considered a "neighborhood" of the support of the true parameter, both in terms of dimensionality and the (lack of) extension of the true parameter outside these models.

A different way of writing the approximation Π^∞ is

$$\Pi^\infty(B|Y) = \frac{\sum_{S \in \mathcal{S}_0} (\pi_p(s)/\binom{p}{s})(\lambda/2)^s \int_{B_S} e^{-(1/2)\|Y - X_S\beta_S\|_2^2}\, d\beta_S}{\sum_{S \in \mathcal{S}_0} (\pi_p(s)/\binom{p}{s})(\lambda/2)^s \int e^{-(1/2)\|Y - X_S\beta_S\|_2^2}\, d\beta_S}, \tag{2.8}$$

where $B_S = \{\beta_S : (\beta_S, 0_{S^c}) \in B\}$ is the intersection (and not projection) of $B \subset \mathbb{R}^p$ with the subspace $\mathbb{R}^S$. To see this, decompose $Y - X_S\beta_S = (Y - X_S\widehat{\beta}_{(S)}) + X_S(\widehat{\beta}_{(S)} - \beta_S)$, and observe that the two summands are orthogonal. The Lebesgue integral $d\beta_S$ can be interpreted as an improper prior on the parameter β_S of model S, and the expression as a mixture of the corresponding posterior distributions, with model weights proportional to the prior weights times $(\lambda/2)^s(2\pi)^{s/2} \int e^{-(1/2)\|Y - X_S\beta_S\|_2^2}\, d\beta_S$. It follows that the Laplace priors g_S on the nonzero coordinates wash out from the components of the posterior. On the other hand, they are still visible in the weights through the factors $(\lambda/2)^s$. In general, this influence is mild in the sense that these factors will not change the relative weights of the models much.

THEOREM 6 (Bernstein–von Mises, small lambda regime). *If λ satisfies* (2.1), *and π_p satisfies* (2.2), *then for every $c_0 > 0$ and any s_n with $s_n\lambda\sqrt{\log p}/\|X\| \to 0$,*

$$\sup_{\substack{\beta^0:\phi(S_0)\geq c_0\\ |S_0|\leq s_n,\overline{\psi}(S_0)\geq c_0}} \mathbb{E}_{\beta^0}\|\Pi(\cdot|Y) - \Pi^{\infty}(\cdot|Y)\|_{\mathrm{TV}} \to 0.$$

COROLLARY 2 (Limit under strong model selection). *Under the combined assumptions of Corollary* 1 *and Theorem* 6,

$$\sup_{\substack{\beta^0\in\widetilde{\mathrm{B}}:\phi(S_0)\geq c_0\\ |S_0|\leq s_n,\widetilde{\psi}(S_0)\geq c_0}} \mathbb{E}_{\beta^0}\|\Pi(\cdot|Y) - \mathcal{N}(\widehat{\beta}_{(S_0)}, (X_{S_0}^t X_{S_0})^{-1}) \otimes \delta_{S_0^c}\|_{\mathrm{TV}} \to 0.$$

The distributional results imply that the spread in the posterior distribution gives a correct (conservative) quantification of remaining uncertainty on the parameter. One way of making this precise is in terms of *credible sets* for the individual parameters β_j. The marginal posterior distribution of β_j is a mixture $\hat{\pi}_j\delta_0 + \hat{H}_j$ of a point mass at zero and a continuous component $\hat{H}_j$. Thus a reasonable *upper 0.975 credible limit* for β_j is equal to

$$\hat{R}_j = \begin{cases} \hat{H}_j^{-1}(0.975), & \text{if } 0.975 \leq \hat{H}_j(0), \\ 0, & \text{if } \hat{H}_j(0) \leq 0.975 \leq \hat{H}_j(0) + \hat{\pi}_j, \\ \hat{H}_j^{-1}(0.975 - \hat{\pi}_j), & \text{if } \hat{H}_j(0) + \hat{\pi}_j \leq 0.975. \end{cases}$$

It is not difficult to see that under the conditions of Corollary 2, $\mathbb{P}_{\beta^0}(\beta_j^0 \leq \hat{R}_j) \to 0.975$ if $j \in S_0$ and $\mathbb{P}_{\beta^0}(\beta_j^0 = 0) \to 1$ if $j \notin S_0$.

3. The LASSO is not fully Bayesian.

The LASSO (cf. [44])

$$\hat{\beta}_\lambda^{\mathrm{LASSO}} = \arg\min_{\beta\in\mathbb{R}^p}\bigl[\|Y - X\beta\|_2^2 + 2\lambda\|\beta\|_1\bigr] \tag{3.1}$$

is the posterior mode for the prior that models the coordinates β_i as an i.i.d. sample from a Laplace distribution with scale parameter λ, and thus also possesses a Bayesian flavor. It is well known to have many desirable properties: it is computationally tractable; with appropriately tuned smoothing parameter λ it attains good reconstruction rates; it automatically leads to sparse solutions; by small adaptations it can be made consistent for model selection under standard conditions. However, as a Bayesian object it has a deficit: in the sparse setup the full posterior distribution corresponding to the LASSO prior does not contract at the same speed as its mode. Therefore the full posterior distribution is useless for uncertainty quantification, the central idea of Bayesian inference.

We prove this in the following theorem, which we restrict to the sequence model of Example 1, that is, model (1.1) with $X = I$ the identity matrix. In this setting the LASSO estimator is known to attain the (near) minimax rate $s\log n$ for the square Euclidean loss over the "nearly black bodies" $\{\beta : |S_\beta| \le s\}$, and a near minimax rate over many other sparsity classes as well, if the regularity parameter λ is chosen of the order $\sqrt{2\log n}$. The next theorem shows that for this choice the LASSO posterior distribution $\Pi_\lambda^{\mathrm{LASSO}}(\cdot|Y)$ puts no mass on balls of radius of the order $\sqrt{n}/(\log n)^{1/2}$, which is substantially bigger than the minimax rate $(s\log n)^{1/2}$ (except for extremely dense signals).

Intuitively, this is explained by the fact that the parameter λ in the Laplace prior must be large in order to shrink coefficients β_i to zero, but at the same time reasonable so that the Laplace prior can model the nonzero coordinates. That these conflicting demands do not affect the good behavior of the LASSO estimators must be due to the special geometric, sparsity-inducing form of the posterior mode, not to the Bayesian connection.

THEOREM 7. *Assume that we are in the setting of Example 1. For any $\lambda = \lambda_n$ such that $\sqrt{n}/\lambda_n \to \infty$, there exists $\delta > 0$ such that, as $n \to \infty$,*

$$\mathbb{E}_{\beta^0=0}\Pi_{\lambda_n}^{\mathrm{LASSO}}\biggl(\beta : \|\beta\|_2 \le \delta\sqrt{n}\biggl(\frac{1}{\lambda_n} \wedge 1\biggr)\bigg|Y\biggr) \to 0.$$

4. Prediction for arbitrary design.

The vector $X\beta$ is the mean vector of the observation Y in (1.1), and one might guess that this is estimable without identifiability conditions on the regression matrix X. In this section we show that the posterior distribution based on the prior (1.2) can indeed solve this *prediction problem* at (nearly) optimal rates under no condition on the design matrix X. These results are inspired by [17] and Theorem 8 below can be seen as a full Bayesian version of the results on the PAC-Bayesian point estimators in the latter paper; see also [36] for prediction results for mixtures of least-squares estimators.

We are still interested in the sparse setting, and hence the regression matrix X still intervenes by modeling the unknown mean vector $\mathbb{E}Y$ as a linear combination of a small set of its columns.

First, we consider the case of priors (1.2) that model the mean vector indirectly by modeling the set of columns and the coefficients of the linear combination. The prior $\pi_p(s)$ comes in through the constant

$$C_\pi = \sum_{s=0}^{p} 9^s \binom{p}{s}^{1/2} \sqrt{\pi_p(s)}. \tag{4.1}$$

For the choice of prior on coordinates β_i, the best results are obtained with heavy-tailed densities g. In general the rate depends on the Kullback–Leibler divergence between the measure with distribution function G_{S_0} (corresponding to the prior density g_{S_0}) and the same measure shifted by $\beta_{S_0}^0$. Let KL be the Kullback–Leibler divergence, and set

$$D_{\beta^0} = \frac{\binom{p}{s_0}}{\pi_p(s_0)} e^{\mathrm{KL}(G_{S_0}(\cdot-\beta_{S_0}^0), G_{S_0}) + (1/2)\int \|X\beta_{S_0}\|_2^2 \, dG_{S_0}(\beta_{S_0})}. \tag{4.2}$$

THEOREM 8. *For any prior π_p and C_π as in* (4.1), *any density g that is symmetric about* 0, *any $\beta^0, \beta^* \in \mathbb{R}^p$ and $r \geq 1$*,

$$\mathbb{E}_{\beta^0}\Pi\Big(\|X(\beta-\beta^0)\|_2 > 7\|X(\beta^*-\beta^0)\|_2 + 4\sqrt{\log(C_\pi^2 D_{\beta^*})} + 8\sqrt{r}|Y\Big) \lesssim e^{-r}.$$

If the prior on the dimension satisfies (2.2) with $A_4 > 1$, then C_π is bounded in p, and the rate for squared error loss is determined by

$$\begin{aligned}\rho_n(\beta^0) &:= \log D_{\beta^0}\\ &\lesssim |S_{\beta^0}|\log p + \mathrm{KL}(G_{S_0}(\cdot - \beta^0_{S_0}), G_{S_0}) + \frac{1}{2}\int \|X\beta_{S_0}\|_2^2\, dG_{S_0}(\beta_{S_0}).\end{aligned}$$

This rate might be dominated by the Kullback–Leibler divergence for large signal β^0. However, for heavy tailed priors g the induced constraints on the signal to achieve the good rate $|S_{\beta^0}|\log p$ are quite mild. Consider the prior distribution (1.2) with g_S a product of $|S|$ univariate densities g of the form

$$g(x) \propto \frac{\lambda}{1+|\lambda x|^\mu}, \qquad x \in \mathbb{R}, \lambda > 0, \mu > 3. \tag{4.3}$$

COROLLARY 3. *If π_p satisfies* (2.3) *with $a \geq 1$, and g is of the form* (4.3) *with $\lambda = \|X\|$ and $\mu > 3$, then for sufficiently large M*,

$$\sup_{\beta^0}\mathbb{E}_{\beta^0}\Pi(\beta \in \mathbb{R}^p : \|X\beta - X\beta^0\|_2^2 > M\rho_n(\beta^0)|Y) \to 0,$$

for $\rho_n(\beta) = |S_\beta|\log p \vee \sum_{i\in S_\beta}\log(1+\|X\|^\mu|\beta_i|^\mu)$.

REMARK 1. The constant 7 in Theorem 8 can be improved to $4+\delta$, for an arbitrary $\delta > 0$, by a slight adaptation of the argument. Using PAC-Bayesian techniques Dalalyan and Tsybakov [17] obtain an oracle inequality with leading constant 1 for a so-called pseudo-posterior mean: the likelihood in (1.4) is raised to some power, which amounts to replacing the 1/2 factor by $1/\beta$. The "inverse temperature" β must be taken large enough; the case $\beta = 2$ corresponding to the Bayes posterior as considered here is not included; see also [30].

Theorem 8 and its corollary address the question of achieving prediction with no condition on X, and the same rate is achieved as in Section 2 with the same type of priors, up to some slight loss incurred only for true vectors β^0 with very large entries. As shown in the corollary, this slight dependence on β^0 can be made milder with flatter priors. We now consider a different approach specifically targeted at the prediction problem and which enables to remove dependency on the size of the coordinates of β^0 completely.

Because the prediction problem is concerned only with the mean vector, and the columns of X will typically be linearly dependent, it is natural to define the prior distribution directly on the corresponding subspaces. For any $S \subset \{1, \dots, p\}$, let $\mathcal{Y}_S := \{X\beta, S_\beta \subseteq S\}$ be the subspace of $\mathbb{R}^n$ generated by the columns X^j, $j \in S$ of X. Let $\mathcal{V}$ denote the collection of all *distinct* subspaces $\mathcal{Y}_S$.

Define a (improper) prior Ξ on $\mathbb{R}^n$ by first selecting an integer t in $\{0, 1, \dots, n\}$ according to a prior π_n, next given t selecting a subspace $V \in \mathcal{V}$ of dimension t uniformly at random among subspaces in $\mathcal{V}$ of dimension t; finally, let Ξ given V be defined as Lebesgue measure on V if $\dim(V) \geq 1$, and let Ξ be the Dirac mass at $\{0\}$ for $V = \{0\}$. Note that the posterior distribution $\Xi[\cdot|Y]$ is a well-defined probability measure on $\mathbb{R}^n$.

We choose, for a fixed $d \geq 4$ (the numerical constant 4 is for simplicity),

$$\pi_n(t) := \pi_{n,p}(t) = \frac{e^{-dt \log p}}{\sum_{t=0}^{n} e^{-dt \log p}}, \qquad t = 0, 1, \dots, n. \tag{4.4}$$

Let $V^0 := \mathcal{Y}_{S_{\beta^0}}$ and t_0 be the dimension of V^0.

THEOREM 9. *Let Ξ be the improper prior on $\mathbb{R}^n$ defined above with π_n as in* (4.4). *For M large enough*,

$$\sup_{\beta^0} \mathbb{E}_{\beta^0} \Xi\left[\gamma \in \mathbb{R}^n, \|\gamma - X\beta^0\|_2^2 > M(t_0 \vee 1) \log p | Y\right] \to 0.$$

The result is uniform in $\beta^0 \in \mathbb{R}^p$. Also, note that $t_0 \leq |S_{\beta^0}| = s_0$ and that one may have $t_0 = o(s_0)$. The obtained rate thus may improve on the previous prediction rates. It has a simple interpretation: up to an additional logarithmic factor, it is the rate of the natural estimate $\gamma^* = \mathrm{Proj}_{V^0} Y$ if the true subspace V^0 is known, where Proj_{V^0} denotes the orthogonal projection in $\mathbb{R}^n$ into the subspace V^0.

5. Computational algorithms. In this section we survey computational methods to compute posterior distributions in the regression model (1.1) based on model selection priors (1.2). In most cases, this is a "spike and slab" prior, as discussed in Example 5, implemented with auxiliary 0–1 variables that indicate whether a parameter β_j is included in the model or not. The slab distribution is typically chosen a scale mixture of Gaussian distributions, which may include the Laplace law, which is an exponential mixture. Most implementations also allow an unknown error variance (which is taken to be unity in the present paper), with the inverse gamma distribution as the favorite prior.

For low-dimensional regression problems, computation of the posterior given mixture priors was studied by many authors, including [18, 21, 26, 27, 34, 41, 48]. Higher-dimensional settings have been considered recently: most of the following papers have appeared in the last five years, and a number of them are preprints.

Several authors [7, 8, 25, 31, 35, 43] have implemented *MCMC* schemes to simulate from the posterior distribution, coupled with *stochastic search* algorithms that limit the model space, so as to alleviate the curse of dimensionality. Besides computation time, monitoring the convergence of the samplers is an issue. For higher dimensions it is impossible to sample from the complete model space, but this should also not be necessary, as in sparse situations the posterior will concentrate on lower-dimensional spaces, as is also apparent from our theoretical results. Bottolo et al. [7] provide ready-made software, which runs on dimensions up to several thousands. The same authors have also exploited hardware solutions, such as graphical processing units, to speed up computations in genomic data analyses.

Sequential Monte Carlo methods or *particle filters* can be viewed as MCMC schemes that can more readily incorporate correct moves in the model space that ensure good approximation to the posterior distribution. In [39, 42] such methods are shown to perform well for model selection in regression models with up to hundreds of covariates.

The *shrinkage-thresholding Metropolis adjusted Langevin algorithm* (or STMALA) introduced in [40] is another variation on earlier MCMC algorithms, targeted to work for $p > n$, in, for instance, imaging applications. It jointly samples a model and a regression vector in this model, using proposals based on the gradient of the logarithm of the smooth part of the posterior distribution (as in MALA) combined with applying a shrinkage-thresholding operator to set coordinates to zero. Geometric convergence of the algorithm, which is capable of moving between rather distant models, is guaranteed for slab prior densities of the form $\propto \exp(-\lambda\|\beta\|_1 - \mu\|\beta\|_2^2)$, where $\mu > 0$. Illustrations showing good practical performance are given in [40] (Section 5.2) for values of (n, p) equal to $(100, 200)$ or $(39, 300)$.

An alternative to simulation from the exact posterior is to compute an exact, analytic approximation to the posterior. A relatively simple and computationally efficient *variational Bayes approximation* is proposed in [47] and is shown to perform satisfactorily, but examples in the paper are limited to cases where $p \leq n$.

By relaxing the spike at zero to a Gaussian distribution with small variance, Ročková and George [38] succeeded in reducing computations of aspects of the posterior distribution, such as means and moments, to iterations of an efficient *EM-algorithm*. They show good performance with exponentially decreasing priors on model dimension, as considered in the present paper.

Closely related to the spike and slab prior is *exponential weighting*, where each of the 2^p models is given a prior weight, which is then updated with the likelihood function. A survey and numerical simulations in high-dimensional settings using the Metropolis–Hastings algorithm can be found in [37]. Stable reconstructions in dimensions up to $p = 500$, $n = 200$ and sparsity level $s_0 = 20$ are shown to require usually no more than 2000 iterations.

An (empirical, pseudo-) Bayes approach with a spike and Gaussian slabs centered at the least square solutions of the underlying model is implemented in [33]. The algorithm, which can be initialized at the LASSO estimator, is shown to perform well for n up to 100 and p up to 1000. Because the slabs are centered on data-based quantities, the target of this algorithm is different from the posterior distribution in the present paper. However, since the prior puts mass on all models, its computational complexity is comparable to the procedure in the present paper.

For the sequence model of Example 1, an algorithm to compute posterior quantities such as modes and quantiles based on *generating polynomials* is implemented in [16]. This is efficient in terms of computation time, but requires large memory. Up to $n = p = 500$ standard software and hardware suffice. The method may be extended to other designs by making suitable transformations [5].

6. Proofs for Section 2. Denote by $p_{n,\beta}$ the density of the $\mathcal{N}(X\beta, I)$-distribution, and the corresponding log likelihood ratios by

$$\Lambda_{n,\beta,\beta^*}(Y) = \frac{p_{n,\beta}}{p_{n,\beta^*}}(Y) = e^{-(1/2)\|X(\beta-\beta^*)\|_2^2+(Y-X\beta^*)^t X(\beta-\beta^*)}. \tag{6.1}$$

LEMMA 2. *For p sufficiently large and any $\beta^* \in \mathbb{R}^p$, with support S_* and $s_* := |S_*|$, and Π given by* (1.2) *with g_S a product of Laplace densities with scale λ, we have, almost surely,*

$$\int \Lambda_{n,\beta,\beta^*}(Y)\, d\Pi(\beta) \ge \frac{\pi_p(s_*)}{p^{2s_*}} e^{-\lambda\|\beta^*\|_1} e^{-1}.$$

PROOF. For $s_* = 0$ the right-hand side is $\pi_p(0)e^{-1}$, while the left-hand side is bounded below by $\Lambda_{n,0,0}\pi_p(0) = \pi_p(0)$, by (1.2). Thus we may assume that $s_* \ge 1$.

First we prove that for any set S and $s = |S| > 0$,

$$\int_{\|\beta_S\|_1 \le r} g_S(\beta_S)\, d\beta_S = e^{-\lambda r} \sum_{k=s}^{\infty} \frac{(\lambda r)^k}{k!} \ge e^{-\lambda r} \frac{(\lambda r)^s}{s!}. \tag{6.2}$$

If $(L_i)_{i=1,\dots,s}$ are i.i.d. random variables with the Laplace distribution with scale parameter λ, then $(|L_i|)_{i=1,\dots,s}$ are i.i.d. exponential variables of the same scale. Hence the left-hand side of the display, which is equal to $\mathbb{P}(\sum_{i=1}^s |L_i| \le r)$, is the probability that the first s events of a Poisson process of intensity λ occur before time r. This is identical to the probability that the Poisson process has s or more events in $[0, r]$, which is the sum in the display.

By (1.2), the left-hand side of the lemma is bounded below by

$$\begin{aligned} &\frac{\pi_p(s_*)}{\binom{p}{s_*}} \int \Lambda_{n,\beta,\beta^*}(Y) g_{S_*}(\beta_{S_*})\, d\beta_{S_*} \\ &\qquad \ge \frac{\pi_p(s_*)}{\binom{p}{s_*}} e^{-\lambda\|\beta^*\|_1} \int e^{-(1/2)\|Xb_{S_*}\|_2^2+(Y-X\beta^*)^t Xb_{S_*}} g_{S_*}(b_{S_*})\, db_{S_*}, \end{aligned}$$

by (6.1), the change of variables $\beta_{S_*} - \beta^*_{S_*} \to b_{S_*}$ and the inequality $g_{S_*}(\beta_{S_*}) \ge e^{-\lambda\|\beta^*\|_1} g_{S_*}(b_{S_*})$. The finite measure μ defined by the identity $d\mu =$

$\exp(-\frac{1}{2}\|Xb_{S_*}\|_2^2)g_{S_*}(b_{S_*})\,db_{S_*}$ is symmetric about zero, and hence the mean of b_{S_*} relative to μ is zero. Let $\bar{\mu}$ denote the normalized probability measure corresponding to μ, that is, $\bar{\mu} := \mu/\mu(\mathbb{R}^{|S_*|})$. Let $\mathbb{E}_{\bar{\mu}}$ denote the expectation operator with respect to $\bar{\mu}$. Define $Z(b_{S_*}) := (Y - X\beta^*)^t Xb_{S_*}$. By Jensen's inequality $\mathbb{E}_{\bar{\mu}} \exp(Z) \geq \exp(\mathbb{E}_{\bar{\mu}} Z)$. However, $\mathbb{E}_{\bar{\mu}} Z = 0$, by the just mentioned symmetry of μ. So the last display is bounded below by

$$\frac{\pi_p(s_*)}{\binom{p}{s_*}} e^{-\lambda\|\beta^*\|_1} \int e^{-(1/2)\|Xb_{S_*}\|_2^2} g_{S_*}(b_{S_*})\,db_{S_*},$$

almost surely. Using that $\|X\beta\|_2 = \|\sum_{i=1}^p \beta_i X_{\cdot,i}\|_2 \leq \|\beta\|_1\|X\|$, and then (6.2), we find that the integral in the last display is bounded below by

$$e^{-1/2}\int_{\|X\|\|b_{S_*}\|_1\leq 1} g_{S_*}(b_{S_*})\,db_{S_*} \geq e^{-1/2}\frac{e^{-\lambda/\|X\|}(\lambda/\|X\|)^{s_*}}{s_*!}.$$

With (2.1), $e^{-\lambda/\|X\|}(\lambda/\|X\|)^{s_*}$ is bounded from below by $e^{-1/2}p^{-s_*}$, if $\lambda/\|X\| \leq 1/2$ and by $e^{-4\sqrt{\log p}}2^{-s_*}$, if $\lambda/\|X\| \geq 1/2$. Since $s_* > 0$ and $e^{-4\sqrt{\log p}}$ decays to zero slower than any polynomial power of p, we find $e^{-\lambda/\|X\|}(\lambda/\|X\|)^{s_*} \geq e^{-1/2}p^{-s_*}$ in both cases, provided that p is sufficiently large. The lemma follows upon substituting these bounds and the bound $\binom{p}{s}s! \leq p^s$ in the display. □

LEMMA 3. *For any $\beta, \beta^* \in \mathbb{R}^p$ and random variable $0 \leq U = U(Y) \leq 1$,*

$$\mathbb{E}_\beta U \leq e^{(1/2)\|X(\beta-\beta^*)\|_2^2}(\mathbb{E}_{\beta^*}U)^{1/2}.$$

PROOF. Write the left-hand side as $\mathbb{E}_{\beta^*}[U\Lambda_{n,\beta,\beta^*}]$, and use the Cauchy–Schwarz inequality; see, for example, [16], Lemma 6.1. □

LEMMA 4.

$$\mathbb{P}_{\beta^0}\big(\|X^t(Y - X\beta^0)\|_\infty > 2\sqrt{\log p}\|X\|\big) \leq \frac{2}{p}.$$

PROOF. Under the probability measure $\mathbb{P}_{\beta^0}$ the vector $\varepsilon = Y - X\beta^0$ possesses an n-dimensional standard normal distribution, whence the p coordinates of the vector $X^t\varepsilon$ are normal with variances $(X^tX)_{i,i} \leq \|X\|$. Now $\mathbb{P}(\|X^t\varepsilon\|_\infty > x) \leq \sum_{i=1}^p \mathbb{P}((X^t\varepsilon)_i > x)$, which can be bounded by the tail bound for the normal distribution. □

THEOREM 10 (Dimension, general result). *If λ satisfies* (2.1) *and the prior π_p satisfies* (2.2), *then for any $M > 2$,*

$$\sup_{\beta^0,\beta^*} \mathbb{E}_{\beta^0}\Pi\left(\beta : |S_\beta| > |S_*| + \frac{M}{A_4}\left(1 + \frac{16}{\phi(S_*)^2}\frac{\lambda}{\bar{\lambda}}\right)|S_*| + \frac{M}{A_4}\frac{\|X(\beta^0 - \beta^*)\|_2^2}{\log p}\Big|Y\right)$$
$$\to 0.$$

PROOF. By the definition of $\overline{\lambda}$ in (2.1) and Lemma 4, the complement of the event $\mathcal{T}_0 = \{\|X^t(Y - X\beta^0)\|_\infty \le \overline{\lambda}\}$ has $\mathbb{P}_{\beta^0}$-probability bounded by $2/p$. By combining this with Lemma 3 we see that for any β^* and any measurable set $B \subset \mathbb{R}^p$,

$$\mathbb{E}_{\beta^0}\Pi(B|Y) \le e^{(1/2)\|X(\beta^0-\beta^*)\|_2^2}\bigl(\mathbb{E}_{\beta^*}[\Pi(B|Y)\mathbb{I}_{\mathcal{T}_0}]\bigr)^{1/2} + \frac{2}{p}. \tag{6.3}$$

By Bayes's formula followed by Lemma 2, with $\Lambda_{n,\beta,\beta^*}(Y)$ the likelihood ratio given in (6.1),

$$\begin{aligned}\Pi(B|Y) &= \frac{\int_B \Lambda_{n,\beta,\beta^*}(Y)\,d\Pi(\beta)}{\int \Lambda_{n,\beta,\beta^*}(Y)\,d\Pi(\beta)}\\ &\le \frac{ep^{2s_*}}{\pi_p(s_*)}e^{\lambda\|\beta^*\|_1}\int_B e^{-(1/2)\|X(\beta-\beta^*)\|_2^2+(Y-X\beta^*)^tX(\beta-\beta^*)}\,d\Pi(\beta).\end{aligned} \tag{6.4}$$

Using Hölder's inequality $|\alpha^t\beta| \le \|\alpha\|_\infty\|\beta\|_1$ and the Cauchy–Schwarz inequality, we see that on the event $\mathcal{T}_0$,

$$\begin{aligned}&(Y - X\beta^*)^tX(\beta - \beta^*)\\ &\quad= (Y - X\beta^0)^tX(\beta - \beta^*) + (X\beta^0 - X\beta^*)^tX(\beta - \beta^*)\\ &\quad\le \overline{\lambda}\|\beta - \beta^*\|_1 + \|X(\beta^0 - \beta^*)\|_2^2 + \tfrac{1}{4}\|X(\beta - \beta^*)\|_2^2 =: L(\beta).\end{aligned} \tag{6.5}$$

Therefore, on the event $\mathcal{T}_0$, the expected value under $\mathbb{E}_{\beta^*}$ of the integrand on the right-hand side of (6.4) is bounded above by

$$\begin{aligned}&e^{-(1/2)\|X(\beta-\beta^*)\|_2^2}\mathbb{E}_{\beta^*}\bigl[e^{(1-\lambda/(2\overline{\lambda}))(Y-X\beta^*)^tX(\beta-\beta^*)}\mathbb{I}_{\mathcal{T}_0}\bigr]e^{(\lambda/(2\overline{\lambda}))L(\beta)}\\ &\quad= e^{-(1/2)(1-(1-\lambda/(2\overline{\lambda}))^2)\|X(\beta-\beta^*)\|_2^2}e^{(\lambda/(2\overline{\lambda}))L(\beta)}\\ &\quad\le e^{\|X(\beta^0-\beta^*)\|_2^2}e^{-(\lambda/(8\overline{\lambda}))\|X(\beta-\beta^*)\|_2^2+(\lambda/2)\|\beta-\beta^*\|_1},\end{aligned}$$

where we use that $\lambda \le 2\overline{\lambda}$. It follows that the expected value $\mathbb{E}_{\beta^*}[\Pi(B|Y)\mathbb{I}_{\mathcal{T}_0}]$ under β^* of (6.4) over $\mathcal{T}_0$ is bounded above by

$$\le \frac{ep^{2s_*}}{\pi_p(s_*)}e^{\lambda\|\beta^*\|_1}\int_B e^{\|X(\beta^0-\beta^*)\|_2^2}e^{-(\lambda/(8\overline{\lambda}))\|X(\beta-\beta^*)\|_2^2+(\lambda/2)\|\beta-\beta^*\|_1}\,d\Pi(\beta). \tag{6.6}$$

By the triangle inequality,

$$\begin{aligned}\|\beta^*\|_1 + \tfrac{1}{2}\|\beta - \beta^*\|_1 &\le \|\beta_{S_*}\|_1 + \tfrac{3}{2}\|\beta_{S_*} - \beta^*\|_1 + \tfrac{1}{2}\|\beta_{S_*^c}\|_1\\ &\le -\tfrac{1}{4}\|\beta - \beta^*\|_1 + \|\beta\|_1,\end{aligned}$$

for $7\|\beta_{S_*} - \beta^*\|_1 \le \|\beta_{S_*^c}\|_1$, as is seen by splitting the norms on the right-hand side over S_* and S_*^c. If $\|\beta_{S_*^c}\|_1 < 7\|\beta_{S_*} - \beta^*\|_1$, then we write $3/2 = 2 - 1/2$ and use the definition of the compatibility number $\phi(S_*)$ to find that

$$\begin{aligned}
&\|\beta_{S_*}\|_1 + \frac{3}{2}\|\beta_{S_*} - \beta^*\|_1 + \frac{1}{2}\|\beta_{S_*^c}\|_1 \\
&\quad \le \|\beta_{S_*}\|_1 + 2\frac{\|X(\beta - \beta^*)\|_2 s_*^{1/2}}{\|X\|\phi(S^*)} - \frac{1}{2}\|\beta_{S_*} - \beta^*\|_1 + \frac{1}{2}\|\beta_{S_*^c}\|_1 \\
&\quad \le \|\beta_{S_*}\|_1 + \frac{1}{8\overline{\lambda}}\|X(\beta - \beta^*)\|_2^2 + \frac{8s_*\overline{\lambda}}{\|X\|^2\phi(S_*)^2} - \frac{1}{4}\|\beta - \beta^*\|_1 + \|\beta\|_1.
\end{aligned}$$

We combine the last three displays to see that (6.6) is bounded above by

$$\frac{ep^{2s_*}}{\pi_p(s_*)} e^{\|X(\beta^0-\beta^*)\|_2^2} e^{8\lambda\overline{\lambda}s_*/(\|X\|^2\phi(S_*)^2)} \int_B e^{-(\lambda/4)\|\beta-\beta^*\|_1+\lambda\|\beta\|_1}\, d\Pi(\beta).$$

For the set $B = \{\beta : |S_\beta| > R\}$ and $R \ge s_*$, the integral in this expression is bounded above by

$$\begin{aligned}
\sum_{S:|S|>R} \frac{\pi_p(s)}{\binom{p}{s}} \left(\frac{\lambda}{2}\right)^s \int e^{-(\lambda/4)\|\beta_S - \beta^*\|_1}\, d\beta_S
&\le \sum_{s=R+1}^{p} \pi_p(s) 4^s \\
&\le \pi_p(s_*) 4^{s_*} \left(\frac{4A_2}{p^{A_4}}\right)^{R+1-s_*} \sum_{j=0}^{\infty} \left(\frac{4A_2}{p^{A_4}}\right)^j,
\end{aligned}$$

by assumption (2.2). Combining the preceding with (6.3), we see that

$$\mathbb{E}_{\beta^0}\Pi(B|Y) \lesssim (4p)^{s_*} e^{\|X(\beta^0-\beta^*)\|_2^2 + 4\lambda\overline{\lambda}s_*/(\|X\|^2\phi(S_*)^2)} \left(\frac{4A_2}{p^{A_4}}\right)^{(R+1-s_*)/2} + \frac{2}{p}.$$

Using that $\overline{\lambda}^{-2} = 4\|X\|^2 \log p$, we can infer the theorem by choosing $R = s_* + MA_4^{-1}(\|X(\beta^0 - \beta^*)\|_2^2/\log p + s_* + 16s_*(\lambda/\overline{\lambda})/\phi(S_*)^2)$ for fixed $M > 2$. □

PROOF OF THEOREM 3. By Theorem 10 the posterior distribution is asymptotically supported on the event $E := \{\beta : |S_\beta| \le D_* \wedge D_0\}$, for

$$D_* = \left(1 + \frac{3}{A_4} + \frac{33}{A_4\phi(S_*)^2}\frac{\lambda}{\overline{\lambda}}\right)s_* + \frac{3}{A_4}\frac{\|X(\beta^0 - \beta^*)\|_2^2}{\log p} \tag{6.7}$$

and D_0 the same expression with β^* replaced by β^0. Thus it suffices to prove that the intersections of the events in the theorem with the event E tends to zero. By combining (6.4), (6.5) and the inequality $\lambda\|\beta^*\|_1 \le 2\overline{\lambda}\|\beta - \beta^*\|_1 + \lambda\|\beta\|_1$, we see that on the event $\mathcal{T}_0 = \{\|X^t(Y - X\beta^0)\|_\infty \le \overline{\lambda}\}$, the variable $\Pi(B|Y)$ is bounded above by

$$\frac{ep^{2s_*}}{\pi_p(s_*)} \int_B e^{-(1/4)\|X(\beta-\beta^*)\|_2^2 + 3\overline{\lambda}\|\beta-\beta^*\|_1 + \|X(\beta^0-\beta^*)\|_2^2 + \lambda\|\beta\|_1}\, d\Pi(\beta).$$

By Definition 2.2 of the uniform compatibility number,

$$(4-1)\overline{\lambda}\|\beta - \beta^*\|_1 \le \frac{4\overline{\lambda}\|X(\beta-\beta^*)\|_2 |S_{\beta-\beta^*}|^{1/2}}{\|X\|\overline{\phi}(|S_{\beta-\beta^*}|)} - \overline{\lambda}\|\beta - \beta^*\|_1 \tag{6.8}$$

$$\le \frac{1}{8}\|X(\beta-\beta^*)\|_2^2 + \frac{32\overline{\lambda}^2 |S_{\beta-\beta^*}|}{\|X\|^2\overline{\phi}(|S_{\beta-\beta^*}|)^2} - \overline{\lambda}\|\beta - \beta^*\|_1. \tag{6.9}$$

Since $|S_{\beta-\beta^*}| \le |S_\beta| + s_* \le D_* \wedge D_0 + s_*$, on the event E and $s_* \le s_0$ by assumption, it follows from (2.5) that for a set $B \subset E$,

$$\begin{aligned} \Pi(B|Y)\mathbb{I}_{\mathcal{T}_0} \le {} & \frac{ep^{2s_*}}{\pi_p(s_*)} e^{\|X(\beta^0-\beta^*)\|_2^2 + 32\overline{\lambda}^2(D_*+s_*)/(\|X\|^2\overline{\psi}(S_0)^2)} \\ & \times \int_B e^{-(1/8)\|X(\beta-\beta^*)\|_2^2 - \overline{\lambda}\|\beta-\beta^*\|_1 + \lambda\|\beta\|_1}\, d\Pi(\beta). \end{aligned} \tag{6.10}$$

Since $\mathbb{P}_{\beta^0}(\mathcal{T}_0) \le 2/p$ it suffices to show that the right-hand side tends to zero for the relevant event B.

Proof of first assertion. On the set $B := \{\beta \in E : \|X(\beta - \beta^0)\|_2 > 4\|X(\beta^* - \beta^0)\|_2 + R\}$, we have $\|X(\beta - \beta^*)\|_2^2 > 9\|X(\beta^* - \beta^0)\|_2^2 + R^2$, by the triangle inequality. Note that $\pi_p(s_*) \ge (A_1 p^{-A_3})^{s_*}\pi_p(0)$. It follows that for the set B, the preceding display is bounded above by

$$\begin{aligned} & \frac{ep^{2s_*}}{\pi_p(s_*)} e^{32\overline{\lambda}^2(D_*+s_*)/(\|X\|^2\overline{\psi}(S_0)^2)} e^{-(1/8)R^2} \int e^{-\overline{\lambda}\|\beta-\beta^*\|_1 + \lambda\|\beta\|_1}\, d\Pi(\beta) \\ & \quad \lesssim p^{(2+A_3)s_*} A_1^{-s_*} e^{32\overline{\lambda}^2(D_*+s_*)/(\|X\|^2\overline{\psi}(S_0)^2)} e^{-(1/8)R^2} \sum_{s=0}^{p} \pi_p(s) 2^s, \end{aligned}$$

by (2.2) and a calculation similar to the proof of Theorem 10. For

$$\begin{aligned} \frac{1}{8}R^2 &= (3 + A_3)s_* \log p + \frac{32\overline{\lambda}^2(D_* + s_*)}{\|X\|^2\overline{\psi}(S_0)^2} \lesssim \frac{\log p(D_* + s_*)}{\overline{\psi}(S_0)^2} \\ &=: R_*^2, \end{aligned}$$

this tends to zero. Thus we have proved that for some sufficiently large constant M,

$$\mathbb{E}_{\beta^0}\Pi(\beta : \|X(\beta - \beta^0)\|_2 \ge 4\|X(\beta^* - \beta^0)\|_2 + MR_*|Y) \to 0.$$

Proof of second assertion. Similar to (6.8),

$$\begin{aligned}&\bar{\lambda}\|\beta-\beta^0\|_1\\&\quad\le\bar{\lambda}\|\beta^*-\beta^0\|_1+\frac{1}{2}\|X(\beta-\beta^*)\|_2^2+\frac{\bar{\lambda}^2|S_{\beta-\beta^*}|}{2\|X\|^2\overline{\psi}(S_0)^2}\\&\quad\le\|X(\beta-\beta^0)\|_2^2+\bar{\lambda}\|\beta^*-\beta^0\|_1+\|X(\beta^*-\beta^0)\|_2^2+\frac{\bar{\lambda}^2|S_{\beta-\beta^*}|}{2\|X\|^2\overline{\psi}(S_0)^2}.\end{aligned}$$

The claim follows now from the first assertion.

Proof of third assertion. Note that $\|X(\beta-\beta^0)\|_2\ge\widetilde{\phi}(|S_{\beta-\beta^0}|)\|X\|\|\beta-\beta^0\|_2\ge\widetilde{\psi}(S_0)\|X\|\|\beta-\beta^0\|_2$. Now, the proof follows from the first assertion. □

PROOF OF THEOREM 6. The total variation distance between a probability measure Π and its renormalized restriction $\Pi_A(\cdot):=\Pi(\cdot\cap A)/\Pi(A)$ to a set A is bounded above by $2\Pi(A^c)$. We apply this to both the posterior measure $\Pi(\cdot|Y)$ and the approximation $\Pi^\infty(\cdot|Y)$, with the set

$$A:=\left\{\beta:\|\beta-\beta^0\|_1\le\frac{Ms_0\sqrt{\log p}}{\|X\|\overline{\psi}(S_0)^2\phi(S_0)^2}\right\},$$

where M is a sufficiently large constant. By Theorem 2 the probability $\Pi(A|Y)$ tends to one under $\mathbb{P}_{\beta^0}$, and at the end of this proof we show that $\Pi^\infty(A|Y)$ tends to one as well. Hence it suffices to prove Theorem 6 with $\Pi(\cdot|Y)$ and $\Pi^\infty(\cdot|Y)$ replaced by their renormalized restrictions to A.

The measure $\Pi_A^\infty(\cdot|Y)$ is by its definition a mixture over measures corresponding to models $S\in\mathcal{S}_0$. By Theorems 1 and 2 the measure $\Pi_A(\cdot|Y)$ is asymptotically concentrated on these models. If $(\tilde{v}_S)$ is the renormalized restriction of a probability vector (v_S) to a set $\mathcal{S}_0$, then, for any probability measures Π_S,

$$\left\|\sum_S\tilde{v}_S\Pi_S-\sum_S v_S\Pi_S\right\|_{\mathrm{TV}}\le\|(\tilde{v}_S)-(v_S)\|_{\mathrm{TV}}\le 2\sum_{S\notin\mathcal{S}_0}v_S,$$

by the preceding paragraph. We infer that we can make a further reduction by restricting and renormalizing the mixing weights of $\Pi(\cdot|Y)$ to $\mathcal{S}_0$. More precisely, define probability measures by

$$\Pi^{(1)}(B|Y)\propto\sum_{S\in\mathcal{S}_0}\frac{\pi_p(s)}{\binom{p}{s}}\left(\frac{\lambda}{2}\right)^s\int_{(B\cap A)_S}e^{-(1/2)\|Y-X_S\beta_S\|_2^2}e^{-\lambda\|\beta_S\|_1}\,d\beta_S,$$

$$\Pi^{(2)}(B|Y)\propto\sum_{S\in\mathcal{S}_0}\frac{\pi_p(s)}{\binom{p}{s}}\left(\frac{\lambda}{2}\right)^s\int_{(B\cap A)_S}e^{-(1/2)\|Y-X_S\beta_S\|_2^2}e^{-\lambda\|\beta^0\|_1}\,d\beta_S.$$

Then it suffices to show that $\mathbb{E}_{\beta^0}\|\Pi^{(1)}(\cdot|Y)-\Pi^{(2)}(\cdot|Y)\|_{\mathrm{TV}}\to 0$. (The factor $e^{-\lambda\|\beta^0\|_1}$ in the second formula cancels in the normalization, but is inserted to connect to the remainder of the proof.)

For any sequences of measures (μ_S) and (ν_S), we have

$$\left\| \frac{\sum_S \mu_S}{\|\sum_S \mu_S\|_{\mathrm{TV}}} - \frac{\sum_S \nu_S}{\|\sum_S \nu_S\|_{\mathrm{TV}}} \right\|_{\mathrm{TV}} \le \frac{2\sum_S \|\mu_S - \nu_S\|_{\mathrm{TV}}}{\|\sum_S \mu_S\|_{\mathrm{TV}}} \le 2\sup_S \left\| 1 - \frac{d\nu_S}{d\mu_S} \right\|_\infty$$

if ν_S is absolutely continuous with respect to μ_S with density $d\nu_S/d\mu_S$, for every S. It follows that

$$\begin{aligned} \|\Pi^{(1)}(\cdot|Y) - \Pi^{(2)}(\cdot|Y)\|_{\mathrm{TV}} &\le 2 \max_{S\in\mathcal{S}_0} \sup_{\beta\in A} \left| e^{\lambda\|\beta_S\|_1 - \lambda\|\beta^0\|_1} - 1 \right| \\ &\le 2 \max_{S\in\mathcal{S}_0} \sup_{\beta\in A} e^{\lambda\|\beta_S - \beta^0\|_1} \lambda \|\beta_S - \beta^0\|_1. \end{aligned}$$

This tends to zero by the definition of A and the assumptions on β^0.

Finally we show that $\Pi^\infty(A|Y) \to 1$. For $\Lambda_{n,\beta,\beta^0}$, the likelihood ratio given in (6.1), we have

$$\Pi^\infty(A^c|Y) = \frac{\int_{A^c} \Lambda_{n,\beta,\beta^0}(Y)\, dU(\beta)}{\int \Lambda_{n,\beta,\beta^0}(Y)\, dU(\beta)}$$

$$\text{for } dU(\beta) = \sum_{S\in\mathcal{S}_0} \frac{\pi_p(s)}{\binom{p}{s}} \left(\frac{\lambda}{2}\right)^s d\beta_S \otimes \delta_{S^c}.$$

By (6.1) the denominator in $\Pi^\infty(\cdot|Y)$ satisfies

$$\begin{aligned} &\int \Lambda_{n,\beta,\beta^0}(Y)\, dU(\beta) \\ &\quad \ge \frac{\pi_p(s_0)}{\binom{p}{s_0}} \left(\frac{\lambda}{2}\right)^{s_0} \int e^{-(1/2)\|X(\beta_{S_0} - \beta^0_{S_0})\|_2^2 + (Y - X\beta^0)^t X(\beta_{S_0} - \beta^0_{S_0})}\, d\beta_{S_0} \\ &\quad \ge \frac{\pi_p(s_0)}{\binom{p}{s_0}} \left(\frac{\lambda}{2}\right)^{s_0} \int e^{-(1/2)\|Xb_{S_0}\|_2^2}\, db_{S_0} = \frac{\pi_p(s_0)}{\binom{p}{s_0}} \left(\frac{\lambda}{2}\right)^{s_0} \frac{(2\pi)^{s_0/2}}{|\Gamma_{S_0}|^{1/2}}, \end{aligned}$$

where $\Gamma_S = X_S^t X_S$, and for the second inequality we use Jensen's inequality similarly as in the proof of Lemma 2.

Using Hölder's inequality $|\alpha^t\beta| \le \|\alpha\|_\infty \|\beta\|_1$, we see that on the event $\mathcal{T}_0 = \{\|X^t(Y - X\beta^0)\|_\infty \le \bar{\lambda}\}$,

$$\begin{aligned} (Y - X\beta^0)^t X(\beta - \beta^0) &\le \bar{\lambda}\|\beta - \beta^0\|_1 \\ &\le 2\frac{\bar{\lambda}\|X(\beta - \beta^0)\|_2 |S_{\beta-\beta^0}|^{1/2}}{\|X\|\bar{\phi}(|S_{\beta-\beta^0}|)} - \bar{\lambda}\|\beta - \beta^0\|_1 \\ &\le \frac{1}{2}\|X(\beta - \beta^0)\|_2^2 + \frac{2\bar{\lambda}^2 |S_{\beta-\beta^0}|}{\|X\|^2 \bar{\phi}(|S_{\beta-\beta^0}|)^2} - \bar{\lambda}\|\beta - \beta^0\|_1. \end{aligned}$$

Since $\overline{\lambda}(|S_{\beta-\beta^0}|) \geq \overline{\psi}(|S_0|)$ for every $S_\beta \in \mathcal{S}_0$, it follows that on $\mathcal{T}_0$ the numerator in $\Pi^\infty(A^c|Y)$ is bounded above by

$$e^{(2\overline{\lambda}^2|S_{\beta-\beta^0}|/(\|X\|^2\overline{\psi}(S_0)^2))-(\overline{\lambda}Ms_0\sqrt{\log p}/(2\|X\|\overline{\psi}(S_0)^2\phi(S_0)^2))}\int e^{-(1/2)\overline{\lambda}\|\beta-\beta^0\|_1}\,dU(\beta)$$
$$\leq e^{(8|S_{\beta-\beta^0}|\log p/(\overline{\psi}(S_0)^2))-(Ms_0\log p/(2\overline{\psi}(S_0)^2\phi(S_0)^2))}\sum_{s=0}^{p}\pi_p(s)4^s.$$

It follows that $\Pi^\infty(A^c|Y)$ is bounded above by

$$\frac{\binom{p}{s_0}}{\pi_p(s_0)}\left(\frac{2}{\lambda}\right)^{s_0}\frac{|\Gamma_{S_0}|^{1/2}}{(2\pi)^{s_0/2}}e^{(8|S_{\beta-\beta^0}|\log p/\overline{\psi}(S_0)^2)-(Ms_0\log p/(2\overline{\psi}(S_0)^2\phi(S_0)^2))}\sum_{s=0}^{p}\pi_p(s)4^s.$$

By Jensen's inequality applied to the logarithm $|\Gamma_S| \leq (s^{-1}\operatorname{tr}(\Gamma_S))^s \leq \|X\|^{2s}$, and hence $|\Gamma_S|^{1/2}/\lambda^s \leq p^s$, by (2.1). The prior mass $\pi_p(s)$ can be bounded below by powers of p^{-s} by (2.2). This shows that the display tends to zero for sufficiently large M. □

PROOF OF THEOREM 4. Let Σ be the collection of all sets $S \in \mathcal{S}_0$ such that $S \supset S_0$ and $S \neq S_0$. In view of Theorem 6 it suffices to show that $\Pi^\infty(\beta : S_\beta \in \Sigma|Y) \to 0$.

Note that due to $A_4 > 1$, any set in $S \in \mathcal{S}_0$ has cardinality smaller $6s_0$. By (2.7), with $\Gamma_S = X_S^tX_S$,

$$\Pi^\infty(\beta : S_\beta \in \Sigma|Y) \leq \sum_{S\in\Sigma}\widehat{w}_S$$
$$\leq \sum_{s=s_0+1}^{6s_0}\frac{\pi_p(s)\binom{p}{s_0}\binom{p-s_0}{s-s_0}}{\pi_p(s)\binom{p}{s}}\max_{\substack{S\in\Sigma,\\|S|=s}}\frac{|\Gamma_{S_0}|^{1/2}}{|\Gamma_S|^{1/2}}\left(\lambda\sqrt{\frac{\pi}{2}}\right)^{s-s_0}$$
$$\times e^{(1/2)\|X\widehat{\beta}_{(S)}\|_2^2-(1/2)\|X\widehat{\beta}_{(S_0)}\|_2^2}.$$

We shall show below that the factors on the right-hand side can be bounded as follows: for any fixed $r > 2$,

$$\lambda^{s-s_0}|\Gamma_{S_0}|^{1/2}|\Gamma_S|^{-1/2} \leq (4\sqrt{\log p})^{s-s_0}\widetilde{\psi}(S_0)^{s_0-s}, \tag{6.11}$$

$$\mathbb{P}\big(\|X_S\widehat{\beta}_{(S)}\|_2^2 - \|X_{S_0}\widehat{\beta}_{(S_0)}\|_2^2 \leq r(s-s_0)\log p,\ \text{for all } S \in \Sigma\big) \to 1. \tag{6.12}$$

Combining these estimates with assumption (2.2) shows that for $\mathcal{T}$, the event in the second relation,

$$\Pi^\infty(\beta : S_\beta \in \Sigma|Y)\mathbb{I}_{\mathcal{T}} \leq \sum_{s=s_0+1}^{6s_0}(A_1p^{-A_4})^{s-s_0}\binom{s}{s_0}\left(\frac{\sqrt{8\pi\log p}}{\widetilde{\psi}(S_0)}\right)^{s-s_0}p^{r(s-s_0)/2}.$$

For $s_0 \le p^a$ we have $\binom{s}{s_0} = \binom{s}{s-s_0} \le s^{s-s_0} \le (6p^a)^{s-s_0}$. Thus the expression tends to zero if $a - A_4 + r/2 < 0$. Since r can be chosen arbitrarily close to 2, this translates into $a < A_4 - 1$.

To prove bound (6.11), we apply the interlacing theorem to the principal submatrix Γ_{S_0} of Γ_S to see that $\lambda_j(\Gamma_{S_0}) \le \lambda_j(\Gamma_S)$, for $j = 1, \dots, s_0$, where $\lambda_1 \ge \lambda_2 \ge \cdots$ denote the eigenvalues in decreasing order, whence

$$\begin{aligned} |\Gamma_{S_0}| = \prod_{j=1}^{s_0} \lambda_j(\Gamma_{S_0}) \le \prod_{j=1}^{s_0} \lambda_j(\Gamma_S) &\le \lambda_{\min}(\Gamma_S)^{s_0-s} |\Gamma_S| \\ &\le (\widetilde{\phi}(|S|)\|X\|)^{2(s_0-s)} |\Gamma_S|. \end{aligned}$$

Assertion (6.11) follows upon combining this with (2.1).

To bound the probability of the event $\mathcal{T}$ in (6.12), we note that by the projection property of the least squares estimator, for $S \supset S_0$ the difference $\|X_S\widehat{\beta}_{(S)}\|_2^2 - \|X_{S_0}\widehat{\beta}_{(S_0)}\|_2^2$ is the square length of the projection of Y onto the orthocomplement of the range of X_{S_0} within the range of X_S, a subspace of dimension $s - s_0$. Because the mean $X\beta^0$ of $Y = X\beta^0 + \varepsilon$ is inside the smaller of these ranges, it cancels under the projection, and we may use the projection of the standard normal vector ε instead. Thus the square length possesses a chi-square distribution with $s - s_0$ degrees of freedom. There are $N = \binom{p-s_0}{s-s_0}$ models $S \in \Sigma$ that give rise to such a chi-square distribution. Since $\log N \le (s - s_0)\log p \vee 1$, we can apply Lemma 5 with $d = s - s_0$ to give that $\mathbb{P}(\mathcal{T}^c)$ is bounded above by $\sum_{s>s_0} \binom{p-s_0}{s-s_0}^{-(r-2)/4} e^{c(s-s_0)}$. This tends to zero as $p \to \infty$, due to $\binom{p-s_0}{s-s_0} \ge (p-s)^{s-s_0} \ge (p/2)^{s-s_0}$, where the last inequality follows from $s_0/p \le s_0\lambda/\|X\| \to 0$. □

LEMMA 5. *For every $r > 2$, there exists a constant c independent of $N \ge 2$ and $d \ge 1$ such that for any variables $Q_1, \dots, Q_N$ that are marginally $\chi^2(d)$ distributed*,

$$\mathbb{P}\Bigl(\max_{1\le i\le N} Q_i > r\log N\Bigr) \le \Bigl(\frac{1}{N}\Bigr)^{(r-2)/4} e^{cd}.$$

PROOF. By Markov's inequality, for any $u > 0$,

$$\mathbb{P}\Bigl(\max_{1\le i\le N} Q_i > r\log N\Bigr) \le e^{-ur\log N}\mathbb{E}\max_{1\le i\le N} e^{uQ_i} \le N^{-ur}N\sqrt{1-2u}^{-d}.$$

The results follows upon choosing $u = 1/4 + 1/(2r)$, giving $ur - 1 = (r-2)/4$ and $1 - 2u = 1/2 - 1/r$. □

PROOF OF THEOREM 5. *Proof of first two assertions*. Because $\|\beta_{S_0} - \beta^0\|_1 \le \|\beta - \beta^0\|_1$, the posterior probability of the set

$$\Bigl\{\beta : \|\beta_{S_0} - \beta^0\|_1 > \frac{M}{\overline{\psi}(S_0)^2} \frac{|S_0|\sqrt{\log p}}{\|X\|\phi(S_0)^2}\Bigr\}$$

tends to zero by Theorem 3. This implies the first assertion. The second assertion follows similarly from the second assertion of Theorem 3.

Proof of third assertion. First we prove that the largest coefficient in absolute value, say β_m^0, is selected by the posterior if this is above the threshold. By Theorem 6 it is enough to show that $\mathbb{E}_{\beta^0}\Pi^\infty(\beta : m \in S_\beta|Y) \to 1$. For any given set S with $m \notin S$, let $S_m := S \cup \{m\}$ and $s = |S|$. Then

$$\Pi^\infty(\beta : m \notin S_\beta|Y) = \sum_{S\in\mathcal{S}_0 : m\notin S} \widehat{w}_S.$$

We shall bound this further by showing that $\widehat{w}_S \ll \widehat{w}_{S_m}$, for every S in the sum. The quotient of these weights is equal to

$$\begin{aligned}\frac{\widehat{w}_{S_m}}{\widehat{w}_S} &= \lambda\sqrt{\frac{\pi}{2}}\frac{\pi_p(s+1)}{\pi_p(s)}\frac{\binom{p}{s}}{\binom{p}{s+1}}\frac{|\Gamma_S|^{1/2}}{|\Gamma_{S_m}|^{1/2}}e^{(1/2)\|X_{S_m}\widehat{\beta}_{(S_m)}\|_2^2-(1/2)\|X_S\widehat{\beta}_{(S)}\|_2^2} \\ &\gtrsim \lambda p^{-A_3}\frac{s+1}{p-s}\frac{|\Gamma_S|^{1/2}}{|\Gamma_{S_m}|^{1/2}}e^{(1/2)\|X_{S_m}\widehat{\beta}_{(S_m)}\|_2^2-(1/2)\|X_S\widehat{\beta}_{(S)}\|_2^2},\end{aligned}$$

in view of (2.2). By the interlacing theorem, the eigenvalues $\overline{\lambda}_i$ in increasing order of the matrices Γ_S and Γ_{S_m} satisfy $\overline{\lambda}_i(\Gamma_{S_m}) \le \overline{\lambda}_i(\Gamma_S) \le \overline{\lambda}_{i+1}(\Gamma_{S_m})$, for any $1 \le i \le s$. This implies that $|\Gamma_S|/|\Gamma_{S_m}| \ge \overline{\lambda}_{s+1}(\Gamma_{S_m})^{-1}$. Since $\|X\beta\|_2 \le \|X\|\|\beta\|_1 \le \sqrt{|S_\beta|}\|X\|\|\beta\|_2$, for any β, the largest eigenvalue $\overline{\lambda}_{s+1}(\Gamma_{S_m})$ is at most $(s+1)\|X\|^2$. Combining this with (2.1), we conclude that the preceding display is bounded below by

$$\begin{aligned}&\frac{\lambda}{\|X\|}p^{-A_3-1}e^{(1/2)\|X_{S_m}\widehat{\beta}_{(S_m)}\|_2^2-(1/2)\|X_S\widehat{\beta}_{(S)}\|_2^2} \\ &\qquad\ge p^{-A_3-2}e^{(1/2)\|X_{S_m}\widehat{\beta}_{(S_m)}\|_2^2-(1/2)\|X_S\widehat{\beta}_{(S)}\|_2^2}.\end{aligned}$$

By definition of the least squares estimator, the difference of the square norms in the exponent is the square length of the projection of $Y = X\beta^0 + \varepsilon$ onto the orthocomplement F_S of the range of X_S in the range of X_{S_m}, the one-dimensional space spanned by the vector $X_m - P_S X_m$, where P_S denotes the projection onto the range of X_S. If, with an abuse of notation, P_{F_S} is the projection onto F_S, then

$$\begin{aligned}\|X_{S_m}\widehat{\beta}_{(S_m)}\|_2^2 - \|X_S\widehat{\beta}_{(S)}\|_2^2 &= \|P_{F_S}Y\|_2^2 \ge \frac{1}{2}\|P_{F_S}X\beta^0\|_2^2 - \|P_{F_S}\varepsilon\|_2^2 \\ &= \frac{\langle X\beta^0, X_m - P_S X_m\rangle^2}{2\|X_m - P_S X_m\|_2^2} - \frac{\langle \varepsilon, X_m - P_S X_m\rangle^2}{\|X_m - P_S X_m\|_2^2}.\end{aligned}\tag{6.13}$$

We shall show that the first term on the right is large if $|\beta_m^0|$ is large, and the second is small with large probability.

We start by noting that for $j \notin S$ and any S,

$$\begin{aligned}\|P_SX_j\|_2^2 &= \langle X_j, X_S\Gamma_S^{-1}X_S^tX_j\rangle \le \frac{1}{\widetilde{\phi}(s)^2\|X\|^2}\|X_S^tX_j\|_2^2 \\ &= \frac{1}{\widetilde{\phi}(s)^2\|X\|^2}\sum_{i\in S}(X^tX)_{i,j}^2 \le \frac{s\,\mathrm{mc}(X)^2\|X\|^2}{\widetilde{\phi}(s)^2}.\end{aligned} \tag{6.14}$$

It follows from the definitions that $\widetilde{\phi}(1)\|X\| \le \|X_j\| \le \|X\|$, for every j. Combined, this shows that $\|X_j - P_SX_j\|_2 \ge 3\|X\|\widetilde{\phi}(1)/4$ if $\sqrt{s}\,\mathrm{mc}(X) \le \widetilde{\phi}(s)\widetilde{\phi}(1)/4$.

We write $X\beta^0 = X_m\beta_m^0 + X_{-m}\beta_{-m}^0$, for X_{-m} the matrix obtained by removing the column X_m from X, and split the first inner product in (6.13) in the two parts

$$\begin{aligned}|\langle X_m\beta_m^0, X_m - P_SX_m\rangle| &= |\beta_m^0|\|X_m - P_SX_m\|_2^2, \\ |\langle X_{-m}\beta_{-m}^0, X_m - P_SX_m\rangle| &= \left|\sum_{j\neq m}\beta_j^0\langle X_j - P_SX_j, X_m - P_SX_m\rangle\right| \\ &\le \sum_{j\neq m, j\notin S}|\beta_j^0|\bigl(\mathrm{mc}(X)\|X\|^2 + \|P_SX_j\|_2\|P_SX_m\|_2\bigr) \\ &\le s_0|\beta_m^0|\left(\mathrm{mc}(X)\|X\|^2 + \frac{s\,\mathrm{mc}(X)^2\|X\|^2}{\widetilde{\phi}(s)^2}\right),\end{aligned}$$

using that $X_j - P_SX_j = 0$ if $j \in S$, the definition of $\mathrm{mc}(X)$ to bound $\langle X_j, X_m\rangle$, the Cauchy–Schwarz inequality on $\langle P_SX_j, X_m\rangle = \langle P_SX_jzP_SX_m\rangle$ and (6.14). Putting the estimates together we find that for $(s_0 \vee s)\,\mathrm{mc}(X) \le \widetilde{\phi}(s)\widetilde{\phi}(1)/4$,

$$\|P_{F_S}X\beta^0\|_2 \ge |\beta_m^0|\|X\|\widetilde{\phi}(1)\tfrac{1}{4}.$$

We can split the random inner product in (6.13) in the two parts $\langle \varepsilon, X_m\rangle$ and $\langle \varepsilon, P_SX_m\rangle$. For $\sqrt{s}\,\mathrm{mc}(X) \le \widetilde{\phi}(s)\widetilde{\phi}(1)/2$,

$$\|P_{F_S}\varepsilon\|_2 \le \frac{|\langle \varepsilon, X_m\rangle|}{3\|X\|\widetilde{\phi}(1)/4} + \frac{|\langle \varepsilon, P_SX_m\rangle|}{3\|X\|\widetilde{\phi}(1)/4}.$$

Each variable $\langle \varepsilon, v\rangle$ is normally distributed with mean zero and variance $\|v\|_2^2$, for any $v \in \mathbb{R}^n$. When m varies over $1, \dots, p$ and S over all subsets of size s that do not contain m, there are p possible variables in the first term and $p\binom{p-1}{s}$ possible variables in the second. For $\widetilde{\phi}(s) \ge \widetilde{\psi}(S_0) \ge c_0$ the variances of the variables in the two terms are of the orders $1/c_0^2$ and $s\,\mathrm{mc}(X)^2/c_0^4$, respectively. Therefore the means of the two suprema are of the orders $\sqrt{\log p}$ and $\sqrt{\log\binom{p}{s}}s^{1/2}\,\mathrm{mc}(X) \le \sqrt{\log p}$, respectively, if $s\,\mathrm{mc}(X) \le 1$. With probability $O(p^{-\mu})$ these variables do not exceed a multiple of their means.

We conclude that for $(s_0 \vee s)\,\mathrm{mc}(X) \le \widetilde{\phi}(s)\widetilde{\phi}(1)/4$ and $\widetilde{\phi}(s) \ge c_0$, the left-hand side of (6.13) is, with probability tending to one, bounded below by

$\|X\|^2(\beta_m^0)^2c_0^2/16 - O(\log p)$, whence for $|\beta_m^0| \geq M\sqrt{\log p}/\|X\|$ for large M, uniformly in S, m,

$$\frac{\widehat{w}_{S_m}}{\widehat{w}_S} \geq p^{-A_3-2}e^{cM^2\log p} \geq p^{\mu},$$

for $\mu > 0$ as large as desired (depending on M) and c a suitable positive constant. So, with overwhelming probability,

$$\Pi^{\infty}(\beta : m \notin S_\beta|Y) \leq p^{-\mu} \sum_{S\in\mathcal{S}_0:m\in S} \widehat{w}_S \leq p^{-\mu}.$$

Thus $\mathbb{E}_{\beta^0}\Pi^{\infty}(m \notin S|Y) \to 0$ at the order $p^{-\mu}$.

Next, for β_{m_2} the second largest coefficient, we consider $\Pi^{\infty}(m_2 \notin S|m_1 \in S, Y)$. By reasoning similar to the preceding, we show that the index m_2 is included asymptotically, etc. □

Acknowledgments. We thank an Associate Editor and four referees for valuable comments. We are also grateful to Amandine Schreck for helpful discussions.

SUPPLEMENTARY MATERIAL

Bayesian linear regression with sparse priors (DOI: 10.1214/15-AOS1334SUPP; .pdf). In the supplement we state a Bernstein–von Mises type result for large lambda and give the remaining proofs.

REFERENCES

[1] ABRAMOVICH, F., BENJAMINI, Y., DONOHO, D. L. and JOHNSTONE, I. M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.* **34** 584–653. MR2281879

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

[3] ARIAS-CASTRO, E. and LOUNICI, K. (2014). Estimation and variable selection with exponential weights. *Electron. J. Stat.* **8** 328–354. MR3195119

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

[5] BELITSER, E. Personal communication.

[6] BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

[7] BOTTOLO, L., CHADEAU-HYAM, M., HASTIE, D. I., LANGLEY, S. R., PETRETTO, E., TIRET, L., TREGOUET, D. and RICHARDSON, S. (2011). Ess++: A C++ objected-oriented algorithm for Bayesian stochastic search model exploration. *Bioinformatics* **27** 587–588.

[8] BOTTOLO, L. and RICHARDSON, S. (2010). Evolutionary stochastic search for Bayesian model exploration. *Bayesian Anal.* **5** 583–618. MR2719668

[9] BÜHLMANN, P., RÜTIMANN, P., VAN DE GEER, S. and ZHANG, C.-H. (2013). Correlated variables in regression: Clustering and sparse estimation. *J. Statist. Plann. Inference* **143** 1835–1858. MR3095072

[10] BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data*: *Methods*, *Theory and Applications*. Springer, Heidelberg. MR2807761

[11] BUNEA, F., TSYBAKOV, A. and WEGKAMP, M. (2007). Sparsity oracle inequalities for the Lasso. *Electron*. *J*. *Stat*. **1** 169–194. MR2312149

[12] CAI, T. T. and JIANG, T. (2011). Limiting laws of coherence of random matrices with applications to testing covariance structure and construction of compressed sensing matrices. *Ann*. *Statist*. **39** 1496–1525. MR2850210

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

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

[15] CASTILLO, I., SCHMIDT-HIEBER, J. and VAN DER VAART, A. (2015). Supplement to "Bayesian linear regression with sparse priors." DOI:10.1214/15-AOS1334SUPP.

[16] 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

[17] DALALYAN, A. S. and TSYBAKOV, A. B. (2007). Aggregation by exponential weighting and sharp oracle inequalities. In *Learning Theory*. *Lecture Notes in Computer Science* **4539** 97–111. Springer, Berlin. MR2397581

[18] DELLAPORTAS, P., FORSTER, J. and NTZOUFRAS, I. (2002). On Bayesian model and variable selection using mcmc. *Stat*. *Comput*. **12** 27–36.

[19] DONOHO, D. L., ELAD, M. and TEMLYAKOV, V. N. (2006). Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans*. *Inform*. *Theory* **52** 6–18. MR2237332

[20] GEORGE, E. I. (2000). The variable selection problem. *J*. *Amer*. *Statist*. *Assoc*. **95** 1304–1308. MR1825282

[21] GEORGE, E. I. and FOSTER, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika* **87** 731–747. MR1813972

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

[23] GRIFFIN, J. E. and BROWN, P. J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Anal*. **5** 171–188. MR2596440

[24] HANS, C. (2009). Bayesian lasso regression. *Biometrika* **96** 835–845. MR2564494

[25] HANS, C., DOBRA, A. and WEST, M. (2007). Shotgun stochastic search for "large p" regression. *J*. *Amer*. *Statist*. *Assoc*. **102** 507–516. MR2370849

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

[27] JI, C. and SCHMIDLER, S. C. (2013). Adaptive Markov chain Monte Carlo for Bayesian variable selection. *J*. *Comput*. *Graph*. *Statist*. **22** 708–728. MR3173738

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

[29] KOLTCHINSKII, V., LOUNICI, K. and TSYBAKOV, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann*. *Statist*. **39** 2302–2329. MR2906869

[30] LEUNG, G. and BARRON, A. R. (2006). Information theory and mixing least-squares regressions. *IEEE Trans*. *Inform*. *Theory* **52** 3396–3410. MR2242356

[31] LI, F. and ZHANG, N. R. (2010). Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *J*. *Amer*. *Statist*. *Assoc*. **105** 1202–1214. MR2752615

[32] LOUNICI, K. (2008). Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electron*. *J*. *Stat*. **2** 90–102. MR2386087

[33] MARTIN, R., MESS, R. and WALKER, S. G. (2014). Empirical Bayes posterior concentration in sparse high-dimensional linear models. Available at arXiv:1406.7718.

[34] MITCHELL, T. J. and BEAUCHAMP, J. J. (1988). Bayesian variable selection in linear regression. *J. Amer. Statist. Assoc.* **83** 1023–1036. MR0997578

[35] RICHARDSON, S., BOTTOLO, L. and ROSENTHAL, J. S. (2011). Bayesian models for sparse regression analysis of high dimensional data. In *Bayesian Statistics* **9** 539–568. Oxford Univ. Press, Oxford. MR3204018

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

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

[38] ROČKOVÁ, V. and GEORGE, E. I. (2014). EMVS: The EM approach to Bayesian variable selection. *J. Amer. Statist. Assoc.* **109** 828–846. MR3223753

[39] SCHÄFER, C. and CHOPIN, N. (2013). Sequential Monte Carlo on large binary sampling spaces. *Stat. Comput.* **23** 163–184. MR3016936

[40] SCHRECK, A., FORT, G., LE CORFF, S. and MOULINES, E. (2013). A shrinkage-thresholding Metropolis adjusted Langevin algorithm for Bayesian variable selection. Available at arXiv:1312.5658.

[41] SCOTT, J. G. and BERGER, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Ann. Statist.* **38** 2587–2619. MR2722450

[42] SHI, M. and DUNSON, D. B. (2011). Bayesian variable selection via particle stochastic search. *Statist. Probab. Lett.* **81** 283–291. MR2764295

[43] STINGO, F. C. and VANNUCCI, M. (2011). Variable selection for discriminant analysis with Markov random field priors for the analysis of microarray data. *Bioinformatics* **27** 495–501.

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

[45] VAN DE GEER, S. and MURO, A. (2014). On higher order isotropy conditions and lower bounds for sparse quadratic forms. *Electron. J. Stat.* **8** 3031–3061. MR3301300

[46] VAN DE GEER, S. A. and BÜHLMANN, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.* **3** 1360–1392. MR2576316

[47] YOU, C., ORMEROD, J. T. and MÜLLER, S. (2014). On variational Bayes estimation and variational information criteria for linear regression models. *Aust. N. Z. J. Stat.* **56** 73–87. MR3200293

[48] YUAN, M. and LIN, Y. (2005). Efficient empirical Bayes variable selection and estimation in linear models. *J. Amer. Statist. Assoc.* **100** 1215–1225. MR2236436

[49] ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701

[50] 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

[51] ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449

I. CASTILLO
CNRS - LPMA
UNIVERSITÉS PARIS VI & VII
BÂTIMENT SOPHIE GERMAIN
75205 PARIS CEDEX 13
FRANCE
E-MAIL: ismael.castillo@math.cnrs.fr

J. SCHMIDT-HIEBER
A. VAN DER VAART
MATHEMATICAL INSTITUTE
LEIDEN UNIVERSITY
NIELS BOHRWEG 1
2333 CA LEIDEN
THE NETHERLANDS
E-MAIL: schmidthieberaj@math.leidenuniv.nl
avdvaart@math.leidenuniv.nl