Submitted to the Annals of Statistics
arXiv: math.PR/0000000

NEEDLES AND STRAW IN A HAYSTACK: POSTERIOR CONCENTRATION FOR POSSIBLY SPARSE SEQUENCES

By Ismaël Castillo * and Aad van der Vaart

We consider full Bayesian inference in the multivariate normal mean model in the situation that the mean vector is sparse. The prior distribution on the vector of means is constructed hierarchically by first choosing a collection of nonzero means and next a prior on the nonzero values. We consider the posterior distribution in the frequentist set-up that the observations are generated according to a fixed mean vector, and are interested in the posterior distribution of the number of nonzero components and the contraction of the posterior distribution to the true mean vector. We find various combinations of priors on the number of nonzero coefficients and on these coefficients that give desirable performance. We also find priors that give suboptimal convergence, for instance Gaussian priors on the nonzero coefficients. We illustrate the results by simulations.

1. Introduction. Suppose that we observe a vector $X = (X_1, \ldots, X_n)$ in $\mathbb{R}^n$ such that

$$X_i = \theta_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1.1}$$

for independent standard normal random variables ε_i and an unknown vector of means $\theta = (\theta_1, \ldots, \theta_n)$. We are interested in Bayesian inference on θ, in the situation that this vector is possibly *sparse*.

Non-Bayesian approaches to this problem have recently been considered by many authors. Golubev [13] obtained results for model selection methods and threshold estimators for the mean-squared risk. Birgé and Massart [4] treated the model within their general context of model selection by penalized least squares. Abramovich et al. in [1] studied the performance of the False Discovery Rate method. The earlier work by Donoho and Johnstone [10] can be viewed as studying the problem within an ℓ_r context. Many authors (see e.g. [3], [22], [21] and references cited there) have investigated the connection to the LASSO or similar methods.

Methods with a Bayesian connection were studied by George and Foster [12], Zhang [20], Johnstone and Silverman [16, 17], Abramovich, Grinshtein

*Work partly supported by a Postdoctoral fellowship from the VU University Amsterdam and ANR Grant Banhdits ANR-2010-BLAN-0113-03.

AMS 2000 subject classifications: Primary 62G05, 62G20

Keywords and phrases: Bayesian estimators, Sparsity, Gaussian sequence model, Mixture priors, Asymptotics, Contraction.

and Pensky [2], and Jiang and Zhang [15]. The papers [12] and [16] considered an empirical Bayes method, consisting of modelling the parameters $\theta_1, \ldots, \theta_n$ a-priori as independently drawn from a mixture of a Dirac measure at 0 and a continuous distribution, determining an appropriate mixing weight by the method of (restricted) marginal maximum likelihood, and finally employing the posterior median or mean. The second paper [2] motivated penalties, applied in a penalized minimum contrast scheme, by prior distributions on the parameters, and derived estimators for the number of nonzero θ_i and the θ_i itself. The first is a posterior mode, but the estimator for θ, called "Bayesian testimation", does not seem itself Bayesian. (In fact, the Gaussian prior for the non-zero parameters in [2] will be seen to perform suboptimally in our fully Bayesian set-up.) The papers [20] and [15] obtain sharp results on (nonparametric) empirical Bayes estimators.

Other related papers include [19], [6], [7], [14], [15], [5].

A penalized minimum contrast estimator can often be viewed as the mode of the posterior distribution, and it is helpful to interprete penalties accordingly. However, the Bayesian approach yields a full posterior distribution, which is a random probability distribution on the parameter space. It has both a location and a spread, and can be marginalized to give posterior distributions for any functions of the parameter vector of interest. It is this object that we study in this paper. Such full Bayesian inference was recently considered by Scott and Berger [18], who discussed various aspects not covered in the present paper, but no concentration results. One example of our results is that the beta-binomial priors in [18], combined with moderately to heavy tailed priors on the nonzero means, yield optimal recovery.

Sparsity can be defined in various ways. Perhaps the most natural definition is the class of *nearly black* vectors, defined as

$$\ell_0[p_n] = \{\theta \in \mathbb{R}^n : \ \#(1 \le i \le n : \theta_i \ne 0) \le p_n\}.$$

Here p_n is a given number, which in theoretical investigations is typically assumed to be $o(n)$, as $n \to \infty$. Sparsity may also mean that many means are small, but possibly not exactly zero. Definitions that make this precise use *strong* or *weak* ℓ_s*-balls*, typically for $s \in (0, 2)$. These are defined as, with $\theta_{[1]} \ge \theta_{[2]} \ge \cdots \ge \theta_{[n]}$ the nonincreasing permutation of the coordinates of $\theta = (\theta_1, \ldots, \theta_n)$,

$$\begin{aligned} \ell_s[p_n] &= \Big\{\theta \in \mathbb{R}^n : \ \frac{1}{n}\sum_{i=1}^n |\theta_i|^s \le \Big(\frac{p_n}{n}\Big)^s\Big\} \\ m_s[p_n] &= \Big\{\theta \in \mathbb{R}^n : \ \frac{1}{n}\max_{1\le i\le n} i|\theta_{[i]}|^s \le \Big(\frac{p_n}{n}\Big)^s\Big\}. \end{aligned}$$

Because the nonzero coefficients in $\ell_0[p_n]$ are not quantitatively restricted, there is no inclusion relationship between this space and the weak and strong balls, although results for the latter can be obtained by projecting them into $\ell_0[p_n]$. On the other hand, the inclusion $\ell_s[p_n] \subset m_s[p_n]$ holds for any $s > 0$.

The extent of the sparsity, measured by the constant p_n, is assumed unknown. Our Bayesian approach starts by putting a prior π_n on this number, a given probability measure on the set $\{0, 1, 2, \ldots, n\}$. Next we complete this to a prior on the set of all possible sequences $\theta = (\theta_1, \ldots, \theta_n)$ in $\mathbb{R}^n$, by given a draw p from π_n choosing a random subset $S \subset \{1, \ldots, n\}$ of cardinality p, and choosing the corresponding coordinates $(\theta_i : i \in S)$ from a density g_S on $\mathbb{R}^S$ and setting the remaining coordinates $(\theta_i : i \in S^c)$ equal to zero. Given this prior, Bayes' rule yields the posterior distribution of θ as usual. We investigate the properties of this posterior distribution, in its dependence on the priors on the dimension and on the nonzero coefficients, in the nonBayesian set-up where X follows (1.1) with θ equal to a fixed, "true" parameter θ_0.

If the true parameter vector θ_0 belongs to $\ell_0[p_n]$, then it is desirable that the posterior distribution concentrates most of its mass on nearly black vectors. One main result of the paper is that this is the case provided the prior probabilities $\pi_n\{p\}$ decrease exponentially fast with the dimension p.

The quality of the reconstruction of the full vector θ can be measured by various distances. A natural one is the Euclidean distance, with square

$$\|\theta - \theta'\|^2 = \sum_{i=1}^{n} (\theta_i - \theta_i')^2.$$

If the indices of the p_n nonzero coordinates of a vector in the model $\ell_0[p_n]$ were known a-priori, then the vector could be estimated with mean square error of the order p_n. In [11] it is shown that, as $n, p_n \to \infty$ with $p_n = o(n)$,

$$\inf_{\hat\theta} \sup_{\theta \in \ell_0[p_n]} P_{n,\theta} \|\hat\theta - \theta\|^2 = 2p_n \log(n/p_n)\big(1 + o(1)\big).$$

Here the infimum is taken over all estimators $\hat\theta = \hat\theta(X)$ and $P_{n,\theta}$ denotes taking the expectation under the assumption that X is $N_n(\theta, I)$-distributed. In other words, the square minimax rate over $\ell_0[p_n]$ is $p_n \log(n/p_n)$, meaning that the unknown identity of the nonzero means needs to lead only to a logarithmic loss.

The Bayesian approach is presumably adopted for the intuition provided by prior modelling, and is not necessarily directed at attaining minimax rates. However, for theoretical investigation it is natural to take the minimax rate as a benchmark, and it is of particular interest to know which

priors yield a posterior distribution that concentrates most of its mass on balls around θ_0 of square radius of order $p_n \log(p_n/n)$, or close relatives as $p_n(\log n)^r$ that loose (only) a logarithmic factor. A second main result of the paper is that the minimax rate is attained for many combinations of priors. It suffices that the priors π_n decrease exponentially with dimension, and give sufficient weight to the true level of sparsity: for some $c > 0$,

$$\pi_n(p_n) \gtrsim \exp\bigl(-cp_n \log(n/p_n)\bigr). \tag{1.2}$$

Furthermore, the priors on the nonzero coordinates should have tails that are not lighter than Laplace, and satisfy a number of other technical properties. If inequality (1.2) fails then the rate of contraction may be slower than minimax; we show that it is not slower than $\log(1/\pi_n(p_n))$. (The word "contraction" is in line with other literature on nonparametric Bayesian procedures; with the present choice of metrics (which grow with n) the rates actually increase to infinity.)

More generally, we consider reconstruction relative to the ℓ^q metric for $0 < q \leq 2$, defined (without qth-root) by

$$d_q(\theta, \theta') = \sum_{i=1}^{n} |\theta_i - \theta_i'|^q. \tag{1.3}$$

For $q < 2$ this "metric" is more sensitive to small variations in the coordinates than the square Euclidean metric, which is d_2. (For $q \leq 1$ the definition gives a true metric d_q; for $1 < q \leq 2$ it does not.) From [11] the minimax rate over $\ell_0[p_n]$ for d_q is known to be of the order

$$r_{n,q}^* = p_n \log^{q/2}(n/p_n). \tag{1.4}$$

We show that the posterior "contraction" rate attains this order under conditions as in the preceding paragraph, and more generally characterize the rate in terms of $\log(1/\pi_n(p_n))$.

Besides nearly black vectors we consider rates of contraction if θ_0 is in a weak ℓ_s-ball. The minimax rate over $m_s[p_n]$ relative to d_q is (see [10])

$$\mu_{n,s,q}^* = n\Big(\frac{p_n}{n}\Big)^s \log^{(q-s)/2}(n/p_n). \tag{1.5}$$

This is shown to be also the rate of posterior contraction under slightly stronger conditions on the priors than before: the prior on dimension must decrease slightly faster than exponential. Under the same conditions we also show that the posterior distribution has exponential concentration, and therefore contracts also in the stronger sense of (any, Euclidean) moments.

A summary of these results is that good priors for the dimension decrease at exponential or, perhaps better, slightly faster rate, and good priors on the nonzero means have tails that are heavier than Laplace. We also show that priors with lighter tails, such as the Gaussian, attain significantly lower contraction rates at true parameter vectors θ_0 that are not close to the origin.

The structure of the article is as follows. In Section 2 we state the main concentration results. A practical algorithm, simulations and some pictures are presented in Section 3. Proofs are gathered at the end of the paper and in the supplementary Appendix [9].

1.1. *Notation.* We denote by $a \wedge b$ and $a \vee b$ the minimum and maximum of two real numbers a, b, and write $a \lesssim b$ if $a \leq Cb$ for a universal constant C. The notation $\triangleq$ means "equal by definition to". We call *support* of a vector $\theta = (\theta_1, \ldots, \theta_n) \in \mathbb{R}^n$ the set of indices of nonzero coordinates, and denote this by $S_\theta = \{i \in \{1, \ldots, n\} : \theta_i \neq 0\}$. We set $\theta_S = (\theta_i : i \in S)$, and let $|S|$ be the cardinality of a set $S \subset \{1, \ldots, n\}$.

2. Main results. Throughout the paper we consider a prior Π_n on $\mathbb{R}^n$ constructed in three steps:

(P1) A *dimension* p is chosen according to a prior probability measure π_n on the set $\{0, 1, 2, \ldots, n\}$.
(P2) Given p a subset $S \subset \{1, \ldots, n\}$ of size $|S| = p$ is chosen uniformly at random from the $\binom{n}{p}$ subsets of size p.
(P3) Given (p, S) a vector $\theta_S = (\theta_i : i \in S)$ is chosen from a probability distribution with Lebesgue density g_S on $\mathbb{R}^p$ (if $p \geq 1$) and this is extended to $\theta \in \mathbb{R}^n$ by setting the remaining coordinates θ_{S^c} equal to 0.

For simplicity we use the same density g_S for every set of a given dimension $|S|$, and will denote this also by $g_{|S|}$. We also assume that the prior on dimension is positive, that is $\pi_n(p) > 0$ for any integer p.

Given the prior Π_n, Bayes' rule yields the *posterior distribution* $B \mapsto \Pi_n(B|\, X)$, the conditional distribution of θ given X if the conditional distribution of X given θ is taken equal to the normal distribution $N_n(\theta, I)$. The probability $\Pi_n(B|\, X)$ of a Borel set $B \subset \mathbb{R}^n$ under the posterior distribution

can be written

$$\frac{\displaystyle\sum_{p=0}^{n} \frac{\pi_n(p)}{\binom{n}{p}} \sum_{|S|=p} \int_{(\theta_S,0)\in B} \prod_{i\in S} \phi(X_i-\theta_i) \prod_{i\notin S} \phi(X_i)\, g_S(\theta_S)\, d\theta_S}{\displaystyle\sum_{p=0}^{n} \frac{\pi_n(p)}{\binom{n}{p}} \sum_{|S|=p} \int \prod_{i\in S} \phi(X_i-\theta_i) \prod_{i\notin S} \phi(X_i)\, g_S(\theta_S)\, d\theta_S}. \tag{2.1}$$

Here $(\theta_S, 0)$ is the vector in $\mathbb{R}^n$ formed by adding coordinates $\theta_i = 0$ to $\theta_S = (\theta_i : i \in S)$, at the positions left open by $S \subset \{1, \ldots, n\}$ (in the correct order of the coordinates and not at the end, as the notation suggests). This expression is somewhat unwieldy; we consider computation in Section 3. The posterior distribution can be marginalized to obtain the posterior distribution of an arbitrary transformation $f(\theta)$ by choosing the set B equal to $B = \{\theta : f(\theta) \in B_0\}$.

The posterior distribution is a random probability distribution on $\mathbb{R}^n$, which we study under the assumption that the vector $X = (X_1, \ldots, X_n)$ is distributed according to a multivariate normal distribution with mean vector θ_0 and covariance matrix the identity. We let $P_{n,\theta_0} T$ denote the expected value of a function $T = T(X)$ under this distribution.

We shall be interested in two aspects of the posterior distribution: its dimensionality and its ability to recover the mean vector θ. Because the conditions are simpler in the case that the nonzero coordinates are independent under the prior, in the first two results we assume that the densities g_S in (P3) are of product form. Concrete examples of priors as in (P1) and (P3) that satisfy the conditions imposed in the theorems are given in Section 2.5.

2.1. *Dimensionality.* In the context of $\ell_0[p_n]$-classes, we say that the prior π_n on dimension has *exponential decrease* if, for some constants $C > 0$ and $D < 1$,

$$\pi_n(p) \le D\pi_n(p-1), \quad p \ge Cp_n. \tag{2.2}$$

If the condition is also satisfied with $C = 0$, we say that the prior on dimension has *strict* exponential decrease.

THEOREM 2.1 (Dimension). *If π_n has exponential decrease* (2.2) *and g_S is a product of $|S|$ copies of a univariate density g, with mean zero and finite second moment, then there exists $M > 0$ such that, as $p_n, n \to \infty$,*

$$\sup_{\theta_0\in\ell_0[p_n]} P_{n,\theta_0} \Pi_n(\theta :\ |S_\theta| > Mp_n |\, X) \to 0.$$

For reasonable priors, we may hope that the posterior distribution spreads mass in the p_n-dimensional subspace that supports a true mean vector $\theta_0 \in \ell_0[p_n]$. The theorem shows that the posterior distribution "overshoots" this space by subspaces of dimension at most a multiple of p_n. Because the overshoot can have a random direction, this does not mean that the posterior distribution concentrates overall on a fixed Mp_n-dimensional subspace. The theorem shows that it concentrates along Mp_n-dimensional coordinate planes, but its support will be far from convex.

Obviously the posterior distribution will concentrate on low-dimensional subspaces if the higher dimensional spaces receive little mass under the prior π_n. By the theorem exponential decrease is sufficient. The next step is to show that exponential decrease is not too harsh: it is compatible with good reconstruction of the full mean vector θ. This then of course requires a lower bound on the prior mass given to the spaces of "correct" dimension; for instance (1.2).

2.2. *Recovery.* Good recovery requires also appropriate prior densities g_S on the nonzero coordinates. Because the statistical problem of recovering θ from a $N_p(\theta, I)$ distributed observation is equivariant in θ, we may hope that the location of the nonzero coordinates of θ_0 does not play a role in its recovery rate. The non-Bayesian procedures considered in for instance [13] indeed fulfill this expectation. However, a Bayesian procedure (with proper priors) necessarily favours certain regions of the parameter space. Depending on the choice of priors g_S in (P3) this may lead to a shrinkage effect, even in the "average" recovery of the parameter as $n \to \infty$, yielding suboptimal behaviour for true parameters θ_0 that are far from the origin. This shrinkage effect can be prevented by choosing priors g_S with sufficiently heavy tails.

Again we first consider the case of independent coordinates. In the following theorem we assume that g_S is a product of $|S|$ densities of the form e^h, for a function $h : \mathbb{R} \to \mathbb{R}$ satisfying

$$|h(x) - h(y)| \lesssim 1 + |x - y|, \qquad \forall x, y \in \mathbb{R}. \tag{2.3}$$

This covers all densities e^h with a uniformly Lipshitz function h, such as the Laplace and Student densities. (For the Student density the following theorem assumes more than 2 degrees of freedom to ensure also finiteness of the second moment). It also covers other smooth densities with polynomial tails, and densities of the form $c_\alpha e^{-|x|^\alpha}$ for some $\alpha \in (0, 1]$, which have a function h that is bounded in a neighbourhood of the origin and uniformly Lipschitz outside the neighbourhood. On the other hand the standard normal density is ruled out. In Theorem 2.8 we shall see that this indeed causes a shrinkage effect.

Recall definition (1.3) of the (square) distance d_q.

THEOREM 2.2 (Recovery). *If π_n has exponential decrease (2.2) and g_S is a product of $|S|$ univariate densities of the form e^h with mean zero and finite second moment and h satisfying (2.3), then for any $q \in (0,2]$, for r_n satisfying*

$$r_n^2 \geq \{p_n \log(n/p_n)\} \vee \log \frac{1}{\pi_n(p_n)}, \tag{2.4}$$

and sufficiently large M, as $p_n, n \to \infty$ such that $p_n/n \to 0$,

$$\sup_{\theta_0 \in \ell_0[p_n]} P_{n,\theta_0} \Pi_n\big(\theta : d_q(\theta,\theta_0) > M r_n^q p_n^{1-q/2} \,|\, X\big) \to 0.$$

For $q = 2$ the theorem refers to the square Euclidean distance d_2, and asserts that the posterior distribution contracts at the rate r_n^2, uniformly over $\ell_0[p_n]$. The first inequality in (2.4) says that this rate is (of course) not faster than the minimax rate $r_{n,2}^* = p_n \log(n/p_n)$. The second shows that it is also limited by the amount of prior mass $\pi_n(p_n)$ put on the true dimension. If this satisfies (1.2), then $\log(1/\pi_n(p_n)) \lesssim r_{n,2}^*$ and the rate r_n^2 is equal to the minimax rate.

Condition (1.2) for every p_n leaves a free margin of a $\log(n/p_n)$-term over just exponential decrease of the prior π_n. If the decrease is still faster than (1.2), then the rate of contraction may be slower. For instance, for $\pi_n(p) \asymp \exp(-p^\alpha)$, for some $\alpha > 1$, the rate for the square Euclidean distance given by the theorem is not better than p_n^α, which is much bigger than $r_{n,2}^*$. In contrast, for $\alpha = 1$ the theorem gives the minimax rate.

For $q \in (0,2)$ we can make similar remarks. The minimax rate $r_{n,q}^*$ over $\ell_0[p_n]$ for d_q is given in (1.4). Because

$$(r_{n,2}^*)^{q/2} p_n^{1-q/2} = r_{n,q}^*,$$

the theorem shows contraction of the posterior distribution relative to d_q at the minimax rate $r_{n,q}^*$ over $\ell_0[p_n]$ under the same conditions that it gives the minimax rate $r_{n.2}^*$ for d_2: (1.2) suffices. Furthermore, if there is less prior mass at p_n, then the rate of contraction will be slower.

In the case that $0 < q < 1$ the result is surprising at first when compared to the finding in [16] that the posterior *median*, or more generally so-called "strict-thresholding rules", attain the convergence rate $r_{n,q}^*$, but the posterior *mean* converges at a *strictly slower* rate (even when $\theta_0 = 0$; see Section 10 in [16] and the remark below). By the preceding theorem the *full* posterior distribution *does* contract at the optimal rate $r_{n,q}^*$, for any $0 < q < 2$. This

is true in particular for the case of binomial priors on dimension considered in [16] with the "best possible" (oracle) choice $\alpha_n = p_n/n$.

The slower convergence of the posterior mean relative to the contraction of the full posterior distribution is made possible by the fact that d_q-balls have astroid-type shapes for $0 < q < 1$, and differ significantly from their convex hull if n is large. The posterior mean, which is in the convex hull of the support of the posterior, can therefore be significantly farther in d_q-distance from θ_0 than the bulk of the distribution. By Theorem 2.1 only few coordinates outside the support of θ_0 are given non-zero values by the posterior. However, the corresponding indices are random and *on average* spread over $\{1, 2, \ldots, n\}$, which makes that the posterior mean at a fixed coordinate is typically non-zero. Adding up all small errors in ℓ^q typically gives a much higher total sum for $q < 1$ than for $q \geq 1$. In contrast the posterior median does not suffer from this averaging effect.

The posterior measure thus provides a unifying point of view on the considered objects. In this perspective for $0 < q < 1$ the posterior mean is a bad representation of the full posterior measure.

REMARK 2.3. From the arguments exposed in [16], it is not hard to check that the posterior mean generally fails to attain the minimax rate over $\ell_0[p_n]$ relative to d_q for $0 < q < 1$. Let us consider the case of $\ell_0[p_n]$ classes with $\theta_0 = 0$. With the notation of [16], the posterior mean $\tilde{\mu}(x, \alpha_n)$ with data $X_1 = x$ for the binomial prior on dimension with parameters (n, α_n) satisfies $|\tilde{\mu}(x, \alpha_n)| \geq C|x|\alpha_n$, by the same reasoning as in the last display of page 1647 in [16] (the weight parameter $\hat{w}$ is fixed here and equals α_n). Hence the ℓ^q-power loss $\sum_i P_{n,\theta_0}|\theta_{0,i} - \tilde{\mu}(X_i, \alpha_n)|^q$ when $\theta_0 = 0$ is bounded from below by a constant times $n\alpha_n^q$. Thus, even for the "oracle" parameter $\alpha_n = p_n/n$, this is much above the minimax risk for any $0 < q < 1$.

2.3. *Dependent priors.* The preceding theorems are also true for priors that render the coordinates θ_i dependent. In the remaining theorems we assume that the densities g_S in (P3) satisfy the conditions, for every $S' \subset S \subset \{1, \ldots, n\}$ and a universal constant c_1,

$$\log g_S(\theta) - \log g_S(\theta') \leq c_1|S| + \frac{1}{64}\|\theta - \theta'\|^2, \quad \forall\, \theta, \theta' \in \mathbb{R}^S, \tag{2.5}$$

$$\big|\log g_S(\theta) - \log g_{S'}(\pi_{S'}\theta)\big| \leq c_1|S| + \frac{1}{64}\|\pi_{S-S'}\theta\|^2, \quad \forall\, \theta \in \mathbb{R}^S. \tag{2.6}$$

Here $\pi_S : \mathbb{R}^n \to \mathbb{R}^S$ is the projection defined by $\pi_S\theta = \theta_S = (\theta_i : i \in S)$. (The constant 64 corresponds to the constant 32 in Lemma 5.1, but has no special significance and can be improved.)

For a partition $S = S_1 \cup S_2$, we denote by $\theta = (\theta_1, \theta_2)$ the corresponding partition of $\theta \in \mathbb{R}^S$ and by $g_{S_1,S_2}(\theta_1, \theta_2) = g_S(\theta)$ the corresponding density. In the next theorem we assume that there exist $C, m_1 > 0$ and, for any S_2, probability densities γ_{S_2} on $\mathbb{R}^{S_2}$ such that for any $S_1 \subset S_2^c$ and $\theta_2 \in \mathbb{R}^{S_2}$,

$$\sup_{\theta_1 \in \mathbb{R}^{S_1}} \frac{g_{S_1,S_2}(\theta_1, \theta_2)}{g_{S_1}(\theta_1)} \leq C m_1^{|S_1|+|S_2|} \gamma_{S_2}(\theta_2), \tag{2.7}$$

This condition expresses that the "mixing between the coordinates within a given subspace" is not too important.

Examples are given in Section 2.5.

THEOREM 2.4 (Recovery). *Suppose π_n has strict exponential decrease, that is satisfies* (2.2) *with $C = 0$ and some $D > 0$. The assertions of Theorems 2.1 and 2.2 are also true if the densities g_S are not product densities, but general densities with finite second moments that satisfy* (2.5)-(2.6)-(2.7) *with $Dm_1 < 1$, and m_1 the constant in* (2.7).

2.4. *Complexity priors.* The next results are designed for application to the particular priors π_n of the form, for positive constants a, b,

$$\pi_n(p) \propto e^{-ap \log(bn/p)}, \tag{2.8}$$

where $\propto$ stands for 'proportional to'. Because $e^{p \log(n/p)} \leq \binom{n}{p} \leq e^{p \log(ne/p)}$, this prior is inversely proportional to the number of models of size p, a quantity that could be viewed as the *model complexity* for a given dimension p. Thus this prior appears particularly suited to the purpose of "downweighting the complexity". Forgetting about the extra component g_S of the prior, we can also consider it an analogue of the penalty "$2p \log(n/p)$" used in model selection in this context by (e.g.) Birgé and Massart in [4]. Every particular model with support S of size $|S| = p$ receives prior probability bounded below and above by expressions of the type $e^{-a_1 p \log(b_1 n/p)}$ from this prior.

Because the *complexity prior* (2.8) has exponential decrease (2.2) when $b > 1+e$ and satisfies (1.2), Theorems 2.1 and 2.4 (or Theorem 2.2) show that the corresponding posterior distribution concentrates on low-dimensional spaces and attains the minimax rate of contraction over $\ell_0[p_n]$ relative to (any) d_q if combined with densities g_S satisfying the conditions of Theorem 2.4. The following theorem relaxes the condition on g_S and gives a more precise result on the contraction of the posterior measure.

The theorem applies more generally to priors on dimension satisfying the upper bound, for some $a, b > 0$, and every $p \in \{0, 1, \ldots, n\}$,

$$\pi_n(p) \lesssim e^{-ap \log(bn/p)}. \tag{2.9}$$

THEOREM 2.5 (Recovery). *If the densities g_S have finite second moments, satisfy* (2.5)-(2.6) *for some constant c_1, and the priors π_n satisfy* (2.9) *for some $a \geq 1$ and $b \geq e^{7+2c_1}$, then, for r_n satisfying* (2.4), *for any $1 \leq p_n \leq n$ and any $r \geq 1$,*

$$\sup_{\theta_0 \in \ell_0[p_n]} P_{n,\theta_0} \Pi_n(\theta : \|\theta - \theta_0\| > 45 r_n + 10 r | X) \lesssim e^{-r^2/10}.$$

Consistent with the preceding findings, the posterior distribution concentrates on Euclidean balls of radius of the order r_n around θ_0. In addition the theorem shows that its "tail" is sub-Gaussian, uniformly in n and uniformly over $\ell_0[p_n]$. As one consequence, for every $l \in \mathbb{N}$,

$$P_{n,\theta_0} \int \|\theta - \theta_0\|^l \, d\Pi_n(\theta | X) \lesssim r_n^l.$$

By Jensen's inequality, this in turn implies the following corollary.

COROLLARY 2.1 (Posterior mean). *Under the conditions of Thm. 2.5,*

$$\forall \, l \in \mathbb{N}, \qquad \sup_{\theta_0 \in \ell_0[p_n]} P_{n,\theta_0} \Big\| \int \theta \, d\Pi_n(\theta | X) - \theta_0 \Big\|^l \lesssim r_n^l.$$

The posterior mean $\int \theta \, d\Pi_n(\theta | X)$ as a point estimator of θ_0 has a risk of the order r_n, relative to every polynomial loss function. In particular, it is rate-minimax over $\ell_0[p_n]$ for the squared ℓ_2-risk.

The posterior coordinate-wise median considered in the simulation study below is another interesting functional of the posterior measure. Under the conditions of Theorem 2.5 and (2.8), the posterior coordinate-wise median is rate-minimax over $\ell_0[p_n]$, *for any* d_q-distance, $0 < q \leq 2$, see [9].

The theorem, with its explicit bound, is also the basis for results on the concentration of the posterior distribution when the true vector is in a weak $m_s[p_n]$-class. Results for the posterior mean and ℓ_2-risk can be obtained as above as a consequence.

THEOREM 2.6 (Recovery, weak class). *If the densities g_S have finite second moments, satisfy* (2.5)-(2.6) *for some constant c_1, and the priors π_n satisfy* (2.9) *for some $a \geq 1$ and $b \geq e^{7+2c_1}$, then, for r_n satisfying*

$$r_n^2 = \min_{1 \leq p \leq n} \Big[\frac{s n^{2/s}}{2-s} \Big(\frac{1}{p} \Big)^{2/s-1} \Big(\frac{p_n}{n} \Big)^2 \vee p \log \frac{n}{p} \vee \log \frac{1}{\pi_n(p)} \Big],$$

for any $1 \leq p_n \leq n$, $s \in (0,2)$ and any $r \geq 1$,

$$\sup_{\theta_0 \in m_s[p_n]} P_{n,\theta_0} \Pi_n(\theta : \|\theta - \theta_0\| > 80 r_n + 20 r | X) \lesssim e^{-r^2/10}.$$

For the "complexity prior" π_n given by (2.8) the third term $\log(1/\pi_n(p))$ in the minimum defining it is smaller than a multiple of the second term, and hence can be omitted. The minimum can then be determined by equating the first two terms, leading to

$$p_n^* \asymp n(p_n/n)^s/\log^{s/2}(n/p_n). \tag{2.10}$$

If $p_n^* \gtrsim 1$, then this value is eligible in the minimum, and the first and second terms evaluated at p_n^* are of the same order, given by

$$r_n^2 \asymp n\Big(\frac{p_n}{n}\Big)^s \log^{1-s/2}\frac{n}{p_n}.$$

This in fact is the minimax rate $\mu_{n,s,2}^*$ for the square Euclidean metric d_2 over the class $m_s[p_n]$ (see (1.5)). Thus the complexity priors combined with densities g_S satisfying (2.5)-(2.6) (in particular product densities satisfying (2.3)) yield contraction at the minimax rate over both the nearly black vectors $\ell_0[p_n]$ and the weak $m_s[p_n]$ classes. For priors on dimension that are significantly smaller than the complexity priors the third term in the minimum must be taken into account and the rate of contraction is smaller than minimax.

The condition $p_n^* \gtrsim 1$ is satisfied as soon as the sparsity coefficient p_n/n is not too small. If the signal is very sparse and has $p_n^* \ll 1$, then the minimum in the definition of r_n^2 is taken at $p \sim 1$, leading to a squared rate of the order $\log n$. This is within a constant of the rate achieved by hard thresholding in that case.

The previous result extends under slightly stronger conditions to d_q-distances with $q > s$. Furthermore, the following theorem shows that p_n^* is indeed an upper bound on the dimensionality of the posterior distribution. For simplicity we only state the result in the case of complexity priors. Recall that $\mu_{n,s,q}^*$, given in (1.5), denotes the minimax rate over the class $m_s[p_n]$ relative to d_q.

THEOREM 2.7 (Dimensionality, recovery, weak class). *Suppose the densities g_S have finite second moments, satisfy* (2.5)-(2.6)-(2.7)*, and π_n satisfies* (2.8) *for sufficiently large $a \geq 1$ and $b > e$. Then for any $s \in (0,2)$, any $q \in (s,2)$ and any (p_n) such that $p_n/n \to 0$ and p_n^* given by* (2.10) *is bounded away from 0, for a sufficiently large constant M,*

$$\sup_{\theta_0 \in m_s[p_n]} P_{n,\theta_0}\Pi_n\big(\theta : |S_\theta| > Mp_n^*|\, X\big) \to 0,$$
$$\sup_{\theta_0 \in m_s[p_n]} P_{n,\theta_0}\Pi_n\big(\theta : d_q(\theta,\theta_0) > M\mu_{n,s,q}^*|\, X\big) \to 0.$$

2.5. *Examples.* In this section we discuss examples of priors on dimension π_n and prior densities g_S on the nonzero coordinates that satisfy the conditions of the preceding theorems.

EXAMPLE 2.1 (Independent Dirac mixtures). Consider the prior on $\theta = (\theta_1, \ldots, \theta_n) \in \mathbb{R}^n$ corresponding to sampling the coordinates θ_i independently from a mixture $(1-\alpha)\delta_0 + \alpha g$ of a Dirac measure at 0 and a univariate density g, for a given $\alpha \in (0,1)$. The coordinates of θ are then independently zero with probability $1-\alpha$, and hence the dimension of the model is binomially distributed with parameters n and α. Furthermore, the nonzero coordinates are distributed according to the product of copies of g. Thus this prior fits in our set-up, with π_n the binomial(n,α)-distribution and g_S a product density.

For a fixed α the coordinates θ_i are independent, under both the prior and the posterior distribution. Furthermore, the posterior distribution of θ_i depends on X_i only.

This prior is considered in [12] and [16], in combination with a Gaussian or a heavy tailed density g, respectively. In the next section we show that Gaussian priors are deficient if the nonzero coordinates of the signal are large. The authors of [16] propose to use the coordinatewise posterior median (or another univariate point estimator) for estimating θ, with the weight parameter α set by a thresholded empirical Bayes method: the parameter is chosen equal to the maximum likelihood estimator of α based on the marginal distribution of X in the Bayesian set-up (i.e. with θ integrated out but with fixed α) subject to the constraint that the resulting posterior median (after plugging in $\hat\alpha$) given an observation in the interval $[-\sqrt{2\log n}, \sqrt{2\log n}]$ is zero. The authors show that the resulting point estimator works remarkably well, in a minimax sense, for various metrics and sparsity classes.

A natural Bayesian approach is to put a prior on α, which yields a mixture of binomials as a prior π_n on the dimension of the model. The independence of the coordinates θ_i is then lost. We discuss this prior further in the following example.

EXAMPLE 2.2 (Binomial and Beta-binomial priors). The binomial (n, α_n) distribution as the prior π_n on dimension gives an expected dimension of $n\alpha_n$. In the sparse setting a small value of α_n is therefore natural. If the sparsity parameter p_n were known, we could consider the choice $\alpha_n = p_n/n$; we shall refer to the corresponding law as *oracle binomial prior*.

Assume that $p_n \to \infty$ with $p_n/n \to 0$. The binomial prior has exponential decrease (2.2) if $\alpha_n \lesssim p_n/n$. The oracle binomial prior $\alpha_n \asymp p_n/n$ is at

the upper end of this range, and also satisfies (1.2), and thus yields the minimax rate of contraction. The choice $\alpha_n = 1/n$ yields $\log \pi_n(p_n)$ of the order $-p_n \log p_n$, and hence attains the minimax rate if p_n is of the order n^a, $a < 1$; for larger p_n it may miss the minimax rate by a logarithmic factor.

A natural Bayesian strategy is to view the unknown "sparsity" parameter α as a hyperparameter and put a prior on it. The classical choice is the Beta prior, leading to the hierachical scheme $\alpha \sim \text{Beta}(\kappa, \lambda)$ and $p|\,\alpha \sim \text{Binomial}(n, \alpha)$, which corresponds to the following prior on p,

$$\pi_n(p) = \binom{n}{p} \frac{B(\kappa + p, \lambda + n - p)}{B(\kappa, \lambda)} \propto \frac{\Gamma(\kappa + p)\Gamma(\lambda + n - p)}{p!(n-p)!}.$$

The mean dimension is $n\kappa/(\kappa + \lambda)$, which suggests to choose the hyper parameters of the Beta distribution so that $\kappa/(\kappa + \lambda)$ is in the range $(c/n, Cp_n/n)$. It is easy to verify that the prior has exponential decrease (2.2), with $C = 1$, if $(\kappa - 1)/p_n < D(\lambda - 1)/(n - p_n + 1) + D - 1$. This suggests to choose small κ and large λ, thus giving a small variance to the Beta distribution.

For $\kappa = 1$ and $\lambda = n+1$ we obtain $\pi_n(p) \propto \binom{2n-p}{n}$. Then $\pi_n(p)/\pi_n(p-1) = (n-p+1)/(2n-p+1)$, showing exponential decrease (2.2), with $D = 1/2$. By a binomial identity the norming constant is equal to $\binom{2n+1}{n}$, and hence

$$\pi_n(p) = \frac{(2n-p)(2n-p-1)\cdots(2n-p-n+1)}{(2n+1)2n\cdots(2n+1-n+1)} \geq \Big(1 - \frac{p+1}{n+2}\Big)^n.$$

For $p_n/n \to 0$, this gives $\pi_n(p_n) \gtrsim e^{-p_n(1+o(1))}$ and hence (1.2) is satisfied. More generally, we may choose $\kappa = 1$, $\lambda = \kappa_1 n + 1$, which leads to $\pi_n(p) \propto \binom{(\kappa_1+1)n-p}{\kappa_1 n}$. The priors given by $\pi_n(p) \propto \binom{2n-p}{n}^{\kappa_1}$, for some $\kappa_1 > 0$ are a further alternative.

Example 2.3 (Poisson priors and hierarchies). The Poisson(α) distribution truncated to $\{0, 1, \ldots, n\}$, yields priors satisfying

$$\pi_n(p) \propto \frac{e^{-\alpha}\alpha^p}{p!} \asymp C e^{-p\log(p/\alpha)} e^p \frac{1}{\sqrt{p}},$$

for $p \to \infty$, by Stirling's approximation. The mean is approximately α, suggesting α in the range $(1, cp_n)$. As $\pi_n(p)/\pi_n(p-1) = \alpha/p$, the prior has exponential decrease (2.2) for $p \geq \alpha/D$.

If we put an exponential (λ) hyperprior on α, then π_n transforms into a shifted geometric distribution (shifted -1 to have support starting at 0) with success probability $\lambda/(1+\lambda)$. A Gamma hyperprior yields a shifted negative binomial. For fixed hyper-hyper parameters both are of the form e^{-Cp} for some constant C, and hence have exponential decrease, and satisfy (1.2).

EXAMPLE 2.4 (Complexity prior). The prior $\pi_n(p) \propto \exp -ap\log(bn/p)$ has exponential decrease and satisfies (1.2). Theorems 2.5, 2.6 and 2.7 show that this prior also gives sparsity and minimax recovery of the parameter over weak ℓ_s-classes. Although our results do not show the opposite assertion that mere exponential decrease is not enough for minimaxity on weak classes (while together with 1.2 it is enough for minimaxity over $\ell_0[p_n]$), this might be a potential advantage of complexity priors over the binomial and Poisson-based priors discussed previously.

EXAMPLE 2.5 (Product prior). Densities g_S that are products of $|S|$ copies of a univariate density with finite second moment of the form $g = e^h$ for $h : \mathbb{R} \to \mathbb{R}$ a function that satisfies (2.3), satisfy (2.5)-(2.6) and (2.7). In this sense Theorem 2.4 is a generalization of Theorem 2.2.

To see this note that for a product density the function g_S takes the form $g_S(\theta) = \exp\{\sum_{i\in S} h(\theta_i)\}$. Hence if (2.3) holds with proportionality constant 1, then the left side of (2.5) is bounded in absolute value by

$$\sum_{i\in S} h(\theta_i) - h(\theta_i') \le |S| + \|\theta - \theta'\|_1 \le |S| + \sqrt{|S|}\,\|\theta - \theta'\| \le 5|S| + \frac{1}{64}\|\theta - \theta'\|^2.$$

Furthermore, the left side of (2.6) is bounded by

$$\sum_{i\in S-S'} |h(\theta_i)| \le |S - S'||h(0)| + \sum_{i\in S-S'} (1 + |\theta_i|) \lesssim |S - S'| + \sum_{i\in S-S'} |\theta_i|.$$

The L_1-norm of $(\theta_i : i \in S - S')$ can be bounded by a linear combination of $|S - S'|$ and the square L_2-norm, as before, and hence the whole expression is bounded by $C|S| + \|\pi_{S-S'}\theta\|^2/64$, for some constant C.

Because a product density g_S is a product of the marginal densities, the validity of condition (2.7) is clear.

EXAMPLE 2.6 (Weakly mixing priors). For $h : \mathbb{R} \to \mathbb{R}$ a function satisfying (2.3) so that e^h is integrable and $G : [0, \infty) \to \mathbb{R}$ a Lipschitz function that is bounded below, consider, for $\theta = (\theta_1, \ldots, \theta_p)$,

$$g_p(\theta) = a_p e^{\sum_{i=1}^p h(\theta_i) - G(\|\theta\|)}.$$

where a_p is the normalizing constant. An example is the prior, for $a > 0$,

$$g_p(\theta) \propto \frac{e^{-\|\theta\|_1}}{1 + a^2\|\theta\|^2}.$$

In the appendix [9] it is shown that priors of this form satisfy (2.5)-(2.6). Furthermore, it is shown that (2.7) is also satisfied, with $m_1 = (1+a)/(1-a)$

if $-h$ is the absolute value of the identity function and the Lipschitz constant a of G is strictly smaller than 1 (i.e. $\big|G(s) - G(t)\big| \leq a|s-t|$ for $a < 1$).

Thus any prior of this form combined with any prior on dimension that decreases exponentially such that $Dm_1 = D(1+a)/(1-a) < 1$, for D the constant in (2.2), gives recovery at the minimax rate over $\ell_0[p_n]$, by Theorem 2.4, and also over $\ell_s[p_n]$ if combined with a complexity prior on dimension satisfying the conditions of Theorems 2.6 and 2.7. For instance, the hierarchical binomial prior $\pi_n(p) \propto \binom{2n-p}{n}$ in Example 2.2 has $D = 1/2$ and hence $a < 1/3$ suffices for contraction over $\ell_0[p_n]$.

2.6. *Lower bounds.* Condition (2.3) (or (2.5)-(2.6)) on the priors g_S for the nonzero coefficients ensures that the posterior does not shrink to the center of the prior too much. In the next theorem we investigate the necessity of conditions of this type. The theorem shows that product priors with marginal densities proportional to $y \mapsto e^{-|y|^\alpha}$ for some $\alpha > 1$ lead to a slow contraction rate for large true vectors θ_0. We formulate this in an asymptotic setting with a sequence of true vectors, written as θ_0^n, tending to infinity. We denote by p_n the number of nonzero coordinates of θ_0^n.

The theorem applies in particular to the normal distribution. For this prior a problem (only) arises if the parameter vector θ_0^n tends to infinity faster than the optimal rate:

$$\|\theta_0^n\|^2 \gg p_n \log(n/p_n).$$

The posterior then puts no mass on balls of radius a multiple of $\|\theta_0^n\|$ around the true parameter. For "small" θ_0^n no problem occurs, because shrinkage to the origin is desirable in that case. However, if the true parameter satisfies $\|\theta_0^n\|^2 \lesssim p_n \log(n/p_n)$, then the estimator that is zero, irrespective of the observations, possesses mean square error of the order the minimax risk for the problem. Thus it is rather poor consolation that the Bayes procedure based on Gaussian priors performs well in this case, as it is no better than the "zero estimator". Gaussian priors really are problematic.

Product priors with marginal density proportional to $y \mapsto e^{-|y|^\alpha}$ give behaviour as the Gaussian prior for every $\alpha \geq 2$. For $\alpha \in (1,2)$ the result is slightly more complicated and involves the quantities

$$\rho_{0,\alpha}^n = \left(\frac{\|\theta_0^n\|_\alpha^\alpha}{\|\theta_0^n\|_2^2} \wedge 1 \right) \|\theta_0^n\|_\alpha \, p_n^{1/2-1/\alpha}, \tag{2.11}$$

where $\|\cdot\|_\alpha$ denotes the usual L_α-norm on $\mathbb{R}^n$ (i.e. $\|\theta\|_\alpha^\alpha = \sum_i |\theta_i|^\alpha$). The numbers $\rho_{0,\alpha}^n$ increase to infinity as θ_0^n tends to infinity at a sufficiently fast rate. For instance $\rho_{0,\alpha}^n$ is of the order $c_n^{\alpha-1} p_n^{1/2-1/\alpha}$ if $\alpha < 2$ and $\theta_0^n = c_n \bar{\theta}_0$

for scalars c_n and fixed vectors $\bar{\theta}_0$. The following theorem shows that if $\rho_{0,\alpha}^n$ increases to infinity faster than the optimal rate $\left(p_n \log(n/p_n)\right)^{1/2}$, then the posterior does not charge balls of radius a small multiple of $\rho_{0,\alpha}^n$.

THEOREM 2.8 (Heavy tails). *Assume that the densities g_S are products of S univariate densities proportional to $y \mapsto e^{-|y|^\alpha}$ and the prior π_n on dimension satisfies* (1.2) *for some $c > 0$.*

(i) If $\alpha \geq 2$ and $\|\theta_0^n\|^2/(p_n \log(n/p_n)) \to \infty$, then for sufficiently small $\eta > 0$, as $n \to \infty$,

$$P_{n,\theta_0^n} \Pi_n(\theta : \|\theta - \theta_0^n\| \leq \eta \, \|\theta_0^n\| \,|\, X^n) \to 0.$$

(ii) If $1 < \alpha < 2$ and $(\rho_{0,\alpha}^n)^2/(p_n \log(n/p_n)) \to \infty$, then for sufficiently small $\eta > 0$, as $n \to \infty$,

$$P_{n,\theta_0^n} \Pi_n(\theta : \|\theta - \theta_0^n\| \leq \eta \, \rho_{0,\alpha}^n \,|\, X^n) \to 0.$$

Theorem 2.8 shows problematic behaviour of the posterior distribution for signals with large energies $\|\theta_0^n\|$. Instead of using fixed priors on the coordinates, we could make them depend on the sample size, for instance Gaussian priors with variance $v_n \to \infty$, or uniform priors on intervals $[-K_n, K_n]$ with $K_n \to \infty$. Such priors will push the "problematic boundary" towards infinity, but the same reasoning as for the theorem will show that shrinkage remains for (very) large θ_0^n.

The above results show that g_S needs to have heavy tails. Another important condition, this time concerning the prior π_n on the dimension k, concerns the amount of mass $\pi_n(p_n)$ at the true dimension. If this quantity is too small, then the Bayes procedure might not be optimal.

THEOREM 2.9. *Suppose also that the prior π_n on dimension in (P1) is decreasing and that there exist integers $d_{1,n} < d_{2,n}$ such that, for some $C > 0$ and a sequence $\underline{\varepsilon_n}$ such that $n\underline{\varepsilon_n}^2 \to \infty$,*

$$\frac{\pi_n(d_{2,n})}{\pi_n(d_{1,n})} \binom{n}{d_{1,n}} \leq e^{-Cn\underline{\varepsilon_n}^2}.$$

Denoting $d_{3,n} = (3d_{2,n} - d_{1,n})/2$, there exists θ_0 in $\ell_0[d_{3,n}]$ such that, for sufficiently small $\eta > 0$, as $n \to \infty$,

$$P_{n,\theta_0^n} \Pi_n(\theta : \|\theta - \theta_0^n\| \leq \eta\sqrt{n}\underline{\varepsilon_n} \,|\, X^n) \to 0.$$

EXAMPLE 2.7 (Prior on dimension in $\exp(-k(\log k)^a)$, with $a \geq 1$). If $\pi_n(k) = r\exp(-k\log^a k)$, with r the appropriate normalizing constant, let us apply the preceding result with the choices $d_{1,n} = p_n/4$, $d_{2,n} = 3p_n/4$, for some sequence $p_n \to \infty$. It holds

$$\begin{aligned}\frac{\pi_n(3p_n/4)}{\pi_n(p_n/4)}\binom{n}{p_n/4} &\leq e^{-\frac{3p_n}{4}\log^a\left(\frac{3p_n}{4}\right)+\frac{p_n}{4}\log^a\left(\frac{p_n}{4}\right)+\frac{p_n}{4}\log(ne)} \\ &\leq e^{-\frac{p_n}{4}\log^a\left(\frac{3p_n}{4}\right)-\frac{p_n}{4}\log^a\left(\frac{3p_n}{4}\right)^{2^{1/a}}+\frac{p_n}{4}\log^a(ne)}\end{aligned}$$

As long as we impose $(3p_n/4)^{2^{1/a}} \geq ne$ and $\log(3p_n/4) \geq 2^{-1/a}\log p_n$ (which holds for large enough n), the last display is at most $\exp(-\frac{p_n}{8}\log^a p_n)$. Theorem 2.9 implies that there is a vector θ_0 in $\ell_0[p_n]$ with

$$P_{n,\theta_0^n}\Pi_n(\theta : \|\theta - \theta_0^n\|^2 \leq \eta p_n \log^a p_n |\, X^n) \to 0,$$

for a small enough constant η. This implies that the corresponding estimator does not reach the optimal rate over the class $\ell_0[p_n]$ as soon as $p_n\log^a p_n$ tends to infinity faster than $p_n\log(n/p_n)$ (take for instance $p_n = n/\exp(\sqrt{\log n})$).

2.7. *Discussion.* We have identified general conditions on the prior that ensure optimal convergence rates for estimating a sparse mean vector in Gaussian noise. In particular, natural fully Bayes estimates (e.g. Beta-binomial prior on dimension) are shown to be adaptive with respect to the unknown smoothing parameter p_n/n.

Especially in high dimensional contexts the full posterior measure and special aspects of it can start to have divergent behaviors. We have seen that for non-convex distances the posterior mean is not a satisfactory projection. It can also happen that the mode and the full posterior behave differently.

In some situations one might want to estimate prior hyperparameters and in this case it is desirable to assess the convergence properties of the resulting plug-ins. To our knowledge, there are only a few works in this direction, see [16], [15]. Potential alternative proofs could consist in obtaining first (suitably uniform) results for the (full) posterior measure and combine them with statement saying that "the plug-in estimate is not too bad". Also, here, one could evaluate the sparsity coefficient $\eta_n = p_n/n$ via the posterior number $\hat{k}_n$ of selected models and plug this estimate in the full posterior for the binominal prior on dimension. Since $\hat{\eta}_n = \hat{k}_n/n$ does not exceed Cp_n with high probability, we have some control of the plug-in into the full posterior. The question of then deriving results for estimates of it (e.g. the mean), remains open.

3. Implementation. In this section we provide an algorithm to compute several functionals of the posterior measure associated to the prior defined by (P1)-(P3), including the posterior mean, marginal posterior quantiles, and the posterior of the number of selected models. The algorithm is exact in that it does not rely on an approximation of the posterior distribution, but computes the exact expressions. We illustrate the posterior quantities through simulations.

We assume that the densities g_S on $\mathbb{R}^S$ are products of S copies of a univariate density g. Because the prior on the number of nonzero coordinates induces dependence, this generally does not entail a factorization of the posterior distribution as a product measure. (An exception is the binomial distribution for π_n).

For all computations we need the denominator of the posterior measure in (2.1) (the "partition function"). For ϕ the standard normal density, and $\psi = \phi * g$ its convolution with the density g, this can be written

$$Q_n := \sum_{p=0}^{n} \frac{\pi_n(p)}{\binom{n}{p}} \sum_{|S|=p} \prod_{i \in S} \psi(X_i) \prod_{i \notin S} \phi(X_i).$$

Naive computation directly from this expression would require a number of operations that grows exponentially with n. However, the sum over all models S of size p (the inner sum in the display) is equal to the coefficient of Z^p in the polynomial

$$Z \mapsto \prod_{i=1}^{n} (\phi(X_i) + \psi(X_i) Z).$$

This polynomial can be computed by a quadratic number of operations by computing the products term by term, and in $n \log^2 n$ operations by a more clever algorithm.

3.1. *Posterior mean.* The posterior mean $\hat{\theta}^{PM} = \int \theta \, d\Pi_n(\theta | \, X)$ is a random vector in $\mathbb{R}^n$. Letting $\zeta(x) = \int t \phi(x - t) g(t) \, dt$, we can write its first coordinate in the form

$$\hat{\theta}_1^{PM} = \frac{1}{Q_n} \sum_{p=1}^{n} \frac{\pi_n(p)}{\binom{n}{p}} \zeta(X_1) \sum_{|S|=p,\, 1 \in S} \prod_{i \in S,\, i \neq 1} \psi(X_i) \prod_{i \notin S} \phi(X_i).$$

The inner sum (over S) is the coefficient of Z^p in the polynomial $Z \mapsto \zeta(X_1) Z \prod_{i=2}^{n} (\phi(X_i) + \psi(X_i) Z)$. Hence it can be computed as before.

3.2. *Coordinatewise quantiles.* The distribution function of the marginal posterior distribution of the first coordinate can be written, for any real u,

$$\Pi((-\infty, u] \times \mathbb{R}^{n-1} | \, X) = (1 - q_{n,1}) 1_{u \geq 0} + q_{n,1} \frac{\psi(X_1, u)}{\psi(X_1)},$$

where $1 - q_{1,n}$ is the posterior probability that the first coordinate is zero, and $\psi(x, u) = \int_{-\infty}^{u} \phi(x - t) g(t) \, dt$. The former probability can be written

$$1 - q_{n,1} = \Pr(\theta_1 = 0 | \, X) = \frac{1}{Q_n} \sum_{p=0}^{n} \frac{\pi_n(p)}{\binom{n}{p}} \sum_{|S|=p,\, 1 \notin S} \prod_{i \in S} \psi(X_i) \prod_{i \notin S} \phi(X_i).$$

Hence it can be computed as before, now involving the polynomial $Z \mapsto \psi(X_1) Z \prod_{i=2}^{n} (\phi(X_i) + \psi(X_i) Z)$.

Given the marginal posterior distribution, we can compute marginal quantiles. For instance, the first component of the coordinatewise median $\hat{\theta}^{med}$ is given by, with $H_{n,1}^{-1}$ the inverse of $H_{n,1}(u) = \psi(X_1, u)/\psi(X_1)$,

$$\hat{\theta}_1^{med} = \left[H_{1,n}^{-1}\Big(1 - \frac{1}{2q_{1,n}}\Big) \vee 0 \right] + \left[H_{n,1}^{-1}\Big(\frac{1}{2q_{n,1}}\Big) \wedge 0 \right].$$

The last display should be understood with the convention $H_{n,1}^{-1}(u) = -\infty$ if $u \leq 0$ and $H_{n,1}^{-1}(u) = \infty$ if $u \geq 1$.

3.3. *Number of nonzero coordinates.* The posterior distribution of the number $|S_\theta|$ of nonzero coordinates of $\theta \in \mathbb{R}^n$ is the random distribution on the set $\{0, 1, \ldots, n\}$ given by

$$\Pi_n(\theta : |S_\theta| = p | \, X) = \frac{1}{Q_n} \frac{\pi_n(p)}{\binom{n}{p}} \sum_{|S|=p} \prod_{i \in S} \psi(X_i) \prod_{i \notin S} \phi(X_i).$$

The same computational scheme applies. In fact the sum will already be computed in the derivation of Q_n.

3.4. *Simulations.* In a small simulation study we considered the prior defined by (P1)-(P3) with g a Laplace density $x \to (a/2)e^{-a|x|}$, with scale parameter $a > 0$, and two priors on dimension, suggested by our theoretical results, given by

$$\pi_n(p) \propto e^{-\kappa p \log(3n/p)}, \tag{3.1}$$

$$\pi_n(p) \propto \binom{2n - p}{n}^{\kappa}. \tag{3.2}$$

Here κ is a real parameter, which for both priors quantifies how fast they decrease to zero with p. In the results shown we used $a = 1$ and $\kappa \in \{0.1, 1\}$.

We simulated signals $\theta = (\theta_1, \ldots, \theta_n)$ of length $n = 500$, for various settings of the sparsity $p_n = \#(\theta_i \neq 0)$ and for signals θ with the nonzero coordinates set equal to a fixed number A. We show the results for $p_n \in \{25, 50, 100\}$ and "signal strength" $A \in \{3, 4, 5\}$.

Tables 1 and 2 report estimates of the mean square errors $\mathrm{E}\,\|\hat\theta - \theta\|^2$ and mean absolute deviation errors $\mathrm{E}\,\|\hat\theta - \theta\|_1$ of eight estimators $\hat\theta$. These estimates are the average (square) error of 100 estimates $\hat\theta_1, \ldots, \hat\theta_{100}$ computed from 100 data vectors simulated independently from model (1.1). The eight estimators include the posterior means *PM1*, *PM2* and coordinatewise medians *PMed1*, *PMed2* associated to the two priors π_n with $\kappa = 0.1$, the empirical Bayes mean *EBM* and median *EBMed* considered in [16] with a standard Laplace prior, and the hard-thresholding *HT* and hard-thresholding-oracle *HTO* estimators, given by

$$\hat\theta_i^{HT} = X_i 1_{|X_i| > \sqrt{2\log n}}, \qquad \hat\theta_i^{HTO} = X_i 1_{|X_i| > \sqrt{2\log n/p_n}}.$$

The last estimator uses the "oracle" value of the sparsity parameter p_n, whereas the other seven estimators do not use this value.

The tables show that the mean and median of the full Bayesian posterior distribution are competitive with the empirical Bayes estimates. The behaviour of the full Bayes and empirical Bayes estimates seems similar, up to a few aspects. In terms of squared risk, empirical Bayes estimates appear to be slightly better for small p_n and small A, while the full Bayes estimates appear to be slightly better for larger signals and larger p_n. For L^1-risk, the full Bayes estimates appear to outperform the EB-estimates in most of the cases. (Additional simulation results, not shown, suggest that the situation becomes less unfavourable for empirical Bayes as the scale parameter a of the Laplace prior is taken smaller than 1). In agreement with [16], in most cases the mean estimates perform not quite as well as the median ones, already in terms of squared-risk.

The parameter a of the Laplace prior plays the same role for the full Bayes as for the empirical Bayes estimates. Although we do not investigate this aspect here, it could be estimated from the data, as is proposed in the EbayesThresh-package, or be treated as a hyperparameter in a full Bayes approach. (A single scale parameter for high-dimensional densities g_S appears to create dependence between the coordinates stronger than allowed by our conditions (2.5)-(2.6), and hence would need further analysis.) Similar remarks pertain to the parameter κ. The choice $\kappa = 0.1$ seemed to be fairly good uniformly over all considered simulations, also for smaller n's.

p_n	**25**			**50**			**100**		
A	**3**	**4**	**5**	**3**	**4**	**5**	**3**	**4**	**5**
PM1	111	96	94	176	165	154	267	302	307
PM2	**106**	92	82	169	165	152	269	280	274
EBM	**103**	96	93	166	177	174	271	312	319
PMed1	129	**83**	73	205	149	**130**	**255**	279	283
PMed2	125	86	**68**	187	148	**129**	273	**254**	**245**
EBMed	110	**81**	72	**162**	148	142	**255**	294	300
HT	175	142	**70**	339	284	135	676	564	252
HTO	136	92	84	206	159	139	306	261	245

TABLE 1

Average square errors of eight estimators computed on 100 data vectors X of length $n = 500$ simulated from model (1.1) with $\theta = (0, 0, \ldots, 0, A, \ldots, A)$, where p_n coordinates indices are equal to A. In every column the smallest value is printed in bold face. The estimators are: PM1*,* PM2*: posterior means for two priors π_n in* (3.1)-(3.2) *and Laplace prior on nonzero coordinates;* PMed1*,* PMed2 *coordinatewise medians for the same priors;* EBM*,* EBMed*: empirical Bayes mean and median for Laplace prior;* HT*,* HTO*: hard-thresholding and hard-thresholding-oracle.*

For further illustration Figure 1 shows marginal 95% credible intervals (orange bars) for the parameters $\theta_1, \ldots, \theta_n$, and marginal posterior medians (red dots) for a single simulation of the data vector, with single strength $A = 5$, $p_n = 100$ and $n = 500$. The observations $X_1, \ldots, X_n$ are indicated by green dots. The credible intervals are defined as intervals between the 2.5% and 97.5% percent quantiles of the marginal posterior distributions of the parameters. The intervals corresponding to zero and nonzero coefficients θ_i are clearly separated, although some of the credible intervals of nonzero θ_i contain the value zero. Also visible is that the posterior medians and the credible intervals surrounding them are shrunk towards zero relative to the observed value X_i, for the zero coordinates θ_i, which is desirable, but also for the nonzero θ_1. Figure 1 (bottom) shows that for $\kappa = 1$ the shrinkage effects are stronger, and the credible intervals shorter.

Since our main goal here is illustration, we only implemented a simple

p_n	**25**			**50**			**100**		
A	**3**	**4**	**5**	**3**	**4**	**5**	**3**	**4**	**5**
PM1	80	101	110	127	145	147	240	268	270
PM2	79	85	87	135	145	144	219	232	232
EBM	95	110	117	191	200	176	260	285	281
PMed1	**51**	43	45	**86**	**80**	78	178	225	230
PMed2	**50**	**40**	37	**86**	**79**	76	**156**	**162**	163
EBMed	**50**	48	45	108	121	97	212	258	257
HT	63	44	**27**	122	86	**53**	244	173	**102**
HTO	53	41	40	91	79	74	157	148	144

TABLE 2

Average absolute deviation errors of eight estimators computed on 100 data vectors X of length $n = 500$ simulated from model (1.1) with $\theta = (0, 0, \ldots, 0, A, \ldots, A)$, where p_n coordinates indices are equal to A. In every column the smallest value is printed in bold face. The priors and estimators are as in Table 1.

version of the algorithm. This computes the polynomials with direct loops and can be improved. This implementation is limited to n of the order 500, not by computing time, but by the appearance of large numbers in the polynomial coefficients that overflow standard memory capacity $(10^{-300}, 10^{300})$. Handling larger n should certainly be possible by improved programming, for instance by computing on a logarithmic scale. Algorithmic complexity appears not to be a major issue.

4. Proof of Theorem 2.1. We first prove the theorem for priors on dimension $\pi_n(p)$ with strict exponential decrease and densities g_S that are not necessarily of product form, but that satisfy (2.7), for $Dm_1 < 1$, and D the constant in (2.2). Thus the proof also covers half of Theorem 2.4. In view of Example 2.5, densities of the product form satisfy (2.7) with $m_1 = 1$, and hence automatically have $Dm_1 < 1$.

Since the true parameter θ_0 is assumed to have p_n nonzero coordinates,

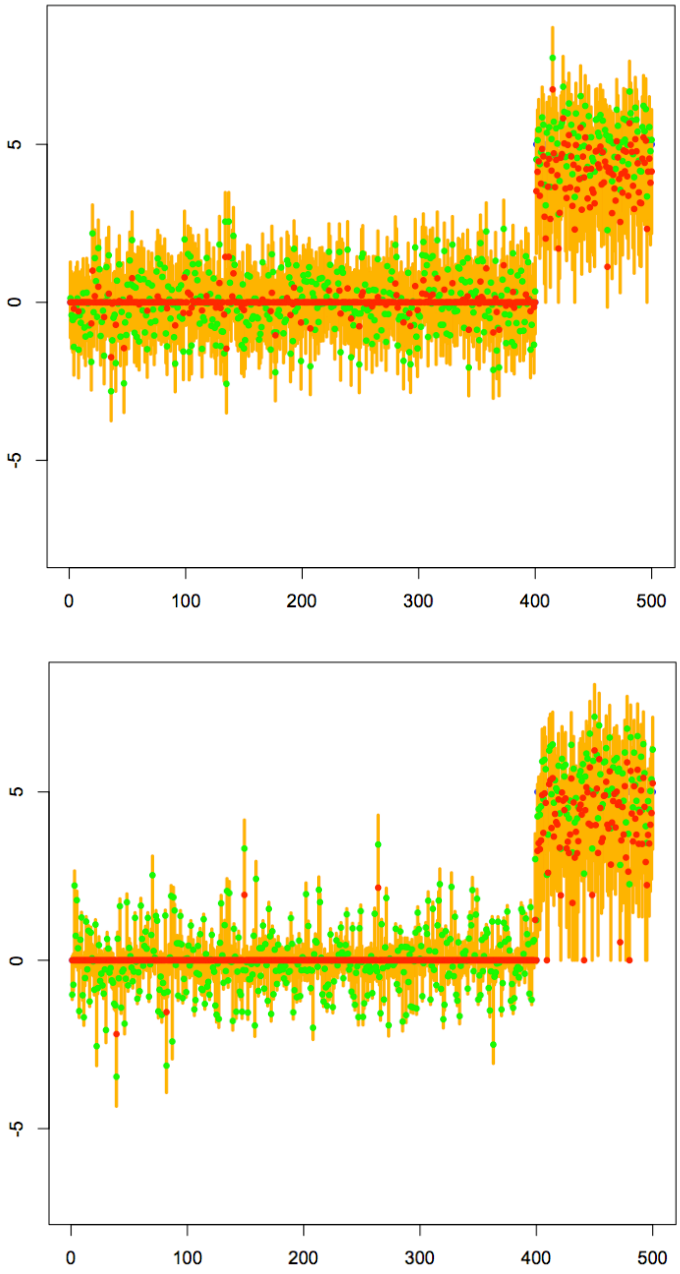

FIG 1. *Marginal posterior medians (red dots) and marginal credible intervals (orange) for the parameters* $\theta_1, \ldots, \theta_n$ *for a single data vector* $X_1, \ldots, X_n$ *simulated according to the model* (1.1) *with* $\theta = (0, 0, \ldots, 0, 5, \ldots, 5)$, *where* $n = 500$ *and the last* $p_n = 100$ *coordinates are nonzero. The data points are indicated by green dots. The prior* g *is the standard Laplace density and* π_n *is as in* (3.2) *with "inverse temperature"* $\kappa_1 = 0.1$ *(TOP graph) and* $\kappa_1 = 1$ *(BOTTOM graph).*

it is sufficient to prove that the intersection of the support S_θ with the complement S_0^c of the support $S_0 \triangleq S_{\theta_0}$ of θ_0 has dimension of the order p_n under the posterior distribution. The following proposition gives an explicit bound on this dimension; it is followed by a lemma that shows that this bound tends to zero under the conditions of the theorems. The idea of the proof of the proposition is to condition on the vector of the coordinates $\pi_{S_0}\theta$ of θ that belong to S_0.

The unconditional density of (S_θ, θ) for θ drawn from the prior Π_n is given by, with δ_0 denoting a "Dirac density at 0",

$$(S, \theta) \mapsto \frac{\pi_n(|S|)}{\binom{n}{|S|}} g_S(\theta_S)\delta_0(\theta_{S^c}).$$

The conditional density of $(S_\theta \cap S_0^c, \theta_{S_0^c})$ given θ_{S_0} is proportional to this expression viewed as function of $(S \cap S_0^c, \theta_{S\cap S_0^c})$. This shows the conditional distribution has the same structure as the prior Π_n, but with sample space $\mathbb{R}^{S_0^c}$ rather than $\mathbb{R}^n$, with the density of the nonzero coordinates of $\theta_{S_0^c}$ given by $g_{S\cap S_0^c|S\cap S_0}(\cdot|\,\theta_{S\cap S_0})$, and the prior on dimension given by,

$$\pi_{n,k}(p) \propto \pi_n(p+k)\frac{\binom{n-p_n}{p}}{\binom{n}{p+k}}, \qquad k = |S_\theta \cap S_0|. \tag{4.1}$$

The extra factor (quotient) on the right arises, because $\pi_{n,k}(p)$ and $\pi_n(p+k)$ are the probabilities of the given dimensions and hence the sums of the probabilities of all subsets of that dimension. Recall also that we assume that $\pi_n(p)$ is positive for any p, which makes the maximum appearing in the following proposition always finite.

PROPOSITION 4.1. *If the densities g_S satisfy* (2.7)*, then, for any $A \geq 1$*

$$\sup_{\theta_0 \in \ell_0[p_n]} P_{n,\theta_0}\Pi_n(\theta : |S_\theta \cap S_{\theta_0}^c| \geq A|\, X) \leq \sum_{p=A}^{n-p_n} m_1^{p_n+p} \max_{0\leq k\leq p_n}\left[\frac{\pi_{n,k}(p)}{\pi_{n,k}(0)}\right].$$

PROOF. For $B = \{\theta : |S_\theta \cap S_0^c| \geq A\}$ and $\Pi_n^{\theta_{S_0}}(\cdot|\, X)$ the marginal distribution of θ_{S_0} if θ is distributed according to the posterior distribution,

$$\begin{aligned}\Pi_n(B|\, X) &= \int \Pi_n(B|\, X, \theta_{S_0} = \bar\theta_1)\, d\Pi_n^{\theta_{S_0}}(\bar\theta_1|\, X)\\ &\leq \sup_{\bar\theta_1 \in \mathbb{R}^{S_0}} \Pi_n(B|\, X, \theta_{S_0} = \bar\theta_1).\end{aligned}$$

In the Bayesian setting the vectors X_{S_0} and $X_{S_0^c}$ are conditionally independent given θ with marginal conditional distributions depending on θ_{S_0} and

$\theta_{S_0^c}$ only, respectively. This implies that the distribution of $\theta_{S_0^c}$ given (X, θ_{S_0}) depends on $(X_{S_0^c}, \theta_{S_0})$ only. The joint distribution of $(X_{S_0^c}, \theta_{S_0^c}, \theta_{S_0})$ can be generated by first generating θ_{S_0} from its marginal distribution derived from Π_n, next generating $\theta_{S_0^c}$ from its conditional given θ_{S_0} derived from Π_n, and finally generating $X_{S_0^c}$ from the $N_{n-p_n}(\theta_{S_0^c}, I)$-distribution. It follows that the conditional distribution of $\theta_{S_0^c}$ given (X, θ_{S_0}) can also be described as the "ordinary" posterior distribution of $\theta_{S_0}^c$ given the observation $X_{S_0^c}$ relative to the prior on $\theta_{S_0^c}$ given by the conditional distribution of $\theta_{S_0^c}$ given θ_{S_0} derived from Π_n. If $\Pi_n(\cdot|\,\bar\theta_1)$ denotes the prior induced on $\mathbb{R}^{S_0^c}$ when conditioning Π_n to the event that $\theta_{S_0} = \bar\theta_1$, and $\bar n_2 = n - p_n$, then

$$\Pi_n(B|\, X, \theta_{S_0} = \bar\theta_1) = \frac{\int_B p_{\bar n_2, \bar\theta_2}(X_{S_0^c})\, d\Pi_n(\bar\theta_2|\,\bar\theta_1)}{\int p_{\bar n_2, \bar\theta_2}(X_{S_0^c})\, d\Pi_n(\bar\theta_2|\,\bar\theta_1)}. \tag{4.2}$$

The denominator of the right side can be bounded below by restricting the integral to the event when, given $\bar\theta_1$, all coordinates of $\bar\theta_2$ are zero,

$$\int p_{\bar n_2, \bar\theta_2}(X_{S_0^c})\, d\Pi_n(\bar\theta_2|\,\bar\theta_1) \geq \Pi_n(\bar\theta_2 = 0|\,\bar\theta_1) p_{\bar n_2, 0_{S_0^c}}(X_{S_0^c}).$$

Let S_2 denote the indices of the nonzero coordinates of $\bar\theta_2 \in \mathbb{R}^{S_0^c}$, θ_2 the vector of their values and $n_2 = |S_2|$, and similarly for S_1, θ_1. Then

$$\begin{aligned}\Pi_n(B|\, X, \theta_{S_0} = \bar\theta_1) &\leq \Pi_n(\bar\theta_2 = 0|\,\bar\theta_1)^{-1} \int_B \frac{p_{\bar n_2, \bar\theta_2}}{p_{\bar n_2, 0_{S_0^c}}}(X_{S_0^c})\, d\Pi_n(\bar\theta_2|\,\bar\theta_1) \\ &\leq \sum_{S_2 \subset S_0^c, |S_2| \geq A} \frac{\Pi_n(S_2|\,\bar\theta_1)}{\Pi_n(S_2 = \emptyset|\,\bar\theta_1)} \int \frac{p_{\bar n_2, \bar\theta_2}}{p_{\bar n_2, 0_{S_0^c}}}(X_{S_0^c})\, d\Pi_n(\bar\theta_2|\,\bar\theta_1, S_2)\end{aligned}$$

With the notation S_1, θ_1, θ_2 introduced above, one obtains

$$\int \frac{p_{\bar n_2, \bar\theta_2}}{p_{\bar n_2, 0_{S_0^c}}}(X_{S_0^c})\, d\Pi_n(\bar\theta_2|\,\bar\theta_1, S_2) = \int \frac{p_{n_2, \theta_2}}{p_{n_2, 0_{S_2}}}(X_{S_2}) \frac{g_{S_1,S_2}(\theta_1, \theta_2)}{\int g_{S_1,S_2}(\theta_1, \theta_2) d\theta_2} d\theta_2.$$

On the other hand, an application of Bayes' formula leads to

$$\frac{\Pi_n(S_2|\,\bar\theta_1)}{\Pi_n(S_2 = \emptyset|\,\bar\theta_1)} = \frac{\Pi_n(S_1, S_2)}{\Pi_n(S_1, S_2 = \emptyset)} \int \frac{g_{S_1,S_2}(\theta_1, \theta_2)}{g_{S_1}(\theta_1)} d\theta_2,$$

and the last ratio of prior probabilities of subsets is equal to

$$\frac{\Pi_n(S_1, S_2)}{\Pi_n(S_1, S_2 = \emptyset)} = \frac{\pi_n(p+k)}{\binom{n}{p+k}} \frac{\binom{n}{k}}{\pi_n(k)} = \frac{\pi_{n,k}(p)}{\pi_{n,k}(0)} \frac{1}{\binom{n-p_n}{p}}.$$

Combining the previous identities and condition (2.7) one obtains that $\Pi_n(B|\, X, \theta_{S_0} = \bar\theta_1)$ is bounded above, uniformly in $\bar\theta_1 \leftrightarrow (S_1, \theta_1)$, by

$$\sum_{p=A}^{n-p_n} \sum_{|S_2|=p} \max_{0\le k\le p_n} \left[\frac{\pi_{n,k}(p)}{\pi_{n,k}(0)}\right] \frac{m_1^{p_n+p}}{\binom{n-p_n}{p}} \int \frac{p_{n_2,\theta_2}}{p_{n_2,0_{S_2}}}(X_{S_2})\gamma_{S_2}(\theta_2)\, d\theta_2.$$

The proposition follows, since $P_{n,\theta_0} p_{n_2,\theta_2}/p_{n_2,0_{S_2}}(X_{S_2}) = 1$. □

LEMMA 4.1. *If π_n satisfies (2.2) with $C = 0$ and a constant D such that $m_1 D < 1$, then $\sum_{p=P_n}^{n-p_n} m_1^{p_n+p} \max_k[\pi_{n,k}(p)/\pi_{n,k}(0)] \to 0$ for P_n bigger than a sufficiently large multiple of p_n and $P_n \to \infty$.*

PROOF. From the expression of $\pi_{n,k}$ in (4.1), simple algebra leads to

$$\frac{\pi_{n,k}(p)}{\pi_{n,k}(0)} = \binom{p+k}{k} \frac{\pi_n(p+k)}{\pi_n(k)} \frac{(n-p_n)\times\cdots\times(n-p_n-p+1)}{(n-k)\times\cdots\times(n-k-p+1)}.$$

Using the assumed strict exponential decrease, the second ratio in the last display is bounded above by $e^{p\log D}$. For any integer k between 0 and p_n, the last factor (ratio) in the last display is bounded above by 1 and $\binom{p+k}{k}$ is bounded above by $\binom{p+p_n}{p_n} \le e^{p_n \log\{e(p+p_n)/p_n\}}$. Since $\log(1+x) \le x/M$, for $M > 0$ as soon as x is larger than a sufficiently large multiple of M, the result follows. □

Combining Proposition 4.1 and Lemma 4.1 concludes the proof of the first half of Theorem 2.4 and of Theorem 2.1 for priors on dimension with strict exponential decrease.

For g_S of the product form and π_n with just exponential decrease ($C > 0$ in (2.2)) such as the oracle binomial prior, we use a slight variant of the above argument. Starting from (4.2), the denominator can be bounded below with the help of Lemma 5.2 (below), applied with $\bar n_2$ instead of n, with $\theta_0 = 0$ and both $\Pi = \tilde\Pi = \Pi_n(\cdot|\, \bar\theta_1)$. This implies that $\Pi_n(B|\, X, \theta_{S_0} = \bar\theta_1)$ is bounded above by

$$e^{\sigma_2^2/2 - \mu_2^T X_{S_0^c}} \int_B \frac{p_{\bar n_2,\bar\theta_2}}{p_{\bar n_2,0_{S_0^c}}}(X_{S_0^c})\, d\Pi_n(\bar\theta_2|\, \bar\theta_1),$$

where $\mu_2 = \int \bar\theta_2 d\Pi_n(\bar\theta_2|\, \bar\theta_1)$ and $\sigma_2^2 = \int \|\bar\theta_2\|^2\, d\Pi_n(\bar\theta_2|\, \bar\theta_1)$. In fact $\mu_2 = 0$, by the assumption that the common density g has zero mean. If m_2 denotes the second moment of g, we have

$$\sigma_2^2 \;=\; \sum_{S_2|S_0^c} \Pi_n(S_2|\, S_1) m_2 |S_2| \le m_2 \sum_{p=0}^{n-p_n} p\pi_{n,k}(p) \triangleq 2\nu_k.$$

This implies that $\Pi_n(B|\, X, \theta_{S_0} = \bar\theta_1)$ is uniformly bounded in $\bar\theta_1$ by

$$\sum_{p=A}^{n-p_n} \sum_{|S_2|=p} \max_{0\le k\le p_n} (\pi_{n,k}(p) e^{\nu_k}) \frac{1}{\binom{n-p_n}{p}} \int \frac{p_{n_2,\theta_2}}{p_{n_2,0_{S_2}}}(X_{S_2}) g_{S_2}(\theta_2)\, d\theta_2.$$

To conclude one takes the P_{n,θ_0}-expectation and uses Lemma 4.2 below.

LEMMA 4.2. *If π_n satisfies (2.2), then $\nu_k \le m_2 D_1 p_n$ with D_1 that depends on C, D in (2.2) only. Furthermore, $\sum_{p=P_n}^{n-p_n} \max_k(\pi_{n,k}(p) e^{\nu_k}) \to 0$ for P_n bigger than a sufficiently large multiple of p_n and $P_n \to \infty$.*

5. Proof of Theorems 2.2 and 2.4. In view of Theorem 2.1 the posterior mass of models of dimension bigger than Ap_n, for a large constant A, tends to zero. Thus it suffices to show concentration around θ_0 in models with $|S_\theta| \le Ap_n$. This is achieved using testing arguments. Proposition 5.1 gives an explicit bound on concentration with respect to the Euclidean metric. General d_q-metrics are next treated by interpolation of metrics.

Let Φ be the standard normal distribution function and $\bar\Phi = 1 - \Phi$.

LEMMA 5.1. *For any $\alpha, \beta > 0$ and any $\theta_0, \theta_1 \in \mathbb{R}^n$ there exists a test ϕ based on $X \sim N(\theta, I)$, such that for every $\theta \in \mathbb{R}^n$ with $\|\theta - \theta_1\| \le \|\theta_0 - \theta_1\|/2 \triangleq \rho$,*

$$\alpha P_{n,\theta_0}\phi + \beta P_{n,\theta}(1-\phi) \le \alpha\bar\Phi\Big(\frac{\rho}{2} + \frac{1}{\rho}\log\frac{\alpha}{\beta}\Big) + \beta\Phi\Big(-\frac{\rho}{2} + \frac{1}{\rho}\log\frac{\alpha}{\beta}\Big).$$

This quantity can be further bounded by $2\sqrt{\alpha\beta}\, e^{-\|\theta_0-\theta_1\|^2/32}$.

We note that the bound of Lemma 5.1, even though valid for every $\alpha, \beta > 0$, is of interest only if α and β are not too different: if $\log\alpha/\beta \le -\|\theta_0 - \theta_1\|^2/32$ or $\log\alpha/\beta \ge \|\theta_0 - \theta_1\|^2/32$, then the trivial tests $\phi = 1$ and $\phi = 0$ give the better bounds α and β, respectively.

LEMMA 5.2. *For any prior probability distribution Π on $\mathbb{R}^n$, any positive measure $\tilde\Pi$ with $\tilde\Pi \le \Pi$, and any $\theta_0 \in \mathbb{R}^m$,*

$$\int \frac{p_{n,\theta}}{p_{n,\theta_0}}(X)\, d\Pi(\theta) \ge \|\tilde\Pi\| e^{-\tilde\sigma^2/2 + \tilde\mu^T(X-\theta_0)},$$

where $\tilde\mu = \int(\theta - \theta_0)\, d\tilde\Pi(\theta)/\|\tilde\Pi\|$ and $\tilde\sigma^2 = \int \|\theta - \theta_0\|^2\, d\tilde\Pi(\theta)/\|\tilde\Pi\|$. Consequently, for any $r > 0$,

$$P_{n,\theta_0}\Big(\int \frac{p_{n,\theta}}{p_{n,\theta_0}}\, d\Pi(\theta) \ge e^{-r^2}\Pi(\theta : \|\theta - \theta_0\| < r)\Big) \ge 1 - e^{-r^2/8}.$$

LEMMA 5.3. *The volume v_p of the p-dimensional Euclidean unit ball satisfies, for every $p \geq 1$, setting $d_1 = 1/\sqrt{\pi}$ and $d_2 = e^{1/6} d_1$,*

$$d_1 (2e\pi)^{p/2} p^{-p/2-1/2} \leq v_p \leq d_2 (2e\pi)^{p/2} p^{-p/2-1/2}.$$

LEMMA 5.4. *Let $S \subset \{1, \ldots, n\}$, $p = |S|$, $j \geq 1$ and $r_n^2 \geq p_n \vee \log \pi_n(p_n)^{-1}$. Let $\theta_{S,j} \in \mathbb{R}^n$ with support S and $2jr_n < \|\theta_{S,j} - \theta_0\| < 2(j+1)r_n$. For some universal constant $c_3 > 0$, we have that*

$$\begin{aligned} &\log \frac{\Pi(\theta \in \mathbb{R}^n : \ S_\theta = S, \ \|\pi_S \theta - \theta_{S,j}\| < jr_n)}{e^{-r_n^2} \Pi(\theta \in \mathbb{R}^n, \ \|\theta - \theta_0\| < r_n)} \\ &\leq c_3(p + p_n) + p \log j + 9(j+1)^2 r_n^2/64 + 7r_n^2/2. \end{aligned}$$

PROOF. Denoting $\beta_{S,j}$ the quantity in the logarithm in the last display,

$$\begin{aligned} \beta_{S,j} &\leq \frac{\Pi(S) G_S(\theta \in \mathbb{R}^S : \|\theta - \pi_S \theta_{S,j}\| < jr_n)}{e^{-r_n^2} \Pi(S_0) G_{S_0}(\theta \in \mathbb{R}^{S_0} : \|\theta - \pi_{S_0} \theta_0\| < r_n)} \\ &\leq \frac{\Pi(S) v_S (jr_n)^{|S|} \max(g_S(\theta) : \|\theta - \pi_S \theta_{S,j}\| < jr_n)}{e^{-r_n^2} \Pi(S_0) v_{S_0} r_n^{|S_0|} \min(g_{S_0}(\theta) : \|\theta - \pi_{S_0} \theta_0\| < r_n)}. \end{aligned}$$

Let us decompose, for any $\theta' \in \mathbb{R}^S$ and $\theta \in \mathbb{R}^{S_0}$,

$$\frac{g_S(\theta')}{g_{S_0}(\theta)} = \frac{g_S(\theta')}{g_{S \cap S_0}(\pi_{S \cap S_0} \theta')} \frac{g_{S \cap S_0}(\pi_{S \cap S_0} \theta')}{g_{S \cap S_0}(\pi_{S \cap S_0} \theta)} \frac{g_{S \cap S_0}(\pi_{S \cap S_0} \theta)}{g_{S_0}(\theta)}.$$

Combining this identity with (2.5) and (2.6), we obtain, with $c_2 = 1/64$,

$$\begin{aligned} \Big| \log \frac{g_S(\theta')}{g_{S_0}(\theta)} \Big| &\leq c_1 |S| + c_1 |S \cap S_0| + c_1 |S_0| \\ &\quad + c_2 \|\pi_{S - S_0} \theta'\|^2 + c_2 \|\pi_{S \cap S_0}(\theta' - \theta)\|^2 + c_2 \|\pi_{S_0 - S} \theta\|^2. \end{aligned}$$

Denoting by $\bar{\theta}, \bar{\theta}'$ the vectors of $\mathbb{R}^n$ with respective supports S_0, S and such that $\pi_{S_0} \bar{\theta} = \theta$, $\pi_S \bar{\theta}' = \theta'$, note that the last line of the previous display is bounded above by $c_2 \|\bar{\theta}' - \bar{\theta}\|^2$. For $\|\theta' - \pi_S \theta_{S,j}\| < jr_n$ and $\|\theta - \pi_{S_0} \theta_0\| < r_n$, we have

$$\|\bar{\theta}' - \bar{\theta}\| \leq \|\bar{\theta}' - \theta_{S,j}\| + \|\theta_{S,j} - \theta_0\| + \|\theta_0 - \bar{\theta}\| \leq 3(j+1) r_n.$$

Due to Lemma 5.3, the quotient $v_p r_n^p / (v_{p_n} r_n^{p_n})$ is bounded by

$$\frac{v_p r_n^{p/2}}{v_{p_n} r_n^{p_n}} \lesssim (2e\pi)^p \left(\frac{\sqrt{p_n}}{r_n} \right)^{p_n} \left(\frac{r_n}{\sqrt{p}} \right)^p.$$

Since $r_n^2 \geq p_n$ by assumption, we have $(\sqrt{p_n}/r_n)^{p_n} \leq 1$ and because the function $p \mapsto p\log(r_n^2/p)$ takes a maximum at $p = r_n^2/e$, we obtain, for some universal constants C, C',

$$\beta_{S,j} \leq j^p e^{Cp + C'p_n + 9c_2(j+1)^2 r_n^2 + (1+1/2e)r_n^2} \Pi(S)/\Pi(S_0).$$

To conclude, one notes that $\Pi(S) \leq 1$ and that $\binom{n}{p_n} \leq (ne/p_n)^{p_n} \leq e^{r_n^2 + p_n}$ by the assumption on r_n, so that $\Pi(S_0) \geq e^{-2r_n^2 - p_n}$. □

PROPOSITION 5.1. *If the densities g_S satisfy* (2.5)-(2.6) *and have finite second moments, then there exist universal constants d_1, d_2, d_3 such that for $M \geq 10$ and $1 \leq A \leq n/(2p_n)$ and r_n^2 satisfying* (2.4) *and $p_n/n \to 0$, as $n \to +\infty$,*

$$\sup_{\theta_0 \in \ell_0[p_n]} P_{n,\theta_0} \Pi_n(\theta : \|\theta - \theta_0\| > Mr_n, |S_\theta| \leq Ap_n \,|\, X) \leq e^{-r_n^2/8} + d_1 \binom{n}{Ap_n} e^{d_2 Ap_n - d_3 (Mr_n)^2}.$$

PROOF. Let $\mathcal{S}_1$ be the collection of subsets $S \subset \{1, 2, \ldots, n\}$ such that $|S| \leq Ap_n$. For each such S and $j = 1, 2, \ldots$ let $\{\theta_{S,j,i} : i \in I_{S,j}\}$ be a maximal jr_n-separated set inside the set $\{\theta \in \mathbb{R}^n : S_\theta = S,\ 2jr_n \leq \|\theta - \theta_0\| \leq 2(j+1)r_n\}$. Because the latter set is within a ball of radius $2(j+1)r_n$ of the projection $\Pi_S\theta_0$ onto the subspace of vectors with support inside S, a volume argument shows that the cardinality of $I_{S,j}$ is at most $9^{|S|}$.

We can partition the set of vectors with exactly support S by assigning each such vector to a closest point $\theta_{S,j,i}$ for some $j = 1, 2, \ldots$, and $i \in I_{S,j}$. The resulting partitioning sets $B_{S,j,i}$ will fit into balls of radius jr_n. For each $\theta_{S,j,i}$ fix a test $\phi_{S,j,i}$ as in Lemma 5.1 with $\alpha = 1$ and the triple (θ_0, θ_1), ρ and β taken equal to the triple $(\theta_0, \theta_{S,j,i})$, jr and $\beta_{S,j,i}$, where the last numbers will be determined later. In view of the second assertion of Lemma 5.2 applied with r equal to r_n, there exist events $\mathcal{A}_n$ such that $P_{n,\theta_0}(\mathcal{A}_n^c) \leq e^{-r_n^2/8}$, on which

$$\int \frac{p_{n,\theta}}{p_{n,\theta_0}} \, d\Pi_n(\theta) \geq e^{-r_n^2} \Pi_n(\theta : \|\theta - \theta_0\| < r_n).$$

We have that

$$
\begin{aligned}
&P_{n,\theta_0}\Pi_n(\theta : \|\theta-\theta_0\| > 2Mr_n, S_\theta \in \mathcal{S}_1 | X)1_{\mathcal{A}_n} \\
&\quad \le \sum_{S\in\mathcal{S}_1}\sum_{j\ge M}\sum_{i\in I_{S,j}} P_{n,\theta_0}\Pi_n(\theta \in B_{S,j,i}| X)1_{\mathcal{A}_n} \\
&\quad \le \sum_{S\in\mathcal{S}_1}\sum_{j\ge M}\sum_{i\in I_{S,j}} \Big(P_{n,\theta_0}\phi_{S,j,i} + P_{n,\theta_0}\Big[(1-\phi_{S,j,i})\frac{\int_{B_{S,j,i}} p_{n,\theta}/p_{n,\theta_0}\, d\Pi(\theta)}{e^{-r_n^2}\Pi\big(\theta : \|\theta-\theta_0\| < r_n\big)}\Big]\Big) \\
&\quad \le \sum_{S\in\mathcal{S}_1}\sum_{j\ge M}\sum_{i\in I_{S,j}} \Big(P_{n,\theta_0}\phi_{S,j,i} + \beta_{S,j,i}\sup_{\theta\in B_{S,j,i}} P_{n,\theta}(1-\phi_{S,j,i})\Big),
\end{aligned}
$$

where we have denoted

$$\beta_{S,j,i} = \frac{\Pi(B_{S,j,i})}{e^{-r_n^2}\Pi\big(\theta : \|\theta-\theta_0\| < r_n\big)}.$$

In view of Lemma 5.1 the term within the triple sum is bounded using by $2\sqrt{\beta_{S,j,i}}e^{-j^2r_n^2/8}$. Since $|S| = p \le Ap_n$ and $p_n/n \to 0$, we can take n large enough in order to have both $c_3(p+p_n) \le r_n^2/10$ and $p\log j \le j^2r_n^2/100$ for any $j \ge 1$. Since $M \ge 10$, we have $j \ge 10$, so we also have $r_n^2 \le j^2r_n^2/100$.

Combination with Lemma 5.4 now yields the bound, for $j \ge 10$,

$$\log\sqrt{\beta_{S,j,i}} \le 2.3j^2r_n^2/100 + 9(j+1)^2r_n^2/128.$$

One easily checks that this is bounded by $(1-d_2)j^2r_n^2/8$, for $d_2 = 1/9$ when $j \ge 10$. Thus the probability at stake is bounded from above by

$$\sum_{p=0}^{Ap_n}\binom{n}{p}\sum_{j\ge M} 2C^p e^{-d_2j^2r_n^2} \le d_1^{Ap_n}e^{-d_2M^2r_n^2}\sum_{p=0}^{Ap_n}\binom{n}{p},$$

for d_1 large enough. By assumption $Ap_n \le n/2$, so each binomial term is bounded by the last one. Using simple algebra this yields the second term in the bound of the theorem. The first term comes from $P_{n,\theta_0}1_{\mathcal{A}_n^c} \le e^{-r_n^2/8}$. $\square$

In view of (2.4) we have $\binom{n}{Ap_n} \le (ne/Ap_n)^{Ap_n} \le e^{d_4r_n^2}$. Therefore, the right side of Proposition 5.1 tends to zero. combination with Theorem 2.1 yields proofs of Theorems 2.2 and 2.4 for d_q the square Euclidean norm d_2.

The theorems for $q \in (0,2)$ are a corollary of the case $q = 2$, by interpolation between the distances. Due to Hölder's inequality, for any θ, θ_0 with $|S_\theta \cup S_0| \le Ap_n$,

$$d_q(\theta,\theta_0) \le \|\theta-\theta_0\|^q(Ap_n)^{1-q/2}.$$

This implies, for any $M > 0$, if $\theta_0 \in \ell_0[p_n]$,

$$P_{n,\theta_0}\Pi_n(d_q(\theta,\theta_0) > Mr_n^q p_n^{1-q/2}|\, X) \le P_{n,\theta_0}\Pi_n(\theta : |S_\theta| > (A-1)p_n|\, X) + P_{\theta_0}^n\Pi(\|\theta-\theta_0\| > M^{1/q}A^{1/2-1/q}r_n|\, X).$$

The first term on the right side tends to zero for sufficiently large A. Next the second tends to zero for sufficiently large M.

6. Proof of Theorem 2.6. The theorem is proved by bounding the (posterior) risk under a vector $\theta_0 \in m_s[p_n]$ by the risk under its projection into $\ell_0[p]$ obtained by setting the smallest $n-p$ coordinates of θ_0 equal to zero. The value p that minimizes the expression that defines the rate r_n^2 is the optimal dimension of a projection, and the complicated expression itself is a trade-off of an approximation error and a rate.

The comparison between θ_0 and its projection θ_1 is made in the following lemma.

LEMMA 6.1. *For any measurable function G and any θ_0, θ_1 in $\mathbb{R}^n$,*

$$P_{n,\theta_0}G \le \sqrt{P_{n,\theta_1}G^2}\, e^{\|\theta_1-\theta_0\|^2/2}.$$

PROOF. In view of the Cauchy-Schwarz inequality,

$$P_{n,\theta_0}G \;\le\; \sqrt{P_{n,\theta_1}G^2}\sqrt{P_{n,\theta_1}\Big(\frac{dP_{n,\theta_0}}{dP_{n,\theta_1}}\Big)^2}.$$

The second integral on the right is equal to $\exp(\|\theta_0-\theta_1\|^2)$. □

Let p_n^* be an index for which the minimum that defines the rate r_n^2 is attained. For given θ_0 belonging to $m_s[p_n]$, let θ_1 denote the vector deduced from θ_0 by keeping unchanged its p_n^* largest components and putting the other ones to 0. By definition θ_1 belongs to $\ell_0[p_n^*]$ and

$$\begin{aligned}\|\theta_1-\theta_0\|^2 &= \sum_{i>p_n^*}|\theta_{0,[i]}|^2 \le \left(\frac{p_n}{n}\right)^2 \sum_{i>p_n^*}\left(\frac{n}{i}\right)^{2/s} \\ &\le \left(\frac{p_n}{n}\right)^2\left(\frac{s}{2-s}\right)n^{2/s}(p_n^*)^{1-2/s} \le r_n^2, \qquad (6.1)\end{aligned}$$

where the first inequality is obtained using the definition of the $m_s[p_n]$-class, and the second follows by comparison of the series with an integral.

Therefore, the triangle inequality implies

$$\Pi_n(\theta : \|\theta-\theta_0\| > 80r_n + 20r|\, X) \le \Pi_n(\theta : \|\theta-\theta_1\| > 79r_n + 20r|\, X).$$

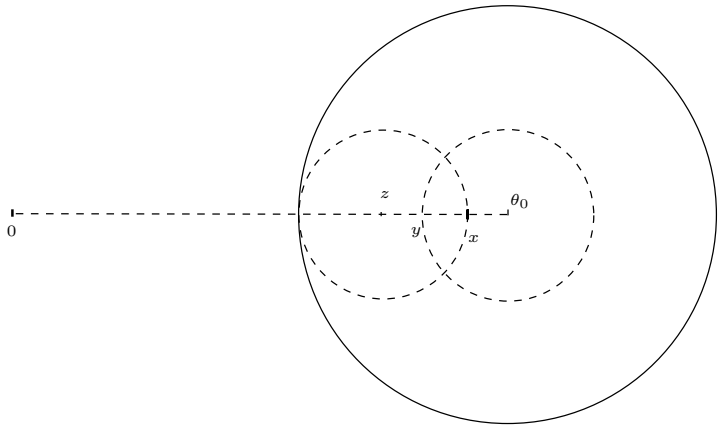

FIG 2. *Idea behind the proof of Theorem 2.8*

By Lemma 6.1 the expectation of the right side under P_{n,θ_0} is bounded by

$$\big(P_{n,\theta_1}\Pi_n(\theta : \|\theta - \theta_1\| > 79r_n + 20r \,|\, X)\big)^{1/2} e^{\|\theta_0 - \theta_1\|^2/2}.$$

Finally apply Theorem 2.5, with r of the theorem taken equal to $3.4r_n + 2r$.

7. Proof of Theorems 2.8 and 2.9. The proof of Theorem 2.8 follows the approach to get lower bound type results introduced in [8], which uses the principle that sets with very little prior mass receive no posterior mass.

LEMMA 7.1. *We have $P_{n,\theta_0}\Pi_n(\theta : \|\theta - \theta_0\| < s_n \,|\, X) \to 0$, for any s_n for which there exist r_n such that*

$$\frac{\Pi_n(\theta : \|\theta - \theta_0\| < s_n)}{\Pi_n(\theta : \|\theta - \theta_0\| < r_n)} = o(e^{-r_n^2}).$$

LEMMA 7.2. *There exist a constant $C > 0$ such that, if $S \subset \{1, \ldots, n\}$ and r_n is a sequence of real numbers such that $r_n^2 \geq |S_{\theta_0}|$, it holds*

$$\frac{v_{|S \cap S_{\theta_0}|}}{v_{|S_{\theta_0}|}} \frac{1}{r_n^{|S_{\theta_0} \setminus S|}} \leq e^{C|S_{\theta_0}|}.$$

PROOF OF THEOREM 2.8. We first consider the (more complicated) case that $1 < \alpha < 2$. For this range of α an application of Hölder's inequality

gives that $\|\theta\|_\alpha \leq \|\theta\|\, p^{1/\alpha-1/2}$, if p is the number of nonzero coordinates of a vector θ. Let us introduce

$$r_n = \left(\frac{\|\theta_0\|_\alpha^\alpha}{\|\theta_0\|^2} \wedge 1\right)\frac{\|\theta_0\|}{8}, \qquad s_n = \frac{\rho_{0,\alpha}^n}{64} = \frac{r_n}{8}\left(\frac{\|\theta_0\|_\alpha}{\|\theta_0\|} p_n^{\frac{1}{2}-\frac{1}{\alpha}}\right).$$

Then $r_n \leq \|\theta_0\|/8$ and $s_n \leq r_n/8$. Also,

$$\begin{aligned}
&\frac{\Pi_n(\theta : \|\theta-\theta_0\| < s_n)}{\Pi_n(\theta : \|\theta-\theta_0\| < r_n)} \\
&\quad = \sum_S \Pi_n(S)\frac{G_S(\theta \in \mathbb{R}^S : \|\theta - \pi_S\theta_0\|^2 + \|\pi_{S_0\setminus S}\theta_0\|^2 < s_n^2)}{\Pi_n(\theta : \|\theta-\theta_0\| < r_n)} \\
&\quad \leq \sum_S \frac{\Pi_n(S)}{\Pi_n(S_0)}\frac{G_{S\cap S_0}(\theta \in \mathbb{R}^{S\cap S_0} : \|\theta - \pi_{S\cap S_0}\theta_0\| \leq s_n)}{G_{S_0}(\theta \in \mathbb{R}^{S_0} : \|\theta - \pi_{S_0}\theta_0\| \leq r_n)} 1_{\|\pi_{S_0\setminus S}\theta_0\| < s_n}.
\end{aligned}$$

Define

$$\theta_B = \left(1 - \frac{r_n - s_n}{\|\theta_0\|}\right)\pi_{S_0}\theta_0^n.$$

Then the ball in $\mathbb{R}^{S_0}$ of radius s_n around θ_B is contained in the ball of radius r_n around $\pi_{S_0}\theta_0$. It follows that the second last display is bounded above by

$$\sum_S \frac{\Pi_n(S)}{\Pi_n(S_0)}\frac{s_n^{|S\cap S_0|} v_{S\cap S_0}}{s_n^{p_n} v_{p_n}}\frac{\sup_{\theta\in A} g_{S\cap S_0}(\theta)}{\inf_{\theta\in B} g_{S_0}(\theta)} 1_{\|\pi_{S_0\setminus S}\theta_0\| \leq s_n}, \tag{7.1}$$

with $A = \{\theta \in \mathbb{R}^{S\cap S_0} : \|\theta - \pi_{S\cap S_0}\theta_0^n\| < s_n\}$ and $B = \{\theta \in \mathbb{R}^{S_0} : \|\theta - \theta_B\| < s_n\}$. We finish the proof by bounding the densities $g_{S\cap S_0}$ and g_{S_0} above and below on the given sets.

If $\theta \in B$, then by the triangle inequality followed by Hölder's inequality,

$$\begin{aligned}
\|\theta\|_\alpha &\leq \|\theta_B\|_\alpha + \|\theta - \theta_B\|_\alpha \\
&\leq \left(1 - \frac{r_n - s_n}{\|\theta_0\|}\right)\|\theta_0\|_\alpha + p_n^{\frac{1}{\alpha}-\frac{1}{2}} s_n \leq \left(1 - \frac{3r_n}{4\|\theta_0\|}\right)\|\theta_0\|_\alpha,
\end{aligned}$$

because $s_n \leq r_n/8$ and $p_n^{\frac{1}{\alpha}-\frac{1}{2}} s_n \leq (r_n/8)\|\theta_0\|_\alpha/\|\theta_0\|$. Similarly, if $\theta \in A$ and $\|\pi_{S_0\setminus S}\theta_0\| < s_n$, then $\|\pi_{S_0\setminus S}\theta_0\|_\alpha < p_n^{1/\alpha-1/2} s_n$ and

$$\begin{aligned}
\|\theta\|_\alpha &\geq \|\theta_0\|_\alpha - \|\theta_0 - \pi_{S\cap S_0}\theta_0\|_\alpha - \|\pi_{S\cap S_0}\theta_0 - \theta\|_\alpha \\
&\geq \|\theta_0\|_\alpha - 2p_n^{\frac{1}{\alpha}-\frac{1}{2}} s_n \geq \|\theta_0\|_\alpha\Big(1 - \frac{r_n}{4\|\theta_0\|}\Big).
\end{aligned}$$

We deduce that, for any S such that $\left\|\pi_{S_0\setminus S}\theta_0\right\| \leq s_n$, denoting by c_α the normalizing constant of the density $x \to c_\alpha \exp(-|x|^\alpha)$,

$$\frac{c_\alpha^{p_n}}{c_\alpha^{|S\cap S_0|}} \frac{\sup_{\theta\in A} g_{S\cap S_0}(\theta)}{\inf_{\theta\in B} g_{S_0}(\theta)} \leq \exp\left[\|\theta_0\|_\alpha^\alpha \left\{\Big(1 - \frac{3r_n}{4\|\theta_0\|}\Big)^\alpha - \Big(1 - \frac{r_n}{4\|\theta_0\|}\Big)^\alpha\right\}\right]$$
$$\leq \exp\left[-2\alpha(5/8)^{\alpha-1} r_n \frac{\|\theta_0\|_\alpha^\alpha}{4\|\theta_0\|}\right] \leq \exp\left[-4\alpha(5/8)^{\alpha-1} r_n^2\right],$$

where to obtain the second last inequality we have used that for any $0 \leq t \leq 1/8$ and $\alpha \geq 1$ it holds $(1-t)^\alpha - (1-3t)^\alpha = \int_1^3 \alpha t(1-ut)^{\alpha-1}\, du \geq 2\alpha t(1-3/8)^{\alpha-1}$. Hence the expression in (7.1) is bounded above by

$$\sum_S \frac{\Pi_n(S)}{\Pi_n(S_0)} (c_\alpha s_n)^{|S\cap S_0|-p_n} \frac{v_{|S\cap S_0|}}{v_{p_n}} e^{-4\alpha(5/8)^{\alpha-1} r_n^2}$$
$$\leq e^{-4\alpha(5/8)^{\alpha-1} r_n^2} \frac{e^{Cp_n}}{\Pi_n(S_0)} \sum_S \Pi_n(S)$$
$$\leq e^{-4\alpha(5/8)^{\alpha-1} r_n^2} e^{Cp_n} e^{cp_n \log(n/p_n)},$$

by Lemma 7.2. The right side is of smaller order than $e^{-r_n^2}$. An application of Lemma 7.1 concludes the proof for the case that $1 < \alpha < 2$.

The proof in the case that $\alpha \geq 2$ follows the same lines, except that we use the inequality $\|\theta\|_\alpha \leq \|\theta\|$, for every $\theta \in \mathbb{R}^p$, without the factor $p^{1/\alpha-1/2}$ that is necessary if $\alpha < 2$. We define $s_n = (r_n/8)\|\theta_0\|_\alpha/\|\theta_0\|$. □

Acknowledgement. The authors would like to thank Subhashis Ghosal for suggesting a simplified argument in the proof of Proposition 4.1.

REFERENCES

[1] Abramovich, F., Benjamini, Y., Donoho, D. L., and Johnstone, I. M. Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist. 34*, 2 (2006), 584–653.

[2] Abramovich, F., Grinshtein, V., and Pensky, M. On optimality of Bayesian testimation in the normal means problem. *Ann. Statist. 35*, 5 (2007), 2261–2286.

[3] Bickel, P. J., Ritov, Y., and Tsybakov, A. B. Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist. 37*, 4 (2009), 1705–1732.

[4] Birgé, L., and Massart, P. Gaussian model selection. *J. Eur. Math. Soc. (JEMS) 3*, 3 (2001), 203–268.

[5] Brown, L. D., and Greenshtein, E. Nonparametric empirical Bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *Ann. Statist. 37*, 4 (2009), 1685–1704.

[6] Cai, T. T., Jin, J., and Low, M. G. Estimation and confidence sets for sparse normal mixtures. *Ann. Statist. 35*, 6 (2007), 2421–2449.

[7] Candes, E., and Tao, T. The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist. 35*, 6 (2007), 2313–2351.

[8] Castillo, I. Lower bounds for posterior rates with Gaussian process priors. *Electronic Journal of Statistics 2* (2008), 1281–1299.

[9] Castillo, I., and van der Vaart, A. W. Supplement to 'Needles and Straws in a Haystack: Posterior concentration for possibly sparse sequences'.

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

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

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

[13] Golubev, G. K. Reconstruction of sparse vectors in white Gaussian noise. *Problemy Peredachi Informatsii 38*, 1 (2002), 75–91.

[14] Huang, J., Ma, S., and Zhang, C.-H. Adaptive Lasso for sparse high-dimensional regression models. *Statist. Sinica 18*, 4 (2008), 1603–1618.

[15] Jiang, W., and Zhang, C.-H. General maximum likelihood empirical Bayes estimation of normal means. *Ann. Statist. 37*, 4 (2009), 1647–1684.

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

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

[18] Scott, J., and Berger, J. Bayes and Empirical-Bayes multiplicity adjustment in the variable-selection problem. *Ann. Statist.* (2010). To appear.

[19] Yuan, M., and Lin, Y. Efficient empirical Bayes variable selection and estimation in linear models. *J. Amer. Statist. Assoc. 100*, 472 (2005), 1215–1225.

[20] Zhang, C.-H. General empirical Bayes wavelet methods and exactly adaptive minimax estimation. *Ann. Statist. 33*, 1 (2005), 54–100.

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

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

CNRS & Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris VI & VII,
E-mail: ismael.castillo@upmc.fr
Department of Mathematics, Faculty of Sciences, VU University Amsterdam,
E-mail: aad@cs.vu.nl