Bernoulli **26**(1), 2020, 191–225
https://doi.org/10.3150/19-BEJ1122

Needles and straw in a haystack: Robust confidence for possibly sparse sequences

EDUARD BELITSER[1] and NURZHAN NURUSHEV[1,2]

[1]*Department of Mathematics, VU Amsterdam, The Netherlands. E-mail: e.n.belitser@vu.nl*
[2]*Korteweg-de Vries Institute for Mathematics, University of Amsterdam, The Netherlands. E-mail: n.nurushev@uva.nl*

In the general *signal+noise* (allowing non-normal, non-independent observations) model, we construct an empirical Bayes posterior which we then use for *uncertainty quantification* for the unknown, possibly sparse, signal. We introduce a novel *excessive bias restriction* (EBR) condition, which gives rise to a new slicing of the entire space that is suitable for uncertainty quantification. Under EBR and some mild *exchangeable exponential moment condition* on the noise, we establish the local (oracle) optimality of the proposed confidence ball. Without EBR, we propose another confidence ball of full coverage, but its radius contains an additional $\sigma n^{1/4}$-term. In passing, we also get the local optimal results for *estimation*, *posterior contraction* problems, and the problem of *weak recovery of sparsity structure*. Adaptive minimax results (also for the estimation and posterior contraction problems) over various sparsity classes follow from our local results.

Keywords: confidence set; empirical Bayes posterior; local radial rate

1. Introduction

The model and the main problem

Suppose we observe $X = X^{(\sigma,n)} = (X_1, \dots, X_n)$:

$$X_i = \theta_i + \sigma\xi_i, \quad i \in [n] = \{1, \dots, n\}, \tag{1}$$

where $\theta = (\theta_1, \dots, \theta_n) \in \mathbb{R}^n$ is an unknown high-dimensional parameter of interest, the ξ_i's are random errors with $\mathrm{E}_\theta \xi_i = 0$, $\mathrm{Var}_\theta(\xi_i) \le C_\xi$, $\sigma > 0$ is the known noise intensity. The goal is to make inference about the parameter θ based on the data X: recovery of θ and *uncertainty quantification* by constructing an *optimal confidence set*. We pursue *robust inference* in the sense that the distribution of the error vector $\xi = (\xi_1, \dots, \xi_n)$ is unknown and may also depend on θ, but assumed to satisfy only certain mild *exchangeable exponential moment condition*; see Condition (A1) in Section 2. We exploit the empirical Bayes approach and derive non-asymptotic results, which imply asymptotic assertions as well if needed. Possible asymptotic regimes are decreasing noise level $\sigma \to 0$, high-dimensional setup $n \to \infty$ (the leading case for high dimensional models), or their combination, for example, $\sigma = n^{-1/2}$ and $n \to \infty$.

Useful inference is not possible without some structure on θ. Popular structural assumptions are *smoothness* and *sparsity*, in this paper we are concerned with the latter. The best studied

1350-7265

problem in the sparsity context is that of estimating θ in the many normal means model, a variety of estimation methods and results are available in the literature: [1,8,14,16,19,30]. However, even an optimal estimator does not reveal how far it is from θ. It is of importance to quantify this uncertainty, which can be seen as the problem of constructing confidence sets for θ.

Bayesian approach and accompanying posterior contraction problem

Many inference methods have Bayesian connections. For example, even some seemingly non Bayesian estimators can be obtained as certain quantities (like posterior mode for penalized minimum contrast estimators) of the (empirical Bayes) posterior distributions resulting from imposing some specific priors on the parameter; cf. [19] and [1]. Although the Bayesian methodology is used or can be related to in constructing many (frequentist) inference procedures, only recently the posterior distributions themselves have been studied in the sparsity context: [7,12,14,21,25, 27,30].

In this paper, for inference on θ we use an empirical Bayes approach. Since any Bayesian approach always delivers a posterior $\pi(\vartheta|X)$ (in the posteriors for θ, we will use the variable ϑ to distinguish it from the "true" θ), an accompanying problem of interest is the contraction of the resulting (empirical Bayes) posterior to the "true" θ from the frequentist perspective of the "true" measure P_θ, the distribution of X from (1). The quality of posterior is characterized by the posterior contraction rate. We pursue a novel local approach by allowing the posterior contraction rate to be a local quantity, that is, depending on the true θ, whereas global minimax rates are typically studied in the literature on Bayesian nonparametrics.

A common Bayesian way to model sparsity structure is by the so called two-groups priors. Such a prior puts positive mass on vectors θ with some exact zero coordinates (zero group) and the remaining coordinates (signal group) are drawn from a chosen distribution. So the marginal prior for each coordinate is a mixture of a continuous distribution and a point-mass at zero. In [14] it is shown that for a suitably chosen two-groups prior, the posterior concentrates around the true θ at the minimax rate (as $n \to \infty$) for two sparsity classes, *nearly black vectors* $\ell_0[p_n]$ with p_n nonzero coordinates and *weak* ℓ_s*-balls* $m_s[p_n]$. As pointed out by [14] (also by [19]), the prior distributions of non-zero coordinates should not have too light tails, otherwise one gets sub-optimal rates. The important Gaussian case is for example excluded. This has to do with the so called *over-shrinkage effect* of the normal prior with a fixed mean, which pushes the posterior too much towards the prior mean, missing the true parameter that in general differs from the prior mean. That is why [19] and [14] discard normal priors on non-zero coordinates and use heavy tailed priors. In the present paper, we show that normal priors are still usable (cf. [21]) and lead to strong local results (even for non-normal models) if combined with empirical Bayes approach.

Uncertainty quantification problem

The main aim in this paper is to construct confidence sets with optimal properties. The size of a confidence set is measured by the smallest radius of a ball containing this set, hence it suffices to consider confidence balls. For the usual norm $\|\cdot\|$ in $\mathbb{R}^n$, a random ball in $\mathbb{R}^n$ is $B(\hat{\theta},\hat{r}) = \{\theta \in \mathbb{R}^n : \|\hat{\theta}-\theta\| \le \hat{r}\}$, where the center $\hat{\theta} = \hat{\theta}(X) : \mathbb{R}^n \mapsto \mathbb{R}^n$ and radius $\hat{r} = \hat{r}(X) : \mathbb{R}^n \mapsto \mathbb{R}_+ =$

$[0, +\infty]$ are measurable functions of the data X. Let us introduce the optimality framework for uncertainty quantification. The goal is to construct such a confidence ball $B(\hat{\theta}, C\hat{r})$ that for any $\alpha_1, \alpha_2 \in (0, 1]$ and some functional $r(\theta) = r_{\sigma,n}(\theta)$, $r : \mathbb{R}^n \to \mathbb{R}_+$, there exist $C, c > 0$ such that

$$\sup_{\theta \in \Theta_0} \mathrm{P}_\theta\big(\theta \notin B(\hat{\theta}, C\hat{r})\big) \le \alpha_1, \qquad \sup_{\theta \in \Theta_1} \mathrm{P}_\theta\big(\hat{r} \ge c r(\theta)\big) \le \alpha_2, \tag{2}$$

for some $\Theta_0, \Theta_1 \subseteq \mathbb{R}^n$. The function $r(\theta)$, called the *radial rate*, is a benchmark for the effective radius of the confidence ball $B(\hat{\theta}, C\hat{r})$. The first expression in (2) is called *coverage relation* and the second *size relation*. Notice that our approach is local (and hence genuinely adaptive) as the radial rate $r(\theta)$ is a function of the "true" parameter θ. The (quadratic) minimax rate over a class Θ_β is defined as $r^2(\Theta_\beta) \triangleq \inf_{\hat{\theta}} \sup_{\theta \in \Theta_\beta} \mathrm{E}_\theta \|\hat{\theta} - \theta\|^2$, where the infimum is taken over all estimators. Recall the common (global) minimax adaptive version of (2): given a *scale* (a family of sets) $\{\Theta_\beta, \beta \in \mathcal{B}\}$, with corresponding minimax rates $r(\Theta_\beta)$ indexed by a structural parameter $\beta \in \mathcal{B}$ (e.g., smoothness or sparsity), the minimax adaptive version of (2) is obtained by taking $\Theta_0 = \Theta_1 = \Theta_\beta$ with the radial rate $r(\theta) = r(\Theta_\beta)$ for all $\theta \in \Theta_\beta$, $\beta \in \mathcal{B}$.

Coming back to our local framework (2), it is desirable to find the smallest $r(\theta)$ and the biggest Θ_0, Θ_1, for which (2) holds. These are contrary requirements, so we have to trade them off against each other. There are different ways of doing this, leading to different optimality frameworks. For example, if we insist on overall uniformity $\Theta_0 = \mathbb{R}^n$, then the results in [20] and [11] (more refined versions are in [3] and [22]) say basically that the radial rate r cannot be of a faster order than $\sigma n^{1/4}$ for every θ and is at least of order $\sigma n^{1/2}$ for some θ. This means that any confidence ball that is optimal with respect to the optimality framework (2) with $\Theta_0 = \mathbb{R}^n$ will necessarily have a big size, even if the true θ happens to lie in a very "good", smooth or sparse, class Θ_1. Many good confidence sets cannot be optimal in this sense (called "honest" in some papers) and effectively excluded from the consideration. For minimax adaptive versions of (2) this means that as soon as we require $\Theta_0 = \Theta_\beta$, $\beta \in \mathcal{B}$ in the coverage relation, the minimax rate $r(\Theta_\beta)$ in the size relation is unattainable even when $\beta \in \mathcal{B} = \{\beta_1, \beta_2\}$ for certain β_1, β_2; cf. [22] for two nearly black classes. Essentially, the overall uniform coverage and optimal size properties can not hold together, it is necessary to sacrifice at least one of these, preferably as little as possible. We argue that it is unreasonable to pursue an optimality framework with the entire space $\Theta_0 = \mathbb{R}^n$ in the coverage relation, because this leads to discarding many good procedures and optimality of uninteresting ones. Instead, it makes sense to sacrifice in the set $\Theta_0 = \mathbb{R}^n \backslash \Theta'$, by removing a preferably small portion of "deceptive parameters" Θ' from $\mathbb{R}^n$ so that that the optimal radial rates become attainable in the size relation with interesting (preferably "massive") sets Θ_1.

This "deceptiveness" phenomenon is well understood for some smoothness structures (e.g., Sobolev scale), especially in global minimax settings; see [4,10,24] and [28]. If we now insist on the optimal size property in (2) for all Θ_β, $\beta \in \mathcal{B}$, the coverage relation in (2) will not hold for all $\Theta_0 = \Theta_\beta$, but only for $\Theta_0 = \Theta_\beta \backslash \Theta'$, with some set of "deceptive parameters" Θ' removed from Θ_β. In [28] such parameters are called "inconvenient truths" and an implicit construction of a $\theta \in \Theta'$ is given. Examples of non-deceptive parameters are the set of *self-similar* parameters $\Theta_0 = \Theta_{\mathrm{ss}}$ introduced by [23] and studied by [9,10,28], and the set of *polished tail parameters* $\Theta_0 = \Theta_{\mathrm{pt}}$ considered by [28]. All the above mentioned papers study global minimax radial rates (i.e., $r(\theta) = r(\Theta_\beta)$ for all $\theta \in \Theta_\beta$) for specific smoothness scales. A local approach, delivering also the adaptive minimax results for many smoothness scales simultaneously, is considered by [2] for posterior contraction rates and by [4] for constructing optimal confidence balls. In [4], yet a more general (than Θ_{ss} and Θ_{pt}) set of non-deceptive parameters was introduced, $\Theta_0 = \Theta_{\mathrm{eb}}$, parameters satisfying the so called *excessive bias restriction* (EBR).

To the best of our knowledge, there are very few papers about adaptive results on uncertainty quantification (2). The case of two nearly black classes is treated by [22], the "general polished tail" condition was introduced in [26] to describe non-deceptive parameters, a more general case of linear regression model (but with the standard normal noise) is recently studied in [5]. A restricted scale of nearly black classes is treated in [29], where effectively a version of our EBR condition is used; more on relation to paper [29] can be found in Supplement [6].

The scope of this paper

In this paper, we propose an empirical Bayes procedure, in fact, two procedures. There are four distinctive features of our approach: *robust*, *local*, *refined* and *EBR*.

First, *robust* means that our results cover also misspecified models, as we allow the ξ_i's to be not necessarily independent normals (a certain type of error misspecification was also mentioned in a remark of the supplement to [12]). Precisely, we use the normal likelihood, whereas the true model (1) does not have to be normal and independence of ξ_i's is not required either, but only satisfying some mild Condition (A1) (see below), called *exchangeable exponential moment condition*. It turned out that, although we use the normal likelihood (whereas the true model may not be normal) in the Bayesian analysis when proving the main results, we can handle the frequentist behavior of the posterior from the perspective of the true measure only on the basis of Condition (A1).

Second, we develop the novel *local* approach, meaning that the radial rate $r(\theta)$ in (2) is allowed to be a function of θ, which, in a way, measures the amount of sparsity for each $\theta \in \mathbb{R}^n$: the smaller $r(\theta)$, the more sparse θ. The local radial rate $r(\theta)$ is constructed as the best rate over a certain family of local rates, therefore called *oracle rate*. We demonstrate that the local approach is more powerful than global in that we do not need to impose any specific sparsity structure, because the proposed local approach automatically exploits the "effective" sparsity of each underlying θ, and our local results imply a whole panorama of the global minimax results for many scales at once. More on this is in Section 3.5.

Third, we derive the local posterior contraction result for the resulting empirical Bayes posterior $\hat{\pi}(\vartheta|X)$ in the *refined non-asymptotic formulation*: $\sup_{\theta\in\mathbb{R}^n} \mathrm{E}_\theta \hat{\pi}(\|\vartheta - \theta\|^2 \geq M_0 r^2(\theta) + M\sigma^2|X) \leq H_0 e^{-m_0 M}$ for some fixed $M_0, H_0, m_0 > 0$ and arbitrary $M \geq 0$, as an exponential non-asymptotic concentration bound in terms of M, uniformly in $\theta \in \mathbb{R}^n$. This formulation provides a rather refined characterization of the quality of the posterior (finer, than, e.g., asymptotically in terms of the dimension n), allowing subtle analysis for various asymptotic regimes. This result is of interest and importance on its own as it actually establishes the contraction of the empirical Bayes posterior with the local rate $r(\theta)$. Besides, we obtain the oracle estimation result (also in similar refined formulation, finer than traditional oracle inequalities) by constructing an estimator, the empirical Bayes posterior mean, which converges to θ with the local rate $r(\theta)$. This result, besides being an ingredient for the uncertainty quantification problem (2), is also of interest on its own as it delivers the same type of (oracle and minimax) estimation results as in [1] and [19] and posterior convergence results as in [14], obtained for different priors.

Next, we construct a confidence ball by using the empirical Bayes posterior quantities. Since we want the size of our confidence sets to be of an oracle rate order, this comes with the price that the coverage property can hold uniformly only over some set of parameters satisfying the so called *excessive bias restriction* (EBR) $\Theta_0 = \Theta_{\mathrm{eb}} \subseteq \mathbb{R}^n$. The main result consists in establishing the optimality (2) of the constructed confidence ball for the optimality framework $\Theta_0 = \Theta_{\mathrm{eb}}$, $\Theta_1 = \mathbb{R}^n$ and the local radial rate $r(\theta)$. The important consequence of our local approach is that a whole panorama of adaptive (global) minimax results (for all considered problems: estimation, posterior contraction rate and uncertainty quantification) over *all* sparsity scales *covered* by $r(\theta)$ follow from our local results; see Section 3.5. We also treat the situation when $\Theta_0 = \mathbb{R}^n$ in (2) by constructing a confidence ball with the radius of the order $\sigma n^{1/4} + r(\theta)$. As we already discussed, the term $\sigma n^{1/4}$ in the size relation is necessary for the uniform coverage to hold. Clearly, this confidence ball will have optimal size only for non-sparse parameters, for which $r(\theta) \geq c\sigma n^{1/4}$.

Although the original motivation of the EBR condition was to remove the deceptive parameters, it turned out to be a useful notion in the context of uncertainty quantification. In effect, the EBR condition leads to a *new sparsity EBR-scale* which gives the slicing of the entire space that is very suitable for uncertainty quantification. This provides a new perspective at the above mentioned "deceptiveness" issue: basically, each parameter is deceptive (or non deceptive) to some extent. It is the structural parameter of the new EBR-scale that measures the deceptiveness amount, and the (mild and controllable) price for handling deceptive parameters is the effective amount of inflating of the confidence ball that matches the amount of deceptiveness needed to provide a high coverage. The EBR condition and EBR-scale are discussed in Supplement [6].

The paper is organized as follows. In Section 2, we introduce the notation and the prior, describe the empirical Bayes procedure, and state the exchangeable exponential moment condition on ξ. In Section 3, we introduce the EBR condition, present the main local results of the paper, and consider a couple of scales for which global minimax results follow. A small simulation study is in Section 4. The proofs of the lemmas and theorems are given in Sections 5 and 6, respectively. In Supplement [6], we discuss the EBR condition, the EBR scale, and some other relarted points. The proofs of three theorems and some concluding remarks are also provided in Supplement [6].

2. Preliminaries

First, we introduce some notation and a family of normal priors (similar to priors from [4] but geared towards modeling sparsity rather than smoothness). Next, by applying the empirical Bayes approach to the normal likelihood, we derive an empirical Bayes posterior which we will use in the construction of the estimator and the confidence ball. We complete this section with introducing the *exchangeable exponential moment condition* on the error vector ξ.

2.1. Notation

Denote the probability measure of X from the model (1) by $\mathrm{P}_\theta = \mathrm{P}_\theta^{(\sigma,n)}$, and by E_θ the corresponding expectation. For notational simplicity, we often skip the dependence on σ and n.

Denote by $\mathbb{1}_E = \mathbb{1}\{E\}$ the indicator function of the event E, by $|S|$ the cardinality of the set S, the difference of sets $S\backslash S_0 = \{s \in S : s \notin S_0\}$. Let $[k] = \{1, \dots, k\}$ and $[k]_0 = \{0\} \cup [k]$ for $k \in \mathbb{N} = \{1, 2, \dots\}$. For $I \subseteq [n]$, define $I^c = [n]\backslash I$. Let $\mathcal{I} = \mathcal{I}_n = \{I : I \subseteq [n]\}$ be the family of all subsets of $[n]$ including the empty set. If the summation range in $\sum_I$ is not specified (for brevity), this means $\sum_{I\in\mathcal{I}}$. Throughout we assume the conventions that $\sum_{i\in\varnothing} a_i = 0$, $\sum_a^b a_i = \sum_{a\le i\le b} a_i$ for any $a_i, a, b \in \mathbb{R}$ and $0\log(c/0) = 0$ (hence, $(c/0)^0 = 1$) for any $c > 0$. Let $\theta^2_{(1)} \le \theta^2_{(2)} \le \dots \le \theta^2_{(n)}$ and $\theta^2_{[1]} \ge \theta^2_{[2]} \ge \dots \ge \theta^2_{[n]}$ (so that $\theta_{(i)} = \theta_{[n+1-i]}$) be the ordered values of $\theta_1^2, \dots, \theta_n^2$. To have some quantity well defined in the sequel, introduce also $0 = \theta^2_{(0)} = \theta^2_{[n+1]}$ and $\theta^2_{[0]} = \theta^2_{(n+1)} = \infty$.

If random quantities appear in a relation, this relation should be understood in P_θ-almost sure sense. Throughout $\phi(x, \mu, \sigma^2)$ will be the density of $\mu + \sigma Z \sim \mathrm{N}(\mu, \sigma^2)$ at point x, where $Z \sim \mathrm{N}(0, 1)$. By convention, $\mathrm{N}(\mu, 0) = \delta_\mu$ denotes a Dirac measure at point μ. We use the usual o, O notation to describe the asymptotic behavior of certain quantities as $n \to \infty$. The symbol $\triangleq$ will refer to equality by definition. Finally, let $\langle x, y\rangle = \sum_i x_i y_i$ denote the usual scalar product between $x, y \in \mathbb{R}^n$.

2.2. Multivariate normal prior

When deriving all the posterior quantities in the Bayesian analysis below, we will use the normal likelihood $\ell(\theta, X) = (2\pi\sigma^2)^{-n/2}\exp\{-\|X - \theta\|^2/2\sigma^2\}$, which is equivalent to imposing the classical high-dimensional normal model $X = (X_i, i \in \mathbb{N}_n) \sim \bigotimes_{i=1}^n \mathrm{N}(\theta_i, \sigma^2)$. Recall however that the "true" model $X \sim \mathrm{P}_\theta$ is not assumed to be normal, but only satisfying Condition (A1).

To model possible sparsity in the parameter θ, the coordinates of θ can be split into two distinct groups of coordinates of θ: for some $I \in \mathcal{I}$, the group of coordinates $\theta_{I^c} = (\theta_i, i \notin I)$ consists of (almost) zeros and $\theta_I = (\theta_i, i \in I)$ is the group of non-zeros coordinates. For any $\theta \in \mathbb{R}^n$ (even "not sparse" one) there is the best (oracle) splitting into two groups, we will come back to this in Section 3. To model sparsity, we propose a prior on θ given I as follows:

$$\pi_I = \bigotimes_{i=1}^n \mathrm{N}\big(\mu_i(I), \tau_i^2(I)\big), \quad \mu_i(I) = \mu_i \mathbb{1}\{i \in I\}, \tau_i^2(I) = \sigma^2 K_n(I)\mathbb{1}\{i \in I\}, \tag{3}$$

and $K_n(I) = (\frac{en}{|I|} - 1)\mathbb{1}\{I \ne \varnothing\}$. The indicators in prior (3) ensure the sparsity of the group I^c. The rather specific choice of $K_n(I)$ is made for the sake of concise expressions in later calculations, many other choices are actually possible. By using normal likelihood $\ell(\theta, X) = (2\pi\sigma^2)^{-n/2}\exp\{-\|X - \theta\|^2/2\sigma^2\}$, the corresponding posterior distribution for θ is readily obtained:

$$\pi_I(\vartheta|X) = \bigotimes_{i=1}^n \mathrm{N}\bigg(\frac{\tau_i^2(I)X_i + \sigma^2\mu_i(I)}{\tau_i^2(I) + \sigma^2}, \frac{\tau_i^2(I)\sigma^2}{\tau_i^2(I) + \sigma^2}\bigg). \tag{4}$$

Next, introduce the prior λ on $\mathcal{I}$, discussed in Supplement [6]. For $\varkappa > 1$, draw a random set from $\mathcal{I}$ with probabilities

$$\lambda_I = c_{\varkappa,n} \exp\Big\{-\varkappa|I|\log\Big(\frac{en}{|I|}\Big)\Big\} = c_{\varkappa,n}\Big(\frac{en}{|I|}\Big)^{-\varkappa|I|}, \quad I \in \mathcal{I}, \tag{5}$$

where $c_{\varkappa,n}$ is the normalizing constant. Since $\big(\frac{n}{k}\big)^k \le \binom{n}{k} \le \big(\frac{en}{k}\big)^k$ and $\binom{n}{0} = 1$,

$$1 = \sum_{I\in\mathcal{I}} \lambda_I = c_{\varkappa,n} \sum_{k=0}^{n} \binom{n}{k}\Big(\frac{en}{k}\Big)^{-\varkappa k} \le c_{\varkappa,n} \sum_{k=0}^{n} \Big(\frac{en}{k}\Big)^{-(\varkappa-1)k} \le c_{\varkappa,n} \sum_{k=0}^{n} e^{-(\varkappa-1)k}, \tag{6}$$

so that $c_{\varkappa,n} \ge 1 - e^{1-\varkappa} > 0$, $n \in \mathbb{N}$. Combining (3) and (5) gives the mixture prior on θ: $\pi = \sum_{I\in\mathcal{I}} \lambda_I \pi_I$. This leads to the marginal distribution of X: $\mathrm{P}_X = \sum_{I\in\mathcal{I}} \lambda_I \mathrm{P}_{X,I}$, with $\mathrm{P}_{X,I} = \bigotimes_{i=1}^n \mathrm{N}(\mu_i(I), \sigma^2 + \tau_i^2(I))$, and the posterior of θ is

$$\pi(\vartheta|X) = \pi_\varkappa(\vartheta|X) = \sum_{I\in\mathcal{I}} \pi_I(\vartheta|X)\pi(I|X), \tag{7}$$

where $\pi_I(\vartheta|X)$ is defined by (4) and the posterior $\pi(I|X)$ for I is

$$\pi(I|X) = \frac{\lambda_I \mathrm{P}_{X,I}}{\mathrm{P}_X} = \frac{\lambda_I \prod_{i=1}^n \phi(X_i, \mu_i(I), \sigma^2 + \tau_i^2(I))}{\sum_{J\in\mathcal{I}} \lambda_J \prod_{i=1}^n \phi(X_i, \mu_i(J), \sigma^2 + \tau_i^2(J))}. \tag{8}$$

2.3. Empirical Bayes posterior

The parameters μ_i are yet to be chosen in the prior. We choose μ_i by using empirical Bayes approach. The marginal likelihood P_X is readily maximized with respect to μ_i: $\tilde{\mu}_i = X_i$, which we then substitute instead of μ_i in the expression (7) for $\pi(\vartheta|X)$, obtaining the *empirical Bayes model averaging* (EBMA) posterior

$$\tilde{\pi}(\vartheta|X) = \tilde{\pi}_\varkappa(\vartheta|X) = \sum_{I\in\mathcal{I}} \tilde{\pi}_I(\vartheta|X)\tilde{\pi}(I|X), \tag{9}$$

where the empirical Bayes conditional posterior (recall that $\mathrm{N}(0,0) = \delta_0$)

$$\tilde{\pi}_I(\vartheta|X) = \bigotimes_{i=1}^n \mathrm{N}\Big(X_i \mathbb{1}\{i \in I\}, \tfrac{K_n(I)\sigma^2 \mathbb{1}\{i\in I\}}{K_n(I)+1}\Big) \tag{10}$$

is obtained from (4) with $\mu_i(I) = X_i \mathbb{1}\{i \in I\}$, and

$$\tilde{\pi}(I|X) = \frac{\lambda_I \mathrm{P}_{X,I}}{\sum_{J\in\mathcal{I}} \lambda_J \mathrm{P}_{X,J}} = \frac{\lambda_I \prod_{i=1}^n \phi(X_i, X_i\mathbb{1}\{i\in I\}, \sigma^2 + \tau_i^2(I))}{\sum_{J\in\mathcal{I}} \lambda_J \prod_{i=1}^n \phi(X_i, X_i\mathbb{1}\{i\in J\}, \sigma^2 + \tau_i^2(J))} \tag{11}$$

is the empirical Bayes posterior for $I \in \mathcal{I}$, obtained from (8) with $\mu_i(I) = X_i \mathbb{1}\{i \in I\}$. Let $\tilde{\mathrm{E}}$ and $\tilde{\mathrm{E}}_I$ be the expectations with respect to the measures $\tilde{\pi}(\vartheta|X)$ and $\tilde{\pi}_I(\vartheta|X)$, respectively. Introduce the EBMA estimator: with $X(I) \triangleq (X_i \mathbb{1}\{i \in I\}, i \in [n])$,

$$\tilde{\theta} = \tilde{\mathrm{E}}(\vartheta|X) = \sum_{I \in \mathcal{I}} \tilde{\mathrm{E}}_I(\vartheta|X)\tilde{\pi}(I|X) = \sum_{I \in \mathcal{I}} X(I)\tilde{\pi}(I|X). \tag{12}$$

Consider an alternative empirical Bayes posterior. First, derive an empirical Bayes variable selector $\hat{I}$ by maximizing $\tilde{\pi}(I|X)$ over $I \in \mathcal{I}$ (any maximizer will do) as follows:

$$\begin{aligned}
\hat{I} &= \arg\max_{I \in \mathcal{I}} \tilde{\pi}(I|X) = \arg\max_{I \in \mathcal{I}} \lambda_I \mathrm{P}_{X,I} \\
&= \arg\max_{I \in \mathcal{I}} \Big\{ -\sum_{i \in I^c} \frac{X_i^2}{2\sigma^2} - \frac{|I|}{2} \log\big(K_n(I)+1\big) + \log \lambda_I \Big\} \\
&= \arg\min_{I \in \mathcal{I}} \Big\{ \sum_{i \in I^c} X_i^2 + (2\varkappa+1)\sigma^2 |I| \log\Big(\frac{en}{|I|}\Big) \Big\} = \big\{ i \in [n] : X_i^2 \geq X_{[\hat{k}]}^2 \big\},
\end{aligned} \tag{13}$$

where $\hat{k} \in [n]_0$ is the minimizer of $\sum_{i=k+1}^n X_{[i]}^2 + (2\varkappa+1)\sigma^2 k \log(en/k)$. This is reminiscent of the penalization procedure from [8] (cf. also [1]). Now plugging in $\hat{I}$ into $\tilde{\pi}_I(\vartheta|X)$ defined by (10) yields *empirical Bayes model selection* (EBMS) posterior and the corresponding EBMS estimator:

$$\check{\pi}(\vartheta|X) = \tilde{\pi}_{\hat{I}}(\vartheta|X), \quad \check{\theta} = \check{\mathrm{E}}(\vartheta|X) = X(\hat{I}) = \big(X_i \mathbb{1}\{i \in \hat{I}\}, i \in [n]\big), \tag{14}$$

where $\check{\mathrm{E}}$ denotes the expectation with respect to the measure $\check{\pi}(\vartheta|X)$.

2.4. Exchangeable exponential moment condition on the errors

The following condition (called *exchangeable exponential moment condition*) on the error vector $\xi = (\xi_1, \ldots, \xi_n)$ will be assumed throughout.

Condition (A1). The random variables ξ_i's from (1) satisfy: $\mathrm{E}_\theta \xi_i = 0$, $\mathrm{Var}_\theta(\xi_i) \leq C_\xi$, $i \in [n]$; and for some $\beta, B > 0$ (without loss of generality assume $C_\xi = 1$ and $\beta \in (0, 1]$),

$$\mathrm{E}_\theta \exp\Big\{\beta \sum_{i \in I} \xi_i^2\Big\} \leq e^{B|I|} \quad \text{for all } I \in \mathcal{I},\ \theta \in \mathbb{R}^n. \tag{A1}$$

The unknown distribution of ξ may depend on θ, in that case (A1) is fulfilled for all $\theta \in \mathbb{R}^n$. The constants $\beta \in (0, 1]$ and $B > 0$ will be fixed in the sequel and we omit the dependence on these constants in all further notation. There is no need to assume $\mathrm{Var}_\theta(\xi_i) \leq C_\xi$ as this follows from (A1), but we provide this just for reader's convenience.

In the proof of Theorem 1 below, we will also need a bound for $\mathrm{E}_\theta\big(\sum_{i\in I}\xi_i^2\big)^2$, $I\in\mathcal{I}$. Condition (A1) ensures such a bound. Indeed, since $x^2\le e^{2x}$ for all $x\ge 0$, by using the Hölder inequality and (A1), we derive that for any $\rho\in(0,B/2]$ and $I\in\mathcal{I}$,

$$\begin{aligned}\mathrm{E}_\theta\Big(\sum_{i\in I}\xi_i^2\Big)^2 &\le \frac{B^2}{(\beta\rho)^2}\mathrm{E}_\theta e^{(2\beta\rho/B)\sum_{i\in I}\xi_i^2}\\ &\le \frac{B^2}{(\beta\rho)^2}\Big(\mathrm{E}_\theta e^{\beta\sum_{i\in I}\xi_i^2}\Big)^{2\rho/B}\le \frac{B^2}{(\beta\rho)^2}e^{2\rho|I|}.\end{aligned}\tag{15}$$

It is interesting to relate Condition (A1) to the so called *sub-Gaussianity* condition on the error vector ξ. The random vector ξ is called *sub-Gaussian* with parameter $\rho>0$ if

$$\mathrm{P}\big(\big|\langle v,\xi\rangle\big|>t\big)\le e^{-\rho t^2}\quad\text{for all } t\ge 0 \text{ and all } v\in\mathbb{R}^n \text{ such that } \|v\|=1.\tag{16}$$

The sub-Gaussianity condition (16) and Condition (A1) are close, but in general incomparable. For example, let $\xi_i=\xi_0$, $i\in[n]$, for some bounded random variable ξ_0 (say, uniform on $[-1,1]$), then Condition (A1) trivially holds, whereas the sub-Gaussianity condition is not fulfilled. It is easy to see that the sub-Gaussianity condition is equivalent to Condition (A1) for independent ξ_i's.

Condition (A1) is clearly satisfied for independent normals $\xi_i\overset{\text{ind}}{\sim}\mathrm{N}(0,1)$ (with $\beta=0.4$ and $B=1$) and for bounded, arbitrarily dependent, ξ_i's. In a way, Condition (A1) prevents too much dependence, but it still allows some interesting cases of dependent ξ_i's. Suppose that the ξ_i's follow an autoregressive model AR(1) with normal white noise:

$$\xi_k=\alpha\xi_{k-1}+\epsilon_k,\quad \epsilon_k\overset{\text{ind}}{\sim}\mathrm{N}(0,1),\ k\in[n];\quad \xi_0=0,\ |\alpha|<1.\tag{17}$$

Let us show that Condition (A1) is fulfilled for the vector $\xi=(\xi_i,i\in[n])$ defined by (17). We have that for any $k>l$, $\xi_k=\alpha^{k-l}\xi_l+\alpha^{k-l-1}\epsilon_{l+1}+\cdots+\epsilon_k=\alpha^{k-l}\xi_l+Z_{k-l}$, where $Z_{k'}\sim\mathrm{N}(0,\sigma_{k'}^2)$ with $\sigma_{k'}^2=1+\alpha^2+\cdots+\alpha^{2(k'-1)}\le\frac{1}{1-\alpha^2}\triangleq\sigma_0^2$. Clearly, for any $I\in\mathcal{I}$, there are $1\le k_1<k_2<\cdots<k_{|I|}\le n$ such that $\sum_{i\in I}\xi_i^2=\sum_{i=1}^{|I|}\xi_{k_i}^2$. Denote $\mathcal{F}_m=\sigma(\xi_{k_i},1\le i\le m)$, $m\in[|I|]$, the σ-algebra generated by $\{\xi_{k_i},1\le i\le m\}$. Choose β and α in such a way that $0<\frac{2\alpha^2}{1-4\beta\sigma_0^2}\le 1$. By using the elementary identity (S4) from Supplement [6], we first evaluate the conditional expectation

$$\begin{aligned}\mathrm{E}\big(e^{\beta(\xi_{k_{m-1}}^2+\xi_{k_m}^2)}|\mathcal{F}_{m-1}\big)&=e^{\beta\xi_{k_{m-1}}^2}\mathrm{E}\big(e^{\beta\xi_{k_m}^2}|\mathcal{F}_{m-1}\big)\le e^{\beta\xi_{k_{m-1}}^2}\mathrm{E}\big(e^{2\beta\xi_{k_m}^2}|\mathcal{F}_{m-1}\big)\\ &=\exp\Big\{\Big(\beta+\frac{2\beta\alpha^{2(k_m-k_{m-1})}}{1-4\beta\sigma_{k_m-k_{m-1}}^2}\Big)\xi_{k_{m-1}}^2-\frac{1}{2}\log\big(1-4\beta\sigma_{k_m-k_{m-1}}^2\big)\Big\}\\ &\le\frac{e^{2\beta\xi_{k_{m-1}}^2}}{(1-4\beta\sigma_0^2)^{1/2}}.\end{aligned}$$

Iterating the above argument, we establish Condition (A1) for the sequence (17):

$$\mathrm{E}\exp\Big\{\beta\sum_{i\in I}\xi_i^2\Big\} = \mathrm{E}\mathrm{E}\Big[\exp\Big\{\beta\sum_{i\in I}\xi_i^2\Big\}\big|\mathcal{F}_{|I|-1}\Big] = \frac{\mathrm{E}\Big[\exp\big\{\beta\sum_{i=1}^{|I|-2}\xi_{k_i}^2\big\}e^{2\beta\xi_{k_{|I|-1}}^2}\Big]}{(1-4\beta\sigma_0^2)^{1/2}}$$
$$\le\cdots\le\big(1-4\beta\sigma_0^2\big)^{-|I|/2} = e^{B|I|},\quad \text{with } B=\log\big(1-4\beta\sigma_0^2\big)^{-1/2}.$$

3. Main results

In this section, we give the main results of the paper. From now on, by $\hat{\pi}(\vartheta|X)$ (with corresponding expectation $\hat{\mathrm{E}}(\cdot|X)$) we denote either EBMA posterior $\tilde{\pi}(\vartheta|X)$ defined by (9) or EBMS posterior $\check{\pi}(\vartheta|X)$ defined by (14), and $\hat{\theta}$ will stand either for $\tilde{\theta}$ defined by (12) or for $\check{\theta}$ defined by (14). Also, $\hat{\pi}(I\in\mathcal{G}|X)$ should be read as $\tilde{\pi}(I\in\mathcal{G}|X)$ in case $\hat{\pi}=\tilde{\pi}$, and as $\mathbb{1}\{\hat{I}\in\mathcal{G}\}$ in case $\hat{\pi}=\check{\pi}$, for all $\mathcal{G}\subseteq\mathcal{I}$ that appear in what follows. Hence, $\hat{\pi}(I|X)=\tilde{\pi}(I|X)$ and $\mathrm{E}_\theta\hat{\pi}(I\in\mathcal{G}|X)=\mathrm{E}_\theta\tilde{\pi}(I\in\mathcal{G}|X)$ in the former case, and $\hat{\pi}(I|X)=\mathbb{1}\{\hat{I}=I\}$ and $\mathrm{E}_\theta\hat{\pi}(I\in\mathcal{G}|X)=\mathrm{P}_\theta(\hat{I}\in\mathcal{G})$ in the latter case.

3.1. Oracle rate

The empirical Bayes posterior $\hat{\pi}(\vartheta|X)$ is a random mixture over $\tilde{\pi}_I(\vartheta|X)$, $I\in\mathcal{I}$. From the P_θ-perspective, each posterior $\tilde{\pi}_I(\vartheta|X)$ (and the corresponding estimator $\tilde{\mathrm{E}}_I(\vartheta|X)=X(I)$) contracts to the true parameter θ with the local rate $R^2(I,\theta)=\sum_{i\in I^c}\theta_i^2+\sigma^2|I|$. Indeed, since $\tilde{\mathrm{E}}_I(\vartheta|X)=X(I)=(X_i\mathbb{1}\{i\in I\}, i\in[n])$, (10) and the Markov inequality yields

$$\mathrm{E}_\theta\tilde{\pi}_I\big(\|\vartheta-\theta\|^2\ge M^2R^2(I,\theta)|X\big)\le\frac{\mathrm{E}_\theta\|X(I)-\theta\|^2+\frac{K_n(I)\sigma^2|I|}{K_n(I)+1}}{M^2R^2(I,\theta)}\le\frac{2}{M^2}.$$

For each $\theta\in\mathbb{R}^n$, among $I\in\mathcal{I}$ there exists the best choice $I_o=I_o(\theta)=I_o(\theta,\sigma)$ (called the *R-oracle*) corresponding to the fastest local rate $R^2(\theta)=R^2(\theta,\mathcal{I})=\min_{I\in\mathcal{I}}R^2(I,\theta)=\sum_{i\in I_o^c}\theta_i^2+\sigma^2|I_o|$. Ideally, we would like to *mimic* the R-oracle, that is, to construct an empirical Bayesian procedure (e.g., $\hat{\pi}(\vartheta|X)$) which performs as good as the oracle empirical Bayes posterior $\tilde{\pi}_{I_o}(\vartheta|X)$ without knowing I_o, uniformly in $\theta\in\mathbb{R}^n$. However, the lower bounds for the estimation problem (hence, also for the posterior contraction problem), obtained by [15] and later by [8], show that it is impossible to mimic the R-oracle and a logarithmic factor is the unavoidable price for the uniformity over $\mathbb{R}^n$ (otherwise this would contradict to the minimax lower bound over the scale of sparsity classes, cf. [8]). Therefore, only a modification of the risk R-oracle where the variance term $\sigma^2|I_o|$ is inflated with the factor $\log(en/|I_o|)$ (thought of as payment for not knowing I_o) is “mimicable”.

The above discussion motivates the following definition. Introduce the family of local rates

$$r^2(I,\theta)=r^2\big(I,\theta,\sigma^2\big)=\sum_{i\in I^c}\theta_i^2+\sigma^2|I|\log\Big(\frac{en}{|I|}\Big)=B(I,\theta)+V(I),\quad I\in\mathcal{I},\tag{18}$$

where $B(I,\theta)=\sum_{i\in I^c}\theta_i^2$ is the bias part of the rate and $V(I)=V(I,\sigma,n)=\sigma^2|I|\log\big(\frac{en}{|I|}\big)$ is the adjusted variance part, the variance term $\sigma^2|I|$ of the rate $R(I,\theta)$ multiplied by the logarithmic factor $\log\big(\frac{en}{|I|}\big)$. There exists the best choice $I_o=I_o(\theta)=I_o(\theta,\sigma^2)=I_o(\theta,\sigma^2,n)\in\mathcal{I}$ (called *oracle*) at which the rate (18) is minimal:

$$r^2(\theta)=r^2_{\mathcal{I}}(\theta)=\min_{I\in\mathcal{I}}r^2(I,\theta)=r^2(I_o,\theta)=B(I_o,\theta)+V(I_o), \tag{19}$$

$r^2(\theta)$ is called the *oracle rate*. Note that the oracle I_o may not be unique (but $|I_o|$ is unique) as some coordinates of θ can coincide, in that case take the one with the earliest coordinates.

Clearly, $I_o=\big\{i\in[n]:\theta_i^2\ge\theta^2_{[k_o]}\big\}$, where $k_o=|I_o|=\arg\min_{k\in[n]_0}\big\{\sum_{i=1}^{n-k}\theta^2_{(i)}+\sigma^2k\log\big(\frac{en}{k}\big)\big\}$. Thus, the oracle I_o classifies the coordinates $(\theta_i, i\in I_o)$ as *significant* and the coordinates $(\theta_i, i\in I_o^c)$ as *insignificant*. The bias part $B(I_o,\theta)=\sum_{i\in I_o^c}\theta_i^2=\sum_{i=k_o+1}^{n}\theta^2_{[i]}=\sum_{i=1}^{n-k_o}\theta^2_{(i)}$ of the oracle rate is called the *excessive bias*. This is the error the oracle makes when setting insignificant coordinates of θ to zero. The "variance" term $V(I_o)=\sigma^2|I_o|\log(\frac{en}{|I_o|})$ is the error the oracle makes when recovering $|I_o|$ significant coordinates, with the log factor as payment for not knowing the locations.

Finally, for $\tau\ge0$, introduce the so called τ*-oracle* $I_o^\tau=I_o^\tau(\theta)=I_o(\theta,\tau\sigma^2)$ and let $i_\tau=|I_o^\tau(\theta)|$ be the corresponding cardinality. A τ-oracle $I_o^\tau(\theta)$ is just the usual oracle defined by (19) with σ^2 substituted by $\tau\sigma^2$, the oracle itself is the τ-oracle with $\tau=1$: $I_o(\theta)=I_o^1(\theta)$ and $i_1=k_o$. Notice that $I_o^{\tau_1}\subseteq I_o^{\tau_2}$ if $\tau_1\ge\tau_2$. For $\tau\downarrow0$, the "limiting" τ-oracle recovers the "true" active index set $I^*=I^*(\theta)=\mathrm{supp}(\theta)\triangleq\{i\in[n]:\theta_i\ne0\}$ in the sense that $I_o^\tau\uparrow I^*$ as $\tau\downarrow0$. More insight on the τ-oracles I_o^τ, $\tau\ge0$, is provided in Supplement [6].

3.2. Contraction results with oracle rate

First, introduce a technical condition on the parameter $\varkappa$ appearing in (5).

Condition (A2). The parameter $\varkappa$ of the prior λ defined by (5) satisfies

$$\varkappa>\bar{\varkappa}\triangleq(12-\beta+4B)/(4\beta), \tag{A2}$$

where β, B are from Condition (A1).

The following theorem establishes that the empirical Bayes posterior $\hat\pi(\vartheta|X)$ contracts to θ with the oracle rate $r(\theta)$ from the frequentist P_θ-perspective, and the empirical Bayes posterior mean $\hat\theta$ converges to θ with the oracle rate $r(\theta)$, uniformly over the entire parameter space.

Theorem 1. *Let Conditions* (A1) *and* (A2) *be fulfilled. Then there exist positive constants* M_0, M_1, H_0, H_1, m_0, m_1 *such that for any* $\theta\in\mathbb{R}^n$ *and any* $M\ge0$,

$$\mathrm{E}_\theta\hat\pi\big(\|\vartheta-\theta\|^2\ge M_0r^2(\theta)+M\sigma^2|X\big)\le H_0e^{-m_0M}, \tag{i}$$

$$\mathrm{P}_\theta\big(\|\hat\theta-\theta\|^2\ge M_1r^2(\theta)+M\sigma^2\big)\le H_1e^{-m_1M}. \tag{ii}$$

Remark 1. Notice that already claim (i) of the above theorem contains an oracle bound for the estimator $\hat{\theta}$. Indeed, by Jensen's inequality, we derive the oracle inequality

$$\begin{aligned} \mathrm{E}_\theta \|\hat{\theta}-\theta\|^2 &\le \mathrm{E}_\theta \hat{\mathrm{E}}\big(\|\vartheta-\theta\|^2|X\big) \\ &\le M_0 r^2(\theta) + H_0 \int_0^{+\infty} e^{-m_0 u/\sigma^2}\,du = M_0 r^2(\theta) + \frac{H_0\sigma^2}{m_0}. \end{aligned} \tag{20}$$

Similarly we can show that also (ii) implies (20). This means that claim (ii) is actually stronger (and more refined) than (20) and therefore requires a separate proof.

The constants $M_0, M_1, H_0, H_1, m_0, m_1 > 0$ in the theorem depend only on β, B and some also on $\varkappa$, the exact expressions can be found in the proof. The above local result implies the minimax optimality over various sparsity scales, see Section 3.5 for more detail on this. In case of independent normal errors, some constants and bounds in the proofs can be sharpened; possible refinements are discussed in Supplement [6].

Remark 2. The non-asymptotic exponential bounds in terms of the constant M from the expression $M'r^2(\theta) + M\sigma^2$ (with some fixed M') in claims (i) and (ii) of the theorem provide a rather refined characterization of the quality of the posterior $\hat{\pi}(\vartheta|X)$ and estimator $\hat{\theta}$, finer than, for example, the traditional oracle inequalities like (20). This refined formulation allows for subtle analysis in various asymptotic regimes ($n \to \infty$, $\sigma \to 0$, or their combination) as we can let M depend in any way on n, σ, or both.

One more technical definition will be used. For constants β, B from Condition (A1), define

$$\bar{\tau} = \bar{\tau}(\varkappa, \beta, B) \triangleq \frac{3(\varkappa\beta + \beta/2 + B)}{\beta}. \tag{21}$$

The next theorem describes the frequentist behavior of the selector $\hat{I}$ and the empirical Bayes posterior for I, saying basically that $\hat{I}$ and $\tilde{\pi}(I|X)$ "live" on a certain set that is, in a sense, almost as good as the oracle $I_o = I_o(\theta)$ defined by (19). For any $\theta \in \mathbb{R}^n$, introduce

$$\bar{I}_o = \bar{I}_o(\theta) \triangleq I_o^{\tau_0}(\theta) = I_o\big(\theta, \tau_0\sigma^2\big), \tag{22}$$

where we fix some $\varrho \in (0,1)$ and $\tau_0 \ge \frac{1+\varrho}{1-\varrho}\bar{\tau}$, with $\bar{\tau}$ defined by (21). For example, we can take $\varrho = 0.1$ and $\tau_0 = \frac{11}{9}\bar{\tau} + 0.1$.

Theorem 2. *Let Condition* (A1) *be fulfilled. The following relations hold for any* $\theta \in \mathbb{R}^n$, $M \ge 0$.

(i) *Let* $\varkappa > \frac{4+\beta+2B}{2\beta}$ (*Condition* (A2) *implies this*). *There exist* $M_0', H_0' > 0$ *such that*

$$\mathrm{E}_\theta \hat{\pi}\bigg(I \in \mathcal{I} : |I|\log\Big(\frac{en}{|I|}\Big) \ge M_0'|I\cap I_o|\log\Big(\frac{en}{|I\cap I_o|}\Big) + M\Big|X\bigg) \le H_0' e^{-M}.$$

(ii) *Let* $\varkappa > \beta^{-1} - \frac{1}{2}$ *(Condition (A2) implies this),* $\bar{\tau}$ *be defined by (21). Fix any* $I' \in \mathcal{I}$*. Then there exist* $H'_1, m'_0 > 0$ *(independent of* θ *and* I'*) such that*

$$\mathrm{E}_\theta \hat{\pi}\Big(I \in \mathcal{I} : \sum_{i \in I' \setminus I} \frac{\theta_i^2}{\sigma^2} \geq \bar{\tau}\big|I \cup I'\big| \log\Big(\frac{en}{|I \cup I'|}\Big) + M\big|X\Big) \leq H'_1 e^{-m'_0 M}.$$

In particular, let $\bar{I}_o$ *be defined by (22), then there exists* $m'_1 > 0$ *such that*

$$\mathrm{E}_\theta \hat{\pi}\Big(I \in \mathcal{I} : |I| \log\Big(\frac{en}{|I|}\Big) \leq \varrho |\bar{I}_o| \log\big(en/|\bar{I}_o|\big) - M\big|X\Big) \leq H'_1 e^{-m'_1 M}. \tag{23}$$

(iii) *Let Condition (A2) be fulfilled,* c_1, c_2, c_3 *be the constants defined in Lemma 2. Then*

$$\mathrm{E}_\theta \hat{\pi}\big(I \in \mathcal{I} : r^2(I, \theta) \geq c_3 r^2(\theta) + M\sigma^2 |X\big) \leq C_0 e^{-c_2 M}, \quad \textit{where } C_0 = \big(1 - e^{1-c_1}\big)^{-1}.$$

We can interpret the above theorem as recovery of the sparsity structure, but in a weak sense. Claim (iii) says that $\hat{\pi}(I|X)$ (i.e., the selector $\hat{I}$ and the posterior $\tilde{\pi}(I|X)$) "lives" on those sets I whose rate $r^2(I,\theta)$ is not far from the oracle rate $r^2(\theta)$. Claims (i) and (ii) roughly mean that $\hat{\pi}(I|X)$ "lives" in a shell between $\bar{I}_o$ and I_o (recall that $\bar{I}_o = I_o^{\tau_0} \subseteq I_o^1 = I_o$ as $\tau_0 > 1$, hence always $|\bar{I}_o| \leq |I_o|$). If $\bar{I}_o$ and I_o are "close" to each other (in the sense that $c|I_o| \leq |\bar{I}_o| \leq |I_o|$ for some $c \in (0,1]$), then $\hat{\pi}(I|X)$ recovers well the oracle structure I_o. In general, the living shell for $\hat{\pi}(I|X)$ can be too wide (when $|\bar{I}_o|$ is much smaller than $|I_o|$), the corresponding θ is then "deceptive". We discuss this below, when introducing the EBR condition.

3.3. Confidence ball under excessive bias restriction (EBR)

Theorem 1 establishes the strong local optimal properties of the empirical Bayes posterior $\hat{\pi}(\vartheta|X)$ and the empirical Bayes posterior mean $\hat{\theta}$, but does not solve the uncertainty quantification problem yet. As is mentioned in the introduction, the uniform coverage and optimal size properties can not hold together in general. This is not an artifact of the method, it is a fundamental, unavoidable problem. It occurs for the so called deceptive parameters θ that have many smallish coordinates, just slightly under the noise level. Clearly, in this case no method can reliably assign those coordinates to the significant set which is needed to control the bias part of the oracle rate. This is possible for the non-deceptive parameters described by the so called EBR condition introduced below.

First, we propose a confidence ball by using the empirical Bayes posterior $\check{\pi}(\vartheta|X)$ defined by (14) (one can construct a confidence ball by using the posterior $\tilde{\pi}(\vartheta|X)$ with similar properties, but with more involved mathematical derivations). Since $\check{\pi}(\vartheta|X) = \bigotimes_{i=1}^n \mathrm{N}(\check{\theta}_i, \check{\sigma}_i^2)$ with $\check{\theta}_i = X_i \mathbb{1}\{i \in \hat{I}\}$ and $\check{\sigma}_i^2 = (1 - |\hat{I}|/en)\sigma^2 \mathbb{1}\{i \in \hat{I}\}$, denoting by $\chi^2_{k,\alpha}$ the $(1-\alpha)$-quantile of χ^2_k-distribution, we have

$$\check{\pi}\Big(\|\vartheta - \check{\theta}\|^2 \leq \sigma^2 \chi^2_{|\hat{I}|,\alpha}\big|X\Big) \geq \check{\pi}\Big(\|\vartheta - \check{\theta}\|^2 \leq \big(1 - |\hat{I}|/en\big)\sigma^2 \chi^2_{|\hat{I}|,\alpha}\big|X\Big) = 1 - \alpha.$$

As $\chi^2_{|\hat{I}|,\alpha} \le M_\alpha|\hat{I}|$ for sufficiently large M_α, a reasonable candidate confidence ball is $B(\check{\theta}, M\sigma|\hat{I}|^{1/2})$. The empirical Bayes posterior $\check{\pi}(\vartheta|X)$ is well concentrated (in fact, in a ball of the size $M\sigma^2|I_o|$), but not around the truth, rather around its mean $\check{\theta}$. Then $B\big(\check{\theta}, M\sigma|\hat{I}|^{1/2}\big)$ cannot have a guaranteed coverage, since otherwise the center $\check{\theta}$ would mimics the R-oracle uniformly in $\theta \in \mathbb{R}^n$, which is impossible as we discussed earlier. To obtain coverage, the radius of any confidence ball must contain a logarithmic factor. This leads us to the inflated credible ball $B(\check{\theta}, M\hat{r})$, where

$$\hat{r}^2 = \hat{r}^2(X) = \sigma^2 + \sigma^2|\hat{I}|\log\big(en/|\hat{I}|\big). \tag{24}$$

According to Theorem 2, $\varrho\sigma^2|\bar{I}_o|\log(en/|\bar{I}_o|) - M\sigma^2 \le \hat{r}^2 \le M_0'\sigma^2|I_o|\log(en/|I_o|) + M\sigma^2$ with large probability, describing a "living shell" for $\hat{r}$. So, $\hat{r}^2$ is at most of the order of the oracle rate $r^2(\theta)$, implying the size property with the oracle radial rate uniformly over $\theta \in \mathbb{R}^n$. But if $|\bar{I}_o|$ is much smaller than $|I_o|$, the living shell for $\hat{r}$ becomes too wide and the coverage property of $B(\check{\theta}, M\hat{r})$ cannot be guaranteed because $\hat{r}$ can be over optimistically too small. This problem will not occur for those (non-deceptive) θ's for which the bias part $B(\bar{I}_o, \theta)$ of the rate $r^2(\bar{I}_o, \theta)$ (see definition (18)) is within a multiple of its variance part $V(\bar{I}_o) = \sigma^2|\bar{I}_o|\log(en/|\bar{I}_o|)$. Then $\sigma^2|\bar{I}_o|\log(en/|\bar{I}_o|)$ is at least of the oracle rate order, which, together with (ii) of Theorem 2, imply that $\hat{r}$ is also at least of the oracle rate order, resulting in a good coverage of the confidence ball $B\big(\check{\theta}, M_2\hat{r} + M\sigma^2\big)$ for some M_2 and sufficiently large M. This discussion motivates introducing the following condition.

Condition EBR. We say that $\theta \in \mathbb{R}^n$ satisfies the *excessive bias restriction* (EBR) condition with structural parameter $t \ge 0$ if $\theta \in \Theta_{\mathrm{eb}}(t)$, where the corresponding set (called the *EBR class*) is

$$\Theta_{\mathrm{eb}}(t) = \Theta_{\mathrm{eb}}(t, \tau_0) = \Big\{\theta \in \mathbb{R}^n : \sum_{i\in\bar{I}_o^c}\theta_i^2 \le t\sigma^2\big[1 + |\bar{I}_o|\log(en/|\bar{I}_o|)\big]\Big\}, \tag{25}$$

where the set $\bar{I}_o = I_o^{\tau_0}$ is defined by (22).

Remark 3. A few remarks about the EBR condition are in order.

(i) The EBR requires basically that the bias part of the rate $r^2(\bar{I}_o, \theta)$ is dominated by a multiple of its variance part (additional σ^2 is needed to handle the case $\bar{I}_o = \varnothing$). This is obviously satisfied also for the rate $r^2(I', \theta)$ with any I' such that $\bar{I}_o \subseteq I'$ (hence, also for I_o as $\bar{I}_o \subseteq I_o$).

(ii) Notice that, for any $\tau > 0$, $\Theta_{\mathrm{eb}}(t_1, \tau) \subseteq \Theta_{\mathrm{eb}}(t_2, \tau)$ for $t_1 \le t_2$, and $\Theta_{\mathrm{eb}}(\tau n, \tau) = \mathbb{R}^n$ by the oracle definition. Hence, $\mathbb{R}^n = \bigcup_{0\le t\le\tau n}\Theta_{\mathrm{eb}}(t, \tau)$, which means that the EBR condition leads to the so called *EBR-scale* $\{\Theta_{\mathrm{eb}}(t, \tau), t \ge 0\}$, discussed in detail in Supplement [6].

(iii) Ideally, the excessive τ_0-bias $B(\bar{I}_o, \theta) = \sum_{i\in\bar{I}_o^c}\theta_i^2$ is zero, which corresponds to "the least deceptive" parameters θ whose insignificant coordinates are true zeros and the significant coordinates are sufficiently distinct from zero. Such examples are given in Supplement [6].

(v) The EBR introduced in [29] and [13] is a version of our EBR (25) adopted to the sparsity scale within the grand space $\ell_0[p_n]$ with $p_n = o(n)$, under the asymptotic regime $n \to \infty$. We elaborate on the EBR (and its relation to the EBR from [29] and [13]) in Supplement [6].

(vi) Interestingly, as is shown in [13], the EBR turns out to be minimal in a certain sense.

The following theorem, which is the main result in the paper, describes the coverage and size properties of the confidence ball based on $\hat{\theta}$ (i.e., either $\check{\theta}$ or $\tilde{\theta}$) and $\hat{r}$.

Theorem 3. *Let Conditions* (A1) *and* (A2) *be fulfilled. Then there exist constants* $M_2, H_2, m_2 > 0$ *such that for any* $t, M \geq 0$, *and with* $\hat{R}^2_M = \hat{R}^2_M(M_2) = (t+1)M_2\hat{r}^2 + (t+2)M\sigma^2$,

$$\sup_{\theta\in\Theta_{\mathrm{eb}}(t)} \mathrm{P}_\theta\big(\theta \notin B(\hat{\theta}, \hat{R}_M)\big) \leq H_2 e^{-m_2 M},$$

$$\sup_{\theta\in\mathbb{R}^n} \mathrm{P}_\theta\bigg(\hat{r}^2 \geq M_0'\sigma^2|I_o|\log\Big(\frac{en}{|I_o|}\Big) + (M+1)\sigma^2\bigg) \leq H_0' e^{-M},$$

where $\Theta_{\mathrm{eb}}(t)$ *is defined by* (25), *and the constants* M_0', H_0' *are defined in Theorem* 2.

Remark 4. Recall that $\bar{I}_o = I_o^{\tau_0}$ from (25) is actually the τ_0-oracle. It may be desirable to impose the EBR condition in terms of the "standard" oracle I_o, rather than the τ_0-oracle. By rewriting (1) as $X\tau_0^{-1/2} = \theta\tau_0^{-1/2} + \sigma\tau_0^{-1/2}\xi$, it is not difficult to see that we can construct a confidence ball with the radius $\sqrt{\tau_0}\hat{R}_M$ satisfying the coverage property as above, but now uniformly over $\Theta_{\mathrm{eb}}(t, 1)$.

Remark 5. To measures the deceptiveness amount in θ, introduce the quantity $b(\theta)$:

$$\begin{aligned} b(\theta) = b(\theta,\tau_0) &= \frac{\sum_{i\in\bar{I}_o^c}\theta_i^2}{\sigma^2 + \sigma^2|\bar{I}_o|\log\big(en/|\bar{I}_o|\big)} \\ &= \frac{\sum_{i=i_1+1}^n \theta_{[i]}^2}{\sigma^2+\sigma^2 i_{\tau_0}\log(en/i_{\tau_0})} = \frac{B(\bar{I}_o,\theta)}{\sigma^2+V(\bar{I}_o,\theta)}. \end{aligned} \tag{26}$$

The EBR condition can be seen as restricting the deceptiveness $b(\theta)$: $\Theta_{\mathrm{eb}}(t) = \{\theta : b(\theta) \leq t\}$.

Note that, when proving Theorem 3, we actually established the following local assertions: there exist constants $M_2, \alpha_1, m_1'', H_2, m_2 > 0$ such that for any $\theta \in \mathbb{R}^n$ and any $\alpha, M \geq 0$

$$\begin{aligned} &\mathrm{P}_\theta\Big(\theta \notin B\Big(\hat{\theta}, \big[\big(b(\theta)+1\big)M_2\hat{r}^2 + \big(b(\theta)+2\big)M\sigma^2\big]^{1/2}\Big)\Big) \\ &\quad\leq H_1\Big(\frac{en}{|I_o|}\Big)^{-\alpha_1|I_o|}e^{-m_1M} + H_1'\Big(\frac{en}{|\bar{I}_o|}\Big)^{-\alpha_1'|\bar{I}_o|}e^{-m_1''M} \leq H_2e^{-m_2M}, \\ &\mathrm{P}_\theta\bigg(\hat{r}^2 \geq \sigma^2\big(M_0'+\alpha\big)|I_o|\log\Big(\frac{en}{|I_o|}\Big) + (M+1)\sigma^2\bigg) \leq H_0'\Big(\frac{ne}{|I_o|}\Big)^{-\alpha|I_o|}e^{-M}, \end{aligned}$$

where all the other constants (H_1, m_1, H_1', α_1', M_0', H_0') are defined in Theorems 1 and 2. Notice that the size relation in Theorem 3 holds uniformly in $\theta \in \mathbb{R}^n$. Although the coverage relation is also uniform in $\theta \in \mathbb{R}^n$, the main (and unavoidable) problem is dependence on $b(\theta)$. The mission of the EBR condition is to provide control over the quantity $b(\theta)$.

3.4. Confidence ball of $n^{1/4}$-radius without EBR

Suppose we want to construct a confidence ball of a full coverage uniformly over the whole space $\mathbb{R}^n$. Recall however that, in view of the above mentioned negative results of [3,11,20] and [22], no data dependent ball can in general provide full coverage and optimal size simultaneously. It turns out that, even when insisting on the full coverage, the size can be still optimal, but only for the parameters with the oracle rate in the range $r^2(\theta) \geq C\sqrt{n}$, i.e., for non-sparse parameters.

An idea is to mimic the quantity $\|\theta - \hat{\theta}\|^2$ by $\hat{R}^2 = \|X - \hat{\theta}\|^2$. Clearly, there is a lot of bias in $\hat{R}^2$, the biggest part of which is due to the term $\sigma^2\|\xi\|^2$ contained in $\hat{R}$. To de-bias for that part, we need to subtract its expectation $\sigma^2 \mathrm{E}\|\xi\|^2 = n\sigma^2$, where we assumed $\mathrm{Var}(\xi_i) = 1$. However, even de-biased quantity $\hat{R}^2$ can only be controlled up to the order $\sigma^2\sqrt{n}$. Thus, a term of the order $\sigma n^{1/4}$ is necessary in the radius to provide coverage uniformly over the whole space $\mathbb{R}^n$.

To handle some technical issues in this case, we impose the following additional condition.

Condition (A3). Besides X given by (1), we also observe $X' \in \mathbb{R}^n$ independent of X, where $X' = \theta + \sigma\xi'$, the random vector ξ' satisfies the following relations: $\mathrm{E}\xi_i' = 0$, $\mathrm{Var}(\xi_i') = 1$, $i \in [n]$;

$$\begin{aligned} &\mathrm{P}\Big(\big|\langle v, \xi'\rangle\big| \geq \sqrt{M}\Big) \leq \psi_1(M) \quad \forall v \in \mathbb{R}^n : \|v\| = 1; \\ &\mathrm{P}\Big(\big|\|\xi'\|^2 - \mathrm{E}\|\xi'\|^2\big| \geq M\sqrt{n}\Big) \leq \psi_2(M). \end{aligned} \tag{A3}$$

Here $\psi_1(M)$, $\psi_2(M)$ are some positive monotonically decreasing functions such that $\psi_1(M) \downarrow 0$ and $\psi_2(M) \downarrow 0$ as $M \uparrow \infty$.

Condition (A3) is satisfied for independent normals $\xi_i \overset{\text{ind}}{\sim} \mathrm{N}(0, 1)$ even if we do not have the second sample X' at our disposal. Indeed, in this case we can "duplicate" the observations by randomization at the cost of doubling the variance as follows: create samples $X' = X + \sigma Z$ and $X'' = X - \sigma Z$, for a $Z = (Z_1, \ldots, Z_n)$ (independent of X) such that $Z_i \overset{\text{ind}}{\sim} \mathrm{N}(0, 1)$. Relations (A3) are then fulfilled with exponential functions $\psi_l(M) = Ce^{-cM}$, $l = 1, 2$, for some $C, c > 0$. If the sub-Gaussianity condition (16) is fulfilled for ξ' (which is the same as Condition (A1) in case of independent ξ_i''s), then $\psi_1(M) = e^{-\rho M}$. By Chebyshev's inequality, we see that the second relation in (A3) is fulfilled with function $\psi_2(M) = cM^{-2}$ for any zero mean independent ξ_i''s with $\mathrm{E}\xi_i'^4 \leq C$.

Coming back to the problem of constructing a confidence ball of full coverage uniformly over $\mathbb{R}^n$, let $\hat{\theta}$ and $\hat{I}$ be defined as before and based on the sample X. We propose to mimic $\|\theta - \hat{\theta}\|^2$ by the de-biased quantity $\|X' - \hat{\theta}\|^2 - n\sigma^2$ plus additional $\sigma^2\sqrt{n}$-order term to control its oscillations, leading us to the following data dependent radius

$$\tilde{R}_M^2 = \Big(\big\| X' - \hat{\theta} \big\|^2 - n\sigma^2 + 2\sigma^2 G_M \sqrt{n} \Big)_+, \quad \text{where } G_M = \sqrt{M(M+M_1)}, \tag{27}$$

$x_+ = x \vee 0$ and the constant M_1 is from Theorem 1. The next theorem establishes the coverage and size properties of the confidence ball $B(\hat{\theta}, \tilde{R}_M)$. The proof is given in Supplement [6].

Theorem 4. *Let Conditions (A1)–(A3) be fulfilled, $\tilde{R}_M^2$ be defined by (27). Then for any $M \geq 0$*

$$\sup_{\theta \in \mathbb{R}^n} \mathrm{P}_\theta\big(\theta \notin B(\hat{\theta}, \tilde{R}_M)\big) \leq \psi_1(M/4) + \psi_2(M) + H_1 e^{-m_1 M},$$

$$\sup_{\theta \in \mathbb{R}^n} \mathrm{P}_\theta\big(\tilde{R}_M^2 \geq g_M(\theta, n)\big) \leq \psi_1(M/4) + \psi_2(M) + 2H_1 e^{-m_1 M},$$

$g_M(\theta, n) = M_1 r^2(\theta) + M\sigma^2 + 4\sigma^2 G_M \sqrt{n}$ and the constants H_1, m_1, M_1 are defined in Theorem 1.

By taking large enough M we can ensure the coverage and size relations uniformly over the entire space $\mathbb{R}^n$. Notice the price for this overall uniformity: the radius of the constructed confidence ball is essentially of the order $\sigma n^{1/4} + r(\theta)$, i.e., at least $\sigma n^{1/4}$ even for very sparse parameters θ. Hence, the radius is of the oracle rate order only when $r(\theta) \geq C\sigma n^{1/4}$, that is, for non-sparse θ's.

3.5. Implications: The minimax results over sparsity classes

In this section, we elucidate the potential strength of the local approach. In particular, we demonstrate how the global adaptive minimax results over certain scales can be derived from the local results. Note that the oracle rate $r(\theta)$ is a local quantity in that it quantifies the level of accuracy of inference about specific θ and originally it is not linked to any particular scale of classes. However, it is always possible to relate the oracle rate to various scales. Precisely, if we want to establish global adaptive minimax results over certain scale, say, $\{\Theta_\beta, \beta \in \mathcal{B}\}$, with corresponding minimax rates $\{r(\Theta_\beta), \beta \in \mathcal{B}\}$ (recall that the minimax rate over Θ_β is $r^2(\Theta_\beta) \triangleq \inf_{\hat{\theta}} \sup_{\theta \in \Theta_\beta} \mathrm{E}_\theta \|\hat{\theta} - \theta\|^2$, where the infimum is taken over all estimators), the only thing we need to show is

$$\sup_{\theta \in \Theta_\beta} r^2(\theta) \leq c r^2(\Theta_\beta), \quad \text{for all } \beta \in \mathcal{B}. \tag{28}$$

If the above property holds, we say the oracle rate $r(\theta)$ *covers* the scale $\{\Theta_\beta, \beta \in \mathcal{B}\}$. In this case, the local results on the estimation, the posterior contraction and the size relation of the confidence ball will immediately imply the corresponding global adaptive minimax results over the covered scale (actually, simultaneously for all scales that are covered by the oracle rate $r(\theta)$). We summarize all these results at once by the following corollary.

Corollary 1. *Let Conditions* (A1)–(A2) *and* (28) *hold. Then for any $M \geq 0$*

$$\sup_{\theta\in\Theta_\beta} \mathrm{E}_\theta \hat{\pi}\big(\|\vartheta-\theta\|^2 \geq M_0 r^2(\Theta_\beta) + M\sigma^2 | X\big) \leq H_0 e^{-m_0 M},$$

$$\sup_{\theta\in\Theta_\beta} \mathrm{P}_\theta\big(\|\hat{\theta}-\theta\|^2 \geq M_1 r^2(\Theta_\beta) + M\sigma^2\big) \leq H_1 e^{-m_1 M},$$

$$\sup_{\theta\in\Theta_\beta} \mathrm{P}_\theta\big(\hat{r}^2 \geq M_0' r^2(\Theta_\beta) + (M+1)\sigma^2\big) \leq H_0' e^{-M}.$$

There is no point in specializing Theorem 2 and the coverage property of Theorem 3 to a particular scale $\{\Theta_\beta, \beta\in\mathcal{B}\}$. Indeed, Theorem 2 holds uniformly over the entire space $\mathbb{R}^n$, hence also over any Θ_β, and the coverage property holds uniformly only over the EBR class $\Theta_{\mathrm{eb}}(t)$ (hence also over $\Theta_{\mathrm{eb}}(t)\cap\Theta_\beta$), whichever scale we consider.

Next, we consider three scales $\{\Theta_\beta, \beta\in\mathcal{B}\}$ for which the adaptive minimax results on the estimation problem, the contraction rate of the empirical Bayes posterior, and the size property of the confidence ball $B\big(\hat{\theta}, (M_2\hat{r}^2+M)^{1/2}\big)$ follow from our local results Theorems 1 and 3. The results for other (covered) scales can also be readily derived.

Nearly black vectors

For $p_n\in[n]$ such that $p_n=o(n)$ as $n\to\infty$, introduce the sparsity class $\ell_0[p_n]=\{\theta\in\mathbb{R}^n : s(\theta)=|I^*(\theta)|\leq p_n\}$, where by $I^*(\theta)$ and $s(\theta)$ we denote the active index set and the sparsity of $\theta\in\mathbb{R}^n$:

$$I^*(\theta)=\big\{i\in[n]:\theta_i\neq 0\big\}, \qquad s(\theta)=\big|I^*(\theta)\big|. \tag{29}$$

Let $a_n \asymp b_n$ mean that $a_n=O(b_n)$ and $b_n=O(a_n)$ as $n\to\infty$. The minimax estimation rate over the class of nearly black vectors $\ell_0[p_n]$ with the sparsity parameter p_n is known to be $r^2(\ell_0[p_n])\asymp\sigma^2 p_n\log(\frac{n}{p_n})$; see [17]. By the definition (19) of the oracle rate $r^2(\theta)$, we have that $r^2(\theta)\leq r^2(I^*(\theta),\theta)$. Then we trivially obtain (28):

$$\sup_{\theta\in\ell_0[p_n]} r^2(\theta) \leq \sup_{\theta\in\ell_0[p_n]} r^2\big(I^*(\theta),\theta\big) \leq \sigma^2 p_n \log\Big(\frac{en}{p_n}\Big) \asymp r^2(\ell_0[p_n]).$$

The last relation, Theorems 1 and 3 immediately imply the adaptive minimax results for the scale $\ell_0[p_n]$. These results are given by Corollary 1 with $\Theta_\beta=\ell_0[p_n]$ and $r^2(\Theta_\beta)=r^2(\ell_0[p_n])\asymp \sigma^2 p_n\log(\frac{en}{p_n})$.

The next theorem describes some "over-dimensionality" (or "undersmoothing") control of the empirical Bayes posterior $\hat{\pi}(I|X)$ from the P_θ-perspective. The proof is given in Supplement [6].

Theorem 5. *Let $s(\theta)$ be defined by* (29)*. Under the conditions of Theorem* 2*, there exist $M_4, m_4>0$ such that for any $M>M_4$ and $\theta\in\mathbb{R}^n$*

$$\mathrm{E}_\theta\hat{\pi}\big(I:|I|>Ms(\theta)|X\big)\leq C_0\exp\Big\{-m_4 s(\theta)\Big[(M-M_4)\log\Big(\frac{en}{s(\theta)}\Big)-M\log M\Big]\Big\}.$$

In particular, there exist constants $M_4', m_4' > 0$ *such that*

$$\mathrm{E}_\theta \hat{\pi}\big(I : |I| > M_4' s(\theta) | X\big) \le C_0 \exp\Big\{-m_4' s(\theta) \log\Big(\frac{en}{s(\theta)}\Big)\Big\}.$$

The above theorem is a local type result, but can readily be specialized to the sparsity class $\theta \in \ell_0[p_n]$ in the minimax sense. If $s(\theta) \ge 1$, the probability bound goes to 0 as $n \to \infty$.

Weak ℓ_s-balls

For $s \in (0, 2)$, the *weak ℓ_s-ball* with the sparsity parameter p_n is defined by

$$m_s[p_n] = \big\{\theta \in \mathbb{R}^n : \theta_{[i]}^2 \le (p_n/n)^2 (n/i)^{2/s},\ i = 1, \dots, n\big\}, \quad p_n = o(\sigma n) \text{ as } n \to \infty,$$

where $\theta_{[1]}^2 \ge \cdots \ge \theta_{[n]}^2$ are the ordered $\theta_1^2, \dots, \theta_n^2$. This scale can be thought of as Sobolev hyper-rectangle for ordered (with unknown locations) coordinates: $m_s[p_n] = \mathcal{H}(\beta, \delta_n) = \{\theta \in \mathbb{R}^n : |\theta_{[i]}| \le \delta_n i^{-\beta}\}$, with $\delta_n = n^{1/s} \frac{p_n}{n}$ and $\beta = 1/s > 1/2$.

Denote $j = O_\theta(i)$ if $\theta_i^2 = \theta_{[j]}^2$, with the convention that in the case $\theta_{i_1}^2 = \cdots = \theta_{i_k}^2$ for $i_1 < \cdots < i_k$ we let $O_\theta(i_{l+1}) = O_\theta(i_l) + 1$, $l = 1, \dots, k-1$. The minimax estimation rate over this class is $r^2(m_s[p_n]) \asymp n\big(\frac{p_n}{n}\big)^s \big[\sigma^2 \log\big(\frac{n\sigma}{p_n}\big)\big]^{1-s/2}$ when $n^{2/s}\big(\frac{p_n}{n}\big)^2 \ge \sigma^2 \log n$, and $r^2(m_s[p_n]) \asymp n^{2/s}\big(\frac{p_n}{n}\big)^2 + \sigma^2$ when $n^{2/s}\big(\frac{p_n}{n}\big)^2 < \sigma^2 \log n$; see [16] and [8]. Then take $I^*(\theta) = \{i : O_\theta(i) \le p_n^*\}$, with $p_n^* = en\big(\frac{p_n}{n\sigma}\big)^s \big[\log\big(\frac{n\sigma}{p_n}\big)\big]^{-s/2}$ in the case $n^{2/s}\big(\frac{p_n}{n}\big)^2 \ge \sigma^2 \log n$, to derive (28):

$$\begin{aligned}
\sup_{\theta \in m_s[p_n]} r^2(\theta) &\le \sup_{\theta \in m_s[p_n]} r^2\big(I^*(\theta), \theta\big) \le \sigma^2 p_n^* \log\Big(\frac{en}{p_n^*}\Big) + n^{2/s}\Big(\frac{p_n}{n}\Big)^2 \sum_{i > p_n^*} i^{-2/s} \\
&\le K_1 \sigma^2 p_n^* \log\Big(\frac{n\sigma}{p_n}\Big) + K_2 n^{2/s}\Big(\frac{p_n}{n}\Big)^2 (p_n^*)^{1-2/s} \\
&\le K n\Big(\frac{p_n}{n}\Big)^s \Big[\sigma^2 \log\Big(\frac{n\sigma}{p_n}\Big)\Big]^{1-s/2} \asymp r^2(m_s[p_n]),
\end{aligned}$$

for some $K = K(s)$. The case $n^{2/s}(\frac{p_n}{n})^2 < \sigma^2 \log n$ is treated similarly by taking $p_n^* = 0$. Theorems 1 and 3 imply the minimax adaptive results for the scale $m_s[p_n]$. These results are obtained by setting $\Theta_\beta = m_s[p_n]$ and $r^2(\Theta_\beta) = r^2(m_s[p_n])$ in Corollary 1.

The following theorem concerns the "over-dimensionality" control for the class $m_s[p_n]$, the proof of this theorem is provided in Supplement [6].

Theorem 6. *Let the conditions of Theorem 2 be fulfilled,* $p_n^* = en\big(\frac{p_n}{n\sigma}\big)^s \big[\log(\frac{n\sigma}{p_n})\big]^{-s/2}$ *and* $n^{2/s}\big(\frac{p_n}{n}\big)^2 \ge \sigma^2 \log n$. *Then there exist constants* $M_5, m_5 > 0$ *such that for any* $M > M_5$ *there exists* $n_0 = n_0(M, s)$ *such that, for all* $n \ge n_0$,

$$\sup_{\theta \in m_s[p_n]} \mathrm{E}_\theta \hat{\pi}\big(I : |I| > M p_n^* | X\big) \le C_0 \exp\Big\{-m_5 (M - M_5) p_n^* \log\Big(\frac{n\sigma}{p_n}\Big)\Big\}.$$

Notice that the exponential upper bound from the last relation converges to zero as $n \to \infty$ because $p_n^* \log\big(\frac{n\sigma}{p_n}\big) \ge e(\sigma^2 \log n)^{s/2}\big[\log\big(\frac{n\sigma}{p_n}\big)\big]^{1-s/2}$.

Remark 6. The same minimax results hold over the so called *strong ℓ_s-ball* $\ell_s[p_n] = \big\{\theta \in \mathbb{R}^n : \frac{1}{n}\sum_{i=1}^n |\theta_i|^s \le \big(\frac{p_n}{n}\big)^s\big\}$, $s \in (0,2)$, since $\ell_s[p_n] \subseteq m_s[p_n] \subseteq \ell_{s'}[p_n]$ for any $s' > s$.

Besov scale

Let $J_0 \in \mathbb{N}$ be such that $2^{J_0+1} = n$. Suppose we observe

$$Y_{jk} = \theta_{jk} + \frac{1}{\sqrt{n}}\xi_{jk}, \quad \xi_{jk} \overset{\text{ind}}{\sim} \mathrm{N}(0,1),\ (jk) \in \mathcal{K} = \big\{(jk) : j \in [J_0]_0, k \in [2^j]\big\}. \tag{30}$$

This model is obtained as a high dimensional "projected" (see (9.57) in [18]) variant of the *orthogonal wavelet transform* of an additive regression function observed in Gaussian noise with $\sigma^2 = n^{-1}$, or just as a sequence version (with respect to some wavelet basis) of the continuous *white noise model*. For details and many interesting connections and relations to the function estimation theory we refer to the very comprehensive and insightful account [18] on this topic. We adopt the notation and conventions from [18].

We can see (30) as $J_0 + 1$ models of type (1), where $\sigma^2 = n^{-1}$ and the j-th model has 2^j observations, $j \in [J_0]_0$. Let $\theta^j = \big(\theta_{jk}, k \in [2^j]\big)$ and $r^2(\theta^j, I_{oj})$ denote the oracle rate in j-th model. Then aggregating the oracle results over these $J_0 + 1 = \log_2 n$ models leads to the results for the whole model (30) with the aggregated oracle rate $r^2(\theta) = \sum_{i \in [J_0]_0} r^2(\theta^j, I_{oj})$. Because of the aggregation, in Theorem 1 we get $\frac{\log_2 n}{n} M$ instead of $\sigma^2 M$ and $(\log_2 n) H_l$ instead of H_l, $l = 0, 1$.

Assume that the true signal θ belongs to a Besov ball

$$\Theta_{p,q}^{\beta}(Q) = \Big\{\theta : \sum_{j \in [J_0]_0} 2^{ajq}\Big(\sum_{k \in [2^j]} \theta_{jk}^p\Big)^{q/p} \le Q^q\Big\}, \quad a = \beta + \frac{1}{2} - \frac{1}{p},$$

for some $p, q, Q > 0$, $\beta \ge 1/p$. The minimax rate over $\Theta_{p,q}^{\beta}(Q)$ is known to be $r^2\big(\Theta_{p,q}^{\beta}(Q)\big) \asymp n^{-\frac{2\beta}{2\beta+1}}$. Now, for any $\theta \in \Theta_{p,q}^{\beta}(Q)$,

$$\begin{aligned} r^2(\theta) &\le \sum_{j \in [J_0]_0} \sum_{k \in I_{oj}^c} \theta_{jk}^2 + \sum_{j \in [J_0]_0} \frac{|I_{oj}|}{n} \log\Big(\frac{e2^j}{|I_{oj}|}\Big) \\ &\le \sum_{j \in [J_0]_0} \min_{0 \le k \le 2^j} \Big(\sum_{l > k} \theta_{j(l)}^2 + \frac{k}{n}\log\big(e2^j/k\big)\Big) \\ &\le Cn^{-\frac{2\beta}{2\beta+1}} \asymp r^2\big(\Theta_{p,q}^{\beta}(Q)\big), \end{aligned}$$

where $\theta_{j(l)}^2$ denotes the l-th largest value among $\{\theta_{jk}^2, j \in [2^k]\}$. The third inequality of the last display follows from Theorem 12.1 in [18] under the assumption $\beta \ge 1/p$. We thus established the relation (28) for the Besov scale, and the global minimax adaptive results for the Besov scale follow by Corollary 1 with $\Theta_\beta = \Theta^\beta_{p,q}(Q)$ and the minimax rate $r^2(\Theta_\beta) = r^2\big(\Theta^\beta_{p,q}(Q)\big) \asymp n^{-\frac{2\beta}{2\beta+1}}$. Recall that we also need to set $\frac{\log_2 n}{n} M$ instead of $\sigma^2 M$ and $(\log_2 n) H_l$ instead of H_l, $l = 0, 1$, because of the aggregation. In this case, the asymptotic regime $n \to \infty$ is of interest. Let us formulate the first claim of Corollary 1 in this case (other claims can be formulated similarly): for some $C > 0$ and any $M \geq 0$,

$$\sup_{\theta\in\Theta^\beta_{p,q}(Q)} \mathrm{E}_\theta \hat{\pi}\Big(\|\vartheta - \theta\|^2 \geq Cn^{-\frac{2\beta}{2\beta+1}} + M\frac{\log_2 n}{n}\big|X\Big) \leq H_0(\log_2 n)e^{-m_0 M}.$$

Take for example $M = M_n = n^{1/(2\beta+1)}/\log_2 n$ to obtain a well interpreted asymptotic relation.

We should mention that there are of course more scales covered by the oracle rate $r^2(\theta)$, one can establish the relation (28) for other scales, for example for smoothness scales (with a log factor in the minimax rate for smoothness scales). Besides, the results can be extended to non-normal, not independent ξ_{jk}'s, but only satisfying Condition (A1).

4. Simulations

Here we present a small simulation study. In the model (1), we used $n = 500$, $\xi_i \overset{\text{ind}}{\sim} \mathrm{N}(0,1)$, $\sigma = 1$, and signals $\theta \in \mathbb{R}^n$ of the form $\theta = \theta(p, A) = (0, \dots, 0, A, \dots, A) = (0 \cdot 1_{n-p}, A1_p)$, where $1_p = (1, \dots, 1) \in \mathbb{R}^p$. The first $n - p$ zero coordinates are "insignificant" and the last p coordinates are "significant". Different sparsity levels $p \in \{25, 50, 100\}$ and "signal strengths" $A \in \{3, 4, 5\}$ are considered. It is easy to compute the oracle $I_o(\theta(p, A)) = \{n-p+1, \dots, n\}$ and oracle rate $r^2(\theta(p, A)) = B(I_o, \theta) + V(I_o) = V(I_o) = |I_o| \log(en/|I_o|) = p \log(en/p)$. Then the excessive bias $B(I_o, \theta(p, A)) = 0$, so that the deceptiveness $b(\theta(p, A), 1) = 0$ and hence the EBR condition (in terms of the "standard" oracle I_o) is satisfied with $t = 0$: $\theta(p, A) \in \Theta_{\mathrm{eb}}(0, 1)$ for all considered $\theta(p, A)$.

In case $\xi_i \overset{\text{ind}}{\sim} \mathrm{N}(0,1)$, Condition (A1) is fulfilled with $\beta = 0.4$, $B = 1$, leading to $\varkappa > 3.24$ in Condition (A2). The bound for $\varkappa$ coming from Condition (A2) is typically too conservative as it is for the general situation of unknown distribution of ξ. For example, if $\xi_i \overset{\text{ind}}{\sim} \mathrm{N}(0,1)$, Condition (A2) can be relaxed to $\varkappa \geq 2.04$; see Supplement [6]. It is desirable to choose $\varkappa$ in a data dependent way. In our simulation study, we choose $\varkappa$ via a *cross-validation* procedure. For that, we create two independent normal samples $X'_i = X_i + \eta_i$ and $X''_i = X_i - \eta_i$, where we generate $\eta_i \overset{\text{ind}}{\sim} \mathrm{N}(0,1)$, independently of ξ. Then X'_i and X''_i are independent random variables with means θ_i and variances 2, thus the observation sample can be duplicated at the cost of doubling the variance. Now we estimate $\varkappa > 0$ as follows: let $\check{\theta}' = \check{\theta}'(\hat{I}') = (X'_i 1\{i \in \hat{I}'\}, i \in [n])$, where $\hat{I}' = \hat{I}'(\varkappa) = \arg\min_{I\in\mathcal{I}}\big\{-\sum_{i\in I}(X'_i)^2 + 2(2\varkappa+1)|I|\log\big(\frac{en}{|I|}\big)\big\}$, then $\hat{\varkappa} = \arg\min_{\varkappa\in(0,\log n]} \|\check{\theta}'(\hat{I}'(\varkappa)) - X''\|^2$. In the simulations below, $\hat{\varkappa}$ turned out to be rather stable, varying primarily in the range [0.4, 0.9], the choice $\varkappa = 0.65$ gave reasonable results (but still worse then data dependent $\hat{\varkappa}$). Recall that, according to (13), $\hat{I}(\varkappa) = \big\{i \in [n] : X^2_i \geq X^2_{[\hat{k}]}\big\}$,

Table 1. The ratio $\overline{\hat{R}}/r(\theta)$ and the frequency $\bar{\alpha}$ of the event that the confidence ball $B(\check{\theta}, \hat{R})$ contains the signal $\theta(p, A)$ computed for 100 vectors X simulated from (1) with $\xi_i \overset{\text{ind}}{\sim} \mathrm{N}(0, 1)$, $n = 500$ and $\sigma = 1$

p	25			50			100		
A	3	4	5	3	4	5	3	4	5
$\overline{\hat{R}}/r(\theta)$	1.81	1.71	1.51	1.78	1.50	1.40	1.51	1.34	1.34
$\bar{\alpha}$	0.98	0.97	0.95	0.99	0.97	1	0.97	1	1

where $\hat{k} = \hat{k}(\varkappa) = \arg\min_{k\in[n]_0}\big\{\sum_{i=k+1}^n X_{[i]}^2 + (2\varkappa + 1)\sigma^2 k \log(en/k)\big\}$. Let $\hat{I} = \hat{I}(\hat{\varkappa})$ and $\check{\theta} = X(\hat{I})$ be defined by (14).

Now consider the confidence ball $B(\check{\theta}, \hat{R})$, where $\hat{R} = \bar{M}\big[(\hat{b}+1)_+\hat{r}^2\big]^{1/2}$, $\hat{r}^2 = |\hat{I}|\log(en/|\hat{I}|)$ given by (24), and $\hat{b} = \frac{\sum_{i\in\hat{I}^c}(X_i^2-1)}{|\hat{I}|\log(en/|\hat{I}|)}$ is an estimate of the deceptiveness $b(\theta, 1)$ defined by (26). The construction of the radius $\hat{R}$ is inspired by the local result formulation in Remark 5. The quantity $\big[(\hat{b} + 1)_+\hat{r}^2\big]^{1/2}$ is an empirical counterpart for the oracle rate $r(\theta)$, but even the oracle rate radius needs to be inflated to ensure coverage. The multiplicative factor $\bar{M}$ is intended to trade-off the size of the ball against its coverage probability. Theoretical inflating factor from Theorem 3 is too conservative, as it is for the general situation of Condition (A1). In this simulation study it is enough to take $\bar{M} = \sqrt{2}$, thus $\hat{R} = \big[2(\hat{b} + 1)_+\hat{r}^2\big]^{1/2}$, which yielded good results for all the cases.

Table 1 shows the performance of the confidence ball $B(\check{\theta}, \hat{R})$. For each $\theta = \theta(p, A)$, with $p \in \{25, 50, 100\}$ and $A \in \{3, 4, 5\}$, we simulated 100 data vectors X from the model (1) and computed two quantities: (1) the ratio $\overline{\hat{R}}/r(\theta)$ of the average of the radius $\overline{\hat{R}}$ to the oracle rate $r(\theta)$ defined by (19); (2) the frequency $\bar{\alpha}$ of the event that confidence ball $B(\check{\theta}, \hat{R})$ contains the signal θ. The former characterizes the size of the confidence ball $B(\check{\theta}, \hat{R})$ relative to the oracle rate, and the latter its coverage. What one can conclude from the table: the higher the signal strength, the smaller the ratio $\overline{\hat{R}}/r(\theta)$; uncertainty quantification does not seem to benefit from the sparsity in terms of relative size (with respect to the oracle rate) and coverage. Of course, the absolute size is certainly better for more sparse signals as the oracle rate is then smaller. In particular, $r(\theta(25, A)) = 9.99$, $r(\theta(50, A)) = 12.85$ and $r(\theta(100, A)) = 16.15$, $A = 3, 4, 5$.

At the first site surprisingly, the constructed confidence ball $B(\check{\theta}, \hat{R})$ appeared to perform well also for deceptive parameters, like $\bar{\theta} = \bar{\theta}(\delta, p, A) = (\delta 1_{n-p}, A1_p)$, with p, A as before and $\delta > 0$. We get very similar (good) results as in Table 1 for *all* $\delta > 0$. Notice that the deceptivenesses may not be zero, for example, $b(\bar{\theta}(0.5, 25, A)) = 1.18$, $b(\bar{\theta}(0.5, 50, A)) = 0.68$, $b(\bar{\theta}(0.5, 100, A)) = 0.38$, $b(\bar{\theta}(0.8, 25, A)) = 3$, etc. The reason for good results even for deceptive signals is that their oracle rates are large relative to $n^{1/4}$. Indeed, even the smallest oracle rate $r(\theta(0, 25, 3)) = 9.99 > (500)^{1/4} = 4.73$, which means that we are essentially in the $n^{1/4}$-situation of Theorem 4 (when both the optimal size and coverage are possible to attain) rather than Theorem 3. Basically, the signals $\bar{\theta}(\delta, p, A)$ are not sparse enough and/or the problem is not high-dimensional enough.

To distill the deceptiveness effect, we created a signal $\theta' = (0, \dots, 0, A_1, \dots, A_p)$ of dimension $n = 500$, with sparsity $p = 10$, $A_i \overset{\text{ind}}{\sim} \mathrm{U}[0, 4]$, $i = 1, \dots, p$. The oracle rate was $r(\theta') = 4.69 < (500)^{1/4} = 4.73$, the deceptiveness was $b(\theta') = 0.64$. Thus, this was a deceptive signal, but not in the $n^{1/4}$-situation anymore. The size was relatively good $\overline{\hat{R}/r(\theta)} = 1.21$, but the coverage was low $\bar{\alpha} = 0.55$. The deceptiveness manifested itself more prominently in the case $n = 1000$, $p = 10$, $A_i \overset{\text{ind}}{\sim} \mathrm{U}[0, 4]$, $i = 1, \dots, p$. The oracle rate was $r(\theta') = 5.52 < (1000)^{1/4} = 5.62$, mild deceptiveness $b(\theta') = 0.47$. The size was still good $\overline{\hat{R}/r(\theta)} = 1.37$ as before, but the coverage was low $\bar{\alpha} = 0.59$.

5. Technical lemmas

First, we provide a couple of technical lemmas used in the proofs of the main results.

Remark 7. Notice that in the below lemma we established the same bound for the both quantities $\mathrm{E}_\theta \hat{\pi}(I|X) = \mathrm{E}_\theta \tilde{\pi}(I|X)$ and $\mathrm{E}_\theta \mathbb{1}\{\hat{I} = I\} = \mathrm{P}_\theta(\hat{I} = I)$. The proofs of the properties of $\breve{\pi}(\vartheta|X)$ and $\breve{\theta}$ are exactly the same as for $\tilde{\pi}(\vartheta|X)$ and $\tilde{\theta}$, with the only difference that everywhere (in the claims and in the proofs) $\hat{\pi}(I \in \mathcal{G}|X)$ should be read as $\tilde{\pi}(I \in \mathcal{G}|X)$ in case $\hat{\pi} = \tilde{\pi}$; and as $\mathbb{1}\{\hat{I} \in \mathcal{G}\}$ in case $\hat{\pi} = \breve{\pi}$, for all $\mathcal{G} \subseteq \mathcal{I}$ that appear in the proof. Hence, $\mathrm{E}_\theta \hat{\pi}(I \in \mathcal{G}|X) = \mathrm{E}_\theta \tilde{\pi}(I \in \mathcal{G}|X)$ in the former case, and $\mathrm{E}_\theta \hat{\pi}(I \in \mathcal{G}|X) = \mathrm{P}_\theta(\hat{I} \in \mathcal{G})$ in the latter case.

Lemma 1. *Let Condition (A1) be fulfilled. Then for any $\theta \in \mathbb{R}^n$ and any $I, I_0 \in \mathcal{I}$,*

$$\mathrm{E}_\theta \hat{\pi}(I|X) \le \Big(\frac{\lambda_I}{\lambda_{I_0}}\Big)^h \exp\Big\{B_h \sum_{i \in I \setminus I_0} \frac{\theta_i^2}{\sigma^2} - A_h \sum_{i \in I_0 \setminus I} \frac{\theta_i^2}{\sigma^2} + C_h |I_0| \log\Big(\frac{en}{|I_0|}\Big) - D_h |I| \log\Big(\frac{en}{|I|}\Big)\Big\},$$

where $h = \frac{2\beta}{3}$, $A_h = \frac{\beta}{6}$, $B_h = \frac{2\beta}{3}$, $C_h = \frac{\beta + B}{3}$ and $D_h = \frac{\beta - 2B}{3}$. If $I \setminus I_0 = \varnothing$, the bound holds also for $h = \beta$ with $A_h = \frac{\beta}{3}$, $B_h = 0$, $C_h = \frac{\beta}{2} + B$, $D_h = \frac{\beta}{2}$. If $I_0 \setminus I = \varnothing$, the bound holds also for $h = \beta$ with $A_h = 0$, $B_h = \beta$, $C_h = \frac{\beta}{2}$, $D_h = \frac{\beta}{2} - B$.

Proof of Lemma 1. Recall that $\mathrm{P}_{X,I} = \phi\big(X_i \mathbb{1}\{i \notin I\}, 0, \sigma^2 + K_n(I)\sigma^2 \mathbb{1}\{i \in I\}\big)$. In case $\hat{\pi}(I|X) = \tilde{\pi}(I|X)$, we get by (11) that, for any $I, I_0 \in \mathcal{I}$ and any $h \in [0, 1]$,

$$\mathrm{E}_\theta \hat{\pi}(I|X) = \mathrm{E}_\theta \tilde{\pi}(I|X) = \mathrm{E}_\theta \frac{\lambda_I \mathrm{P}_{X,I}}{\sum_{J \in \mathcal{I}} \lambda_J \mathrm{P}_{X,J}} \le \mathrm{E}_\theta \Big(\frac{\lambda_I \mathrm{P}_{X,I}}{\lambda_{I_0} \mathrm{P}_{X,I_0}}\Big)^h \tag{31}$$

$$= \mathrm{E}_\theta \bigg[\frac{\lambda_I \prod_{i=1}^n \phi\big(X_i \mathbb{1}\{i \notin I\}, 0, \sigma^2 + K_n(I)\sigma^2 \mathbb{1}\{i \in I\}\big)}{\lambda_{I_0} \prod_{i=1}^n \phi\big(X_i \mathbb{1}\{i \notin I_0\}, 0, \sigma^2 + K_n(I_0)\sigma^2 \mathbb{1}\{i \in I_0\}\big)}\bigg]^h$$

$$= \Big(\frac{\lambda_I}{\lambda_{I_0}}\Big)^h \mathrm{E}_\theta \exp\Big\{\frac{h}{2}\Big[\sum_{i \in I \setminus I_0} \frac{X_i^2}{\sigma^2} - \sum_{i \in I_0 \setminus I} \frac{X_i^2}{\sigma^2} + |I_0| \log\Big(\frac{en}{|I_0|}\Big) - |I| \log\Big(\frac{en}{|I|}\Big)\Big]\Big\}. \tag{32}$$

In case $\hat{\pi}(I|X) = \mathbb{1}\{\hat{I} = I\}$, by the definition (13) of $\hat{I}$ and the Markov inequality, we derive that, for any $I, I_0 \in \mathcal{I}$ and any $h \geq 0$

$$\begin{aligned} \mathrm{E}_\theta \hat{\pi}(I|X) = \mathrm{P}_\theta(\hat{I} = I) &\leq \mathrm{P}_\theta\Big(\frac{\tilde{\pi}(I|X)}{\tilde{\pi}(I_0|X)} \geq 1\Big) \\ &\leq \mathrm{E}_\theta\Big(\frac{\tilde{\pi}(I|X)}{\tilde{\pi}(I_0|X)}\Big)^h = \mathrm{E}_\theta\Big(\frac{\lambda_I \mathrm{P}_{X,I}}{\lambda_{I_0}\mathrm{P}_{X,I_0}}\Big)^h, \end{aligned}$$

which yields exactly the bound (31), and hence the bound (32) again.

Using Hölder's inequality, Condition (A1) and the two elementary facts $X_i^2 \leq 2\theta_i^2 + 2\sigma^2\xi_i^2$ and $-X_i^2 \leq -\frac{\theta_i^2}{2} + \sigma^2\xi_i^2$, we obtain

$$\begin{aligned} &\mathrm{E}_\theta \exp\Big\{\frac{\beta}{3}\Big[\sum_{i\in I\backslash I_0}\frac{X_i^2}{\sigma^2} - \sum_{i\in I_0\backslash I}\frac{X_i^2}{\sigma^2}\Big]\Big\} \\ &\leq \Big(\mathrm{E}_\theta e^{\frac{\beta}{2}\sum_{i\in I\backslash I_0} X_i^2/\sigma^2}\Big)^{2/3}\Big(\mathrm{E}_\theta e^{-\beta\sum_{i\in I_0\backslash I} X_i^2/\sigma^2}\Big)^{1/3} \\ &\leq \exp\Big\{\frac{2\beta}{3}\sum_{i\in I\backslash I_0}\frac{\theta_i^2}{\sigma^2} + \frac{2B}{3}|I\backslash I_0| - \frac{\beta}{6}\sum_{i\in I_0\backslash I}\frac{\theta_i^2}{\sigma^2} + \frac{B}{3}|I_0\backslash I|\Big\}. \end{aligned}$$

Since $|I\backslash I_0| \leq |I| \leq |I|\log(\frac{en}{|I|})$ and $|I_0\backslash I| \leq |I_0| \leq |I_0|\log(\frac{en}{|I_0|})$, the lemma follows for $h = \frac{2\beta}{3}$ from the last display and (32).

If $I\backslash I_0 = \varnothing$, we take $h = \beta$ in (32) and combine this with $\mathrm{E}_\theta \exp\{-\frac{\beta}{2}\sum_{i\in I_0\backslash I}\frac{X_i^2}{\sigma^2}\} \leq \exp\{-\frac{\beta}{3}\sum_{i\in I_0\backslash I}\frac{\theta_i^2}{\sigma^2} + B|I_0\backslash I|\}$, which holds in view of Condition (A1) and $-\frac{X_i^2}{\sigma^2} \leq -\frac{2\theta_i^2}{3\sigma^2} + 2\xi_i^2$, as $(a+b)^2 \geq 2a^2/3 - 2b^2$. If $I_0\backslash I = \varnothing$, we take $h = \beta$ in (32) and combine this with $\mathrm{E}_\theta \exp\{\frac{\beta}{2}\sum_{i\in I\backslash I_0}\frac{X_i^2}{\sigma^2}\} \leq \exp\{\beta\sum_{i\in I\backslash I_0}\frac{\theta_i^2}{\sigma^2} + B|I\backslash I_0|\}$ which holds in view of Condition (A1) and $\frac{X_i^2}{\sigma^2} \leq \frac{2\theta_i^2}{\sigma^2} + 2\xi_i^2$. □

Note that above lemma holds for any set $I_0 \in \mathcal{I}$. By taking $I_0 = I_o$ defined by (19), we obtain the following lemma.

Lemma 2. *Let Conditions* (A1) *and* (A2) *be fulfilled. Then there exist positive constants $c_1 = c_1(\varkappa) > 2$, c_2 and $c_3 = c_3(\varkappa)$ such that for any $\theta \in \mathbb{R}^n$*

$$\mathrm{E}_\theta \hat{\pi}(I|X) \leq \Big(\frac{ne}{|I|}\Big)^{-c_1|I|}\exp\big\{-c_2\sigma^{-2}\big[r^2(I,\theta) - c_3 r^2(\theta)\big]\big\}.$$

Proof of Lemma 2. With constants h, A_h, B_h, C_h, D_h given in Lemma 1, define the constant $c_1 = c_1(\varkappa) = \varkappa h + D_h - A_h = \frac{2\beta\varkappa}{3} + \frac{\beta - 2B}{3} - \frac{\beta}{6} > 2$ as $\varkappa > \bar{\varkappa}$ by Condition (A2). Since $\varkappa h +$

$D_h = c_1 + A_h$, the definition (5) of λ_I entails that

$$\Big(\frac{\lambda_I}{\lambda_{I_0}}\Big)^h \exp\Big\{C_h|I_0|\log\Big(\frac{en}{|I_0|}\Big) - D_h|I|\log\Big(\frac{en}{|I|}\Big)\Big\}$$
$$= \Big(\frac{ne}{|I|}\Big)^{-c_1|I|} \exp\Big\{(\varkappa h + C_h)|I_0|\log\Big(\frac{en}{|I_0|}\Big) - A_h|I|\log\Big(\frac{en}{|I|}\Big)\Big\}.$$

Using the last relation and Lemma 1 with $I_0 = I_o$, we bound

$$\mathrm{E}_\theta \hat{\pi}(I|X)$$
$$\le \Big(\frac{\lambda_I}{\lambda_{I_o}}\Big)^h \exp\Big\{B_h \sum_{i\in I\setminus I_o} \frac{\theta_i^2}{\sigma^2} - A_h \sum_{i\in I_o\setminus I} \frac{\theta_i^2}{\sigma^2} + C_h|I_o|\log\Big(\frac{en}{|I_o|}\Big) - D_h|I|\log\Big(\frac{en}{|I|}\Big)\Big\}$$
$$= \Big(\frac{ne}{|I|}\Big)^{-c_1|I|} \exp\Big\{-A_h \sum_{i\in I_o\setminus I} \frac{\theta_i^2}{\sigma^2} - A_h|I|\log\Big(\frac{en}{|I|}\Big) + B_h \sum_{i\in I\setminus I_o} \frac{\theta_i^2}{\sigma^2} + (\varkappa h + C_h)|I_o|\log\Big(\frac{en}{|I_o|}\Big)\Big\}.$$

The claim of the lemma follows with the constants $c_1 = (4\beta\varkappa + \beta - 4B)/6 > 2$, $c_2 = A_h = \beta/6$, $h = \frac{2\beta}{3}$ and $c_3 = c_3(\varkappa) = \max\{B_h, \varkappa h + C_h\}/A_h = (\varkappa h + C_h)/A_h = 4\varkappa + 2(\beta + B)/\beta$. □

Lemma 3. *Let $Y_1, \ldots, Y_n$ be random variables such that $\mathrm{E}e^{t\sum_{i\in I} Y_i} \le A_{|I|}(t)$ for all $I \in \mathcal{I}$, with some $t > 0$ and $A_k(t)$. Let $Y_{[1]} \ge Y_{[2]} \ge \cdots \ge Y_{[n]}$. Then, for any $k \in \{1, \ldots, n\}$ and $C, c \ge 0$,*

$$\mathrm{P}\Big(\sum_{i=1}^k Y_{[i]} \ge Ck\log\Big(\frac{en}{k}\Big) + c\Big) \le A_k(t)\exp\Big\{-(Ct-1)k\log\Big(\frac{en}{k}\Big) - ct\Big\},$$
$$\mathrm{E}\sum_{i=1}^k Y_{[i]} \le t^{-1}\Big[k\log\Big(\frac{en}{k}\Big) + \log\big(A_k(t)\big)\Big].$$

In particular, if $\xi_1, \ldots, \xi_n \overset{\text{ind}}{\sim} \mathrm{N}(0,1)$, then for any $k = 1, \ldots, n$ and $C, c \ge 0$

$$\mathrm{P}\Big(\sum_{i=1}^k \xi_{[i]}^2 \ge Ck\log\Big(\frac{en}{k}\Big) + c\Big) \le \Big(\frac{en}{k}\Big)^{-(0.4C-2)k} e^{-0.4c}, \quad \mathrm{E}\sum_{i=1}^k \xi_{[i]}^2 \le 6k\log\Big(\frac{en}{k}\Big).$$

Proof. By Jensen's inequality, we derive

$$\exp\Big\{t\mathrm{E}\sum_{i=1}^k Y_{[i]}\Big\} \le \mathrm{E}\exp\Big\{t\sum_{i=1}^k Y_{[i]}\Big\} \le \sum_{I:|I|=k} \mathrm{E}\exp\Big\{t\sum_{i\in I} Y_i\Big\} \le \binom{n}{k} A_k(t).$$

Then $\mathrm{E}\exp\big\{t\sum_{i=1}^{k}Y_{[i]}\big\}\leq\binom{n}{k}A_k(t)\leq\exp\big\{k\log(\frac{en}{k})+\log(A_k(t))\big\}$, where we used $\binom{n}{k}\leq\big(\frac{en}{k}\big)^k$. This and the (exponential) Markov inequality yield the first relation:

$$\mathrm{P}\Big(\sum_{i=1}^{k}Y_{[i]}\geq Ck\log\Big(\frac{en}{k}\Big)+c\Big)\leq A_k(t)\exp\Big\{-(Ct-1)k\log\Big(\frac{en}{k}\Big)-ct\Big\}.$$

The first display implies also the second relation: $\mathrm{E}\sum_{i=1}^{k}Y_{[i]}\leq t^{-1}\big[\log\binom{n}{k}+\log(A_k(t))\big]$.

As to the normal case, for any $I\in\mathcal{I}$ and any $t<\frac{1}{2}$ we have that $\mathrm{E}\exp\big\{t\sum_{i\in I}\xi_i^2\big\}=(1-2t)^{-|I|/2}=A_{|I|}(t)$. Since $A_k(t)\leq e^k\leq e^{k\log(en/k)}$ for any $t\leq(1-e^{-2})/2<0.44$, the first assertion for the normal case follows by taking $t=0.4$. By taking $t=\frac{1}{4}$, the second assertion follows since $\mathrm{E}\sum_{i=1}^{k}\xi_{[i]}^2\leq 4k\log\big(\frac{en}{k}\big)+2k\log 2\leq 6k\log\big(\frac{en}{k}\big)$. □

This lemma is useful if $A_k(t)\leq C_1\big(\frac{en}{k}\big)^{C_2k}$ for some $t,C_1,C_2>0$; in particular, for $Y_i=\xi_i^2$, where the ξ_i's satisfy Condition (A1). Then Lemma 3 applies with $t=\beta$ and $A_k(\beta)=e^{Bk}$:

$$\mathrm{P}\Big(\sum_{i=1}^{k}\xi_{[i]}^2\geq\frac{(1+B)}{\beta}k\log\Big(\frac{en}{k}\Big)+M\Big)\leq\exp\{-\beta M\},\quad k=1,\ldots,n,\ M\geq 0. \tag{33}$$

6. Proofs of the theorems

By C_0, C_1, C_2 etc., denote constants which are different in different proofs.

Proof of Theorem 1. Recall the constants c_1, c_2, c_3 defined in the proof of Lemma 2. Let $M_0=2c_3\big(6+\frac{1+B}{\beta}\big)$. Introduce the subfamily of index sets $\mathcal{S}_M=\mathcal{S}_M(\theta)=\big\{I\in\mathcal{I}:r^2(I,\theta)\leq c_3r^2(\theta)+\frac{\beta}{40(1+B)}M\sigma^2\big\}$, $m=m_M(\theta)=\max\{|I|:I\in\mathcal{S}_M\}$, and the event $A_M=A(\theta)=\big\{\sum_{i=1}^{m}\xi_{[i]}^2\leq\frac{(1+B)}{\beta}m\log\big(\frac{en}{m}\big)+\frac{M}{8}\big\}$. We have

$$\begin{aligned}&\hat{\pi}\big(\|\vartheta-\theta\|^2\geq M_0r^2(\theta)+M\sigma^2|X\big)\\&\quad\leq\mathbb{1}_{A_M^c}+\hat{\pi}\big(I\in\mathcal{S}_M^c|X\big)+\sum_{I\in\mathcal{S}_M}\mathbb{1}_{A_M}\hat{\pi}_I\big(\|\vartheta-\theta\|^2\geq M_0r^2(\theta)+M\sigma^2|X\big)\hat{\pi}(I|X)\\&\quad=T_1+T_2+T_3.\end{aligned}$$

Now we bound the quantities $\mathrm{E}_\theta T_1$, $\mathrm{E}_\theta T_2$ and $\mathrm{E}_\theta T_3$.

First, we bound $\mathrm{E}_\theta T_1$ by using Lemma 3 (see also (33)):

$$\mathrm{E}_\theta T_1=\mathrm{P}_\theta\big(A_M^c\big)=\mathrm{P}\Big(\sum_{i=1}^{m}\xi_{[i]}^2>\frac{(1+B)}{\beta}m\log\Big(\frac{en}{m}\Big)+\frac{M}{8}\Big)\leq\exp\{-\beta M/8\}. \tag{34}$$

Let us bound $\mathrm{E}_\theta T_2$. Since $\binom{n}{k} \le (\frac{en}{k})^k$ and $c_1 > 2$, the following relation holds:

$$\sum_{I\in\mathcal{I}}\Big(\frac{ne}{|I|}\Big)^{-c_1|I|} = \sum_{k=0}^{n}\binom{n}{k}\Big(\frac{en}{k}\Big)^{-c_1 k} \le \sum_{k=0}^{n}\Big(\frac{en}{k}\Big)^{-k(c_1-1)} \le \big(1-e^{1-c_1}\big)^{-1} \triangleq C_0. \tag{35}$$

If $I \in \mathcal{S}_M^c$, then $r^2(I,\theta) > c_3 r^2(\theta) + \frac{\beta}{40(1+B)}M\sigma^2$. Using this, Lemma 2 and (35), we bound $\mathrm{E}_\theta T_2$:

$$\begin{aligned}
\mathrm{E}_\theta T_2 &= \sum_{I\in\mathcal{S}_M^c}\mathrm{E}_\theta\hat{\pi}(I|X) \le \sum_{I\in\mathcal{S}_M^c}\Big(\frac{ne}{|I|}\Big)^{-c_1|I|}\exp\big\{-c_2\sigma^{-2}\big[r^2(I,\theta)-c_3r^2(\theta)\big]\big\}\\
&\le \sum_{I\in\mathcal{I}}\Big(\frac{ne}{|I|}\Big)^{-c_1|I|}\exp\big\{-c_2\beta M/\big(40(1+B)\big)\big\}\\
&\le C_0\exp\big\{-c_2\beta M/\big(40(1+B)\big)\big\}.
\end{aligned}\tag{36}$$

It remains to bound $\mathrm{E}_\theta T_3$. For each $I \in \mathcal{S}_M$, $\sigma^2|I|\log(en/|I|) \le r^2(I,\theta) \le c_3r^2(\theta) + \frac{\beta}{40(1+B)}M\sigma^2$. Since $m = \max\{|I| : I \in \mathcal{S}_M\}$, then $\sigma^2 m\log\big(\frac{en}{m}\big) \le c_3r^2(\theta) + \frac{\beta}{40(1+B)}M\sigma^2$. Thus, for any $I \in \mathcal{S}_M$, the event A_M implies that $\sum_{i\in I}\xi_i^2 \le \sum_{i=1}^m \xi_{[i]}^2 \le \frac{(1+B)}{\beta}m\log\big(\frac{en}{m}\big) + \frac{M}{8} \le \frac{(1+B)}{\beta}c_3\sigma^{-2}r^2(\theta) + \frac{3M}{20}$. Denote for brevity $\Delta_M(\theta) = M_0r^2(\theta) + M\sigma^2$ and recall that $\sum_{i\in I^c}\theta_i^2 \le r^2(I,\theta) \le c_3r^2(\theta) + \frac{\beta}{40(1+B)}M\sigma^2 \le c_3r^2(\theta) + \frac{M}{40}\sigma^2$ for any $I \in \mathcal{S}_M$. Then for any $I \in \mathcal{S}_M$

$$A_M \subseteq \bigg\{\frac{\Delta_M(\theta)}{2} - \sigma^2\sum_{i\in I}\xi_i^2 - \sum_{i\in I^c}\theta_i^2 \ge \Big[\frac{M_0}{2} - \frac{1+B+\beta}{\beta}c_3\Big]r^2(\theta) + \frac{13M\sigma^2}{40}\bigg\}. \tag{37}$$

According to (10), $\tilde{\pi}_I(\vartheta|X) = \bigotimes_{i=1}^n \mathrm{N}\big(X_i(I), \sigma_i^2(I)\big)$, with $X_i(I) = X_i\mathbb{1}\{i \in I\}$ and $\sigma_i^2(I) = \frac{K_n(I)\sigma^2\mathbb{1}\{i\in I\}}{K_n(I)+1}$. Let P_Z be the measure of $Z = (Z_1,\dots,Z_n)$, with $Z_i \overset{\text{ind}}{\sim} \mathrm{N}(0,1)$. By using (37), the fact that $\frac{r^2(\theta)}{\sigma^2} \ge c_3^{-1}\big(m\log\big(\frac{en}{m}\big) - \frac{\beta}{40(1+B)}M\big)$ and Lemma 3 (now applied to the Gaussian case), we obtain that, for any $I \in \mathcal{S}_M$,

$$\begin{aligned}
&\tilde{\pi}_I\big(\|\vartheta-\theta\|^2 \ge M_0r^2(\theta) + M\sigma^2|X\big)\mathbb{1}_{A_M}\\
&\quad= \mathrm{P}_Z\bigg(\sum_{i=1}^n\big(\sigma_i(I)Z_i + X_i(I) - \theta_i\big)^2 \ge \Delta_M(\theta)\bigg)\mathbb{1}_{A_M}\\
&\quad\le \mathrm{P}_Z\bigg(\sum_{i=1}^n\sigma_i^2(I)Z_i^2 \ge \frac{\Delta_M(\theta)}{2} - \sum_{i=1}^n\big(X_i(I)-\theta_i\big)^2\bigg)\mathbb{1}_{A_M}\\
&\quad\le \mathrm{P}_Z\bigg(\sum_{i\in I}\sigma^2Z_i^2 \ge \frac{\Delta_M(\theta)}{2} - \sum_{i\in I}\sigma^2\xi_i^2 - \sum_{i\in I^c}\theta_i^2\bigg)\mathbb{1}_{A_M}
\end{aligned}$$

$$
\begin{aligned}
&\le \mathrm{P}_Z\Big(\sum_{i\in I} Z_i^2 \ge \Big[\frac{M_0}{2} - \Big(\frac{1+B}{\beta}+1\Big)c_3\Big]\frac{r^2(\theta)}{\sigma^2} + \frac{13M}{40}\Big)\\
&\le \mathrm{P}_Z\Big(\sum_{i=1}^m Z_{[i]}^2 \ge \Big(\frac{M_0}{2c_3} - \frac{1+B}{\beta} - 1\Big)\Big[m\log\Big(\frac{en}{m}\Big) - \frac{\beta}{40(1+B)}M\Big] + \frac{13M}{40}\Big)\\
&\le \mathrm{P}_Z\Big(\sum_{i=1}^m Z_{[i]}^2 \ge 5m\log\Big(\frac{en}{m}\Big) + \frac{M}{5}\Big) \le \exp\{-2M/25\},
\end{aligned}
$$

where we also used in the last step that $\frac{M_0}{2c_3} - \frac{1+B}{\beta} - 1 = 5$. Hence,

$$
\begin{aligned}
\mathrm{E}_\theta T_3 &= \mathrm{E}_\theta \sum_{I\in\mathcal{S}_M} \mathbb{1}_{A_M}\tilde{\pi}_I\big(\|\vartheta-\theta\|^2 \ge M_0 r^2(\theta) + M\sigma^2|X\big)\hat{\pi}(I|X)\\
&\le \exp\{-2M/25\}\mathrm{E}_\theta \sum_{I\in\mathcal{I}} \hat{\pi}(I|X) \le \exp\{-2M/25\}.
\end{aligned}
$$

This completes the proof of assertion (i) since, in view of (34), (36) and the last display, we established that $\mathrm{E}_\theta\hat{\pi}\big(\|\vartheta-\theta\|^2 \ge M_0 r^2(\theta) + M\sigma^2|X\big) \le \mathrm{E}_\theta(T_1+T_2+T_3) \le (2+C_0)e^{-m_0 M}$, with constants $M_0 = 2c_3\big(6+\frac{1+B}{\beta}\big)$, $H_0 = 2 + C_0$, $m_0 = \min\big\{\frac{\beta}{8}, \frac{c_2\beta}{40(1+B)}, \frac{2}{25}\big\}$ and C_0 defined in (35).

The proof of assertion (ii) proceeds along similar lines. Recall the constants $c_1 > 2$, c_2, c_3 from Lemma 2 and define $M_1 = 4c_3(1+B+\beta)/\beta$. Introduce the subfamily of sets

$$
\bar{\mathcal{S}}_M = \bar{\mathcal{S}}_M(\theta) = \Big\{I\in\mathcal{I} : r^2(I,\theta) \le 2c_3 r^2(\theta) + \frac{\beta}{6(1+B)}M\sigma^2\Big\},
$$

and the event $\bar{A}_M = \bar{A}_M(\theta) = \big\{\sum_{i=1}^{\bar{m}}\xi_{[i]}^2 \le \frac{(1+B)}{\beta}\bar{m}\log\big(\frac{en}{\bar{m}}\big) + \frac{M}{6}\big\}$, where $\bar{m} = \bar{m}_M(\theta) = \max\{|I| : I\in\bar{\mathcal{S}}_M\}$. Introduce the notation $\bar{\Delta}_M(\theta) = M_1 r^2(\theta) + M\sigma^2$ for brevity. By the definition of $\hat{\theta}$ and the Cauchy–Schwarz inequality, we have that $\|\hat{\theta}-\theta\|^2 \le \sum_{I\in\mathcal{I}}\|X(I)-\theta\|^2\hat{\pi}(I|X)$, where $\|X(I)-\theta\|^2 = \sigma^2\sum_{i\in I}\xi_i^2 + \sum_{i\in I^c}\theta_i^2$. Using this, we derive

$$
\begin{aligned}
&\mathrm{P}_\theta\big(\|\hat{\theta}-\theta\|^2 \ge \bar{\Delta}_M(\theta)\big)\\
&\quad\le \mathrm{P}_\theta\Big(\sum_{I\in\mathcal{I}}\big\|X(I)-\theta\big\|^2\hat{\pi}(I|X) \ge \bar{\Delta}_M(\theta)\Big)\\
&\quad\le \mathrm{P}_\theta\big(\bar{A}_M^c\big) + \mathrm{P}_\theta\Big(\Big\{\sum_{I\in\bar{\mathcal{S}}_M}\Big[\sigma^2\sum_{i\in I}\xi_i^2 + \sum_{i\in I^c}\theta_i^2\Big]\hat{\pi}(I|X) \ge \bar{\Delta}_M(\theta)/2\Big\} \cap \bar{A}_M\Big)\\
&\qquad + \mathrm{P}_\theta\Big(\sum_{I\in\bar{\mathcal{S}}_M^c}\Big[\sigma^2\sum_{i\in I}\xi_i^2 + \sum_{i\in I^c}\theta_i^2\Big]\hat{\pi}(I|X) \ge \bar{\Delta}_M(\theta)/2\Big) = \bar{T}_1 + \bar{T}_2 + \bar{T}_3.
\end{aligned}
$$

Similar to (34), we bound the term $\bar{T}_1$ by Lemma 3 (see also (33)):

$$\bar{T}_1 = \mathrm{P}_\theta\big(\bar{A}_M^c\big) = \mathrm{P}\bigg(\sum_{i=1}^{\bar{m}} \xi_{[i]}^2 > \frac{(1+B)}{\beta}\bar{m}\log\Big(\frac{en}{\bar{m}}\Big) + \frac{M}{6}\bigg) \le \exp\big\{-M\beta/6\big\}.$$

Now we evaluate the term $\bar{T}_2$. Since $\bar{m} = \max\{|I| : I \in \bar{\mathcal{S}}_M\}$, $\sigma^2\bar{m}\log\big(\frac{en}{\bar{m}}\big) \le 2c_3r^2(\theta) + \frac{\beta}{6(1+B)}M\sigma^2$. Then for any $I \in \bar{\mathcal{S}}_M$, the event $\bar{A}_M$ implies that $\sum_{i\in I}\xi_i^2 \le \sum_{i=1}^{\bar{m}}\xi_{[i]}^2 \le \frac{(1+B)}{\beta}\bar{m}\log\big(\frac{en}{\bar{m}}\big) + \frac{M}{6} \le \frac{2c_3(1+B)}{\beta}\frac{r^2(\theta)}{\sigma^2} + \frac{M}{3}$. Also $\sum_{i\in I^c}\theta_i^2 \le r^2(I,\theta) \le 2c_3r^2(\theta) + \frac{\beta}{6(1+B)}M\sigma^2$ for any $I \in \bar{\mathcal{S}}_M$. Hence, for any $I \in \bar{\mathcal{S}}_M$, we obtain the implication

$$\bar{A}_M \subseteq \bigg\{\sigma^2\sum_{i\in I}\xi_i^2 + \sum_{i\in I^c}\theta_i^2 \le \frac{2c_3(1+B+\beta)}{\beta}r^2(\theta) + \Big(\frac{1}{3} + \frac{\beta}{6(1+B)}\Big)M\sigma^2\bigg\}.$$

As $M_1 = 4c_3(1+B+\beta)/\beta$, $\beta \in (0,1]$ and $B > 0$, the last relation entails

$$\begin{aligned}
\bar{T}_2 &= \mathrm{P}_\theta\bigg(\Big\{\sum_{I\in\bar{\mathcal{S}}_M}\Big(\sigma^2\sum_{i\in I}\xi_i^2 + \sum_{i\in I^c}\theta_i^2\Big)\hat{\pi}(I|X) \ge \frac{\bar{\Delta}_M}{2}\Big\} \cap \bar{A}_M\bigg) \\
&\le \mathrm{P}_\theta\bigg(\frac{2c_3(1+B+\beta)}{\beta}r^2(\theta) + \Big(\frac{1}{3} + \frac{\beta}{6(1+B)}\Big)M\sigma^2 \ge \frac{M_1}{2}r^2(\theta) + \frac{M}{2}\sigma^2\bigg) = 0.
\end{aligned}$$

It remains to handle the term $\bar{T}_3$. Applying first the Markov inequality and then the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned}
\bar{T}_3 &\le \frac{\mathrm{E}_\theta\big(\sum_{I\in\bar{\mathcal{S}}_M^c}\big[\sigma^2\sum_{i\in I}\xi_i^2 + \sum_{i\in I^c}\theta_i^2\big]\hat{\pi}(I|X)\big)}{\bar{\Delta}_M(\theta)/2} \\
&\le \frac{\sum_{I\in\bar{\mathcal{S}}_M^c}\Big(\sigma^2\big[\mathrm{E}_\theta\big(\sum_{i\in I}\xi_i^2\big)^2\big]^{1/2}\big[\mathrm{E}_\theta\big(\hat{\pi}(I|X)\big)^2\big]^{1/2} + r^2(I,\theta)\mathrm{E}_\theta\hat{\pi}(I|X)\Big)}{\bar{\Delta}_M(\theta)/2} = T_{31} + T_{32}.
\end{aligned}$$

For any $I \in \bar{\mathcal{S}}_M^c$, we have $c_3r^2(\theta) \le \frac{r^2(I,\theta)}{2} - \frac{\beta}{12(1+B)}M\sigma^2$, yielding the bound

$$\frac{c_2}{2}\big(r^2(I,\theta) - c_3r^2(\theta)\big) \ge C_1r^2(I,\theta) + C_2M\sigma^2 \quad \text{for any } I \in \bar{\mathcal{S}}_M^c, \tag{38}$$

where $C_1 = c_2/4$ and $C_2 = c_2\beta/[24(1+B)]$. By (38) and Lemma 2,

$$\big[\mathrm{E}_\theta\hat{\pi}(I|X)\big]^{1/2} \le \Big(\frac{ne}{|I|}\Big)^{-c_1|I|/2}\exp\big\{-C_1\sigma^{-2}r^2(I,\theta) - C_2M\big\} \quad \text{for any } I \in \bar{\mathcal{S}}_M^c. \tag{39}$$

Since $c_1 > 2$, (35) gives $\sum_{I\in\mathcal{I}}\big(\frac{ne}{|I|}\big)^{-c_1|I|/2} \le (1-e^{-c_1/2})^{-1} \triangleq C_3$. According to (15) with $\rho = \min\{C_1, B/2\}$, $\big[\mathrm{E}\big(\sum_{i\in I}\xi_i^2\big)^2\big]^{1/2} \le \frac{B}{\beta\rho}\exp\{\rho|I|\}$. If $M \in [0,1]$, the claim (ii) holds for any

$H_1 \geq e^{m_1}$. Let $M \geq 1$, then $\sigma^2/\bar{\Delta}_M(\theta) \leq M^{-1} \leq 1$. Besides, $\sigma^{-2} r^2(I,\theta) \geq |I| \log(en/|I|) \geq |I|$. Piecing all these relations together with (39), we derive

$$
\begin{aligned}
T_{31} &\leq \frac{2B}{\beta\rho} \sum_{I \in \bar{\mathcal{S}}_M^c} \exp\big\{\rho|I|\big\}\Big(\frac{ne}{|I|}\Big)^{-c_1|I|/2} \exp\big\{-C_1\sigma^{-2}r^2(I,\theta) - C_2 M\big\} \\
&\leq C_4 \exp\{-C_2 M\},
\end{aligned}
$$

where $C_4 = 2BC_3/(\beta\rho) = 2BC_3/\big(\beta \min\{C_1, B\}\big)$. Finally, by (35), (39) and the facts that $\max_{x\geq 0}\{xe^{-cx}\} \leq (ce)^{-1}$ (for any $c > 0$) and $\sigma^2/\bar{\Delta}_M(\theta) \leq 1$, we bound the term T_{32}:

$$
\begin{aligned}
T_{32} &= \frac{2}{\bar{\Delta}_M(\theta)} \sum_{I \in \bar{\mathcal{S}}_M^c} r^2(I,\theta) \mathrm{E}_\theta \hat{\pi}(I|X) \\
&\leq \frac{2}{\bar{\Delta}_M(\theta)} \sum_{I \in \bar{\mathcal{S}}_M^c} r^2(I,\theta)\Big(\frac{ne}{|I|}\Big)^{-c_1|I|} \exp\big\{-2C_1\sigma^{-2}r^2(I,\theta) - 2C_2 M\big\} \\
&\leq C_5 \exp\{-2C_2 M\},
\end{aligned}
$$

where $C_5 = C_0/(C_1 e)$. The assertion (ii) is proved since we showed that $\mathrm{P}_\theta(\|\hat{\theta} - \theta\|^2 \geq M_1 r^2(\theta) + M\sigma^2) \leq H_1 e^{-m_1 M}$ with $M_1 = 4c_3(1 + B + \beta)/\beta$, $H_1 = \max\big\{1 + C_4 + C_5, e^{m_1}\big\}$, $m_1 = \min\big\{\frac{\beta}{6}, C_2\big\}$. $\square$

Proof of Theorem 2. First, we prove (i). If the inequality $|I \backslash I_o| \log\big(\frac{en}{|I|}\big) < \sum_{i \in I\backslash I_o} \frac{\theta_i^2}{\sigma^2}$ would hold for some $I \in \mathcal{I}$, then

$$
\begin{aligned}
r^2(I \cup I_o, \theta) &= \sum_{i \notin I \cup I_o} \theta_i^2 + \sigma^2 |I \cup I_o| \log\Big(\frac{en}{|I \cup I_o|}\Big) \\
&\leq \sum_{i \notin I \cup I_o} \theta_i^2 + \sigma^2 |I\backslash I_o| \log\Big(\frac{en}{|I|}\Big) + \sigma^2 |I_o| \log\Big(\frac{en}{|I_o|}\Big) \\
&< \sum_{i \notin I \cup I_o} \theta_i^2 + \sum_{i \in I\backslash I_o} \frac{\theta_i^2}{\sigma^2} + \sigma^2 |I_o| \log\Big(\frac{en}{|I_o|}\Big) \\
&= \sum_{i \notin I_o} \theta_i^2 + \sigma^2 |I_o| \log\Big(\frac{en}{|I_o|}\Big) = r^2(\theta),
\end{aligned}
$$

which contradicts the definition of the oracle. Hence, $\sum_{i \in I\backslash I_o} \frac{\theta_i^2}{\sigma^2} \leq |I\backslash I_o| \log\big(\frac{en}{|I|}\big)$ for any $I \in \mathcal{I}$. Define $c_4 = \varkappa\beta - \frac{\beta}{2} - B - 1$ and note that $c_4 > 1$ by the condition of the theorem. Using the relation $\sum_{i \in I\backslash I_o} \frac{\theta_i^2}{\sigma^2} \leq |I\backslash I_o| \log\big(\frac{en}{|I|}\big) \leq |I| \log\big(\frac{en}{|I|}\big)$ and Lemma 1 with $h = \beta$ and $I_0 = I_o \cap I$ (so that $I\backslash I_0 = I\backslash I_o$), we obtain for each $I \in \mathcal{G}_1 = \big\{I \in \mathcal{I} : |I| \log(\frac{en}{|I|}) \geq M_0' |I_0| \log\big(\frac{en}{|I_0|}\big) + M\big\}$

with $M_0' = \varkappa\beta + \frac{\beta}{2}$,

$$
\begin{aligned}
&\mathrm{E}_\theta \hat{\pi}(I|X) \\
&\quad \le \Big(\frac{\lambda_I}{\lambda_{I_0}}\Big)^{\beta} \exp\Big\{\beta \sum_{i\in I\setminus I_0} \frac{\theta_i^2}{\sigma^2} + \frac{\beta}{2}|I_0|\log\Big(\frac{en}{|I_0|}\Big) - \Big(\frac{\beta}{2} - B\Big)|I|\log\Big(\frac{en}{|I|}\Big)\Big\} \\
&\quad \le \Big(\frac{ne}{|I|}\Big)^{-c_4|I|} \exp\Big\{-\Big(\varkappa\beta - \frac{\beta}{2} - B - c_4\Big)|I|\log\Big(\frac{en}{|I|}\Big) + \Big(\beta\varkappa + \frac{\beta}{2}\Big)|I_0|\log\Big(\frac{en}{|I_0|}\Big)\Big\} \\
&\quad = \Big(\frac{ne}{|I|}\Big)^{-c_4|I|} \exp\Big\{-|I|\log\Big(\frac{en}{|I|}\Big) + \Big(\beta\varkappa + \frac{\beta}{2}\Big)|I_0|\log\Big(\frac{en}{|I_0|}\Big)\Big\} \\
&\quad \le \Big(\frac{ne}{|I|}\Big)^{-c_4|I|} \exp\Big\{-\Big(M_0' - \varkappa\beta - \frac{\beta}{2}\Big)|I_0|\log\Big(\frac{en}{|I_0|}\Big) - M\Big\} = \Big(\frac{ne}{|I|}\Big)^{-c_4|I|} e^{-M}.
\end{aligned}
$$

Since $c_4 > 1$, by the same reasoning as in (6) we bound $\sum_{I\in\mathcal{I}} \big(\frac{ne}{|I|}\big)^{-c_4|I|} \le (1 - e^{1-c_4})^{-1} \triangleq H_0'$. Using this and the last display, we finish the proof of (i):

$$
\mathrm{E}_\theta \hat{\pi}(I \in \mathcal{G}_1|X) = \sum_{I\in\mathcal{G}_1} \mathrm{E}_\theta \hat{\pi}(I|X) \le e^{-M} \sum_{I\in\mathcal{I}} \Big(\frac{ne}{|I|}\Big)^{-c_4|I|} \le H_0' e^{-M}.
$$

Next, we prove (ii). Define $\mathcal{G}_2 = \mathcal{G}_2(I') = \big\{I \in \mathcal{I} : \sum_{i\in I'\setminus I} \frac{\theta_i^2}{\sigma^2} \ge \bar{\tau}|I \cup I'|\log\big(\frac{en}{|I\cup I'|}\big) + M\big\}$. Using (5) and Lemma 1 with $h = \beta$ and $I_0 = I_0(I,\theta) = I \cup I'$, we evaluate for each $I \in \mathcal{G}_2$

$$
\begin{aligned}
&\mathrm{E}_\theta \hat{\pi}(I|X) \\
&\quad \le \Big(\frac{\lambda_I}{\lambda_{I_0}}\Big)^{\beta} \exp\Big\{-\frac{\beta}{3} \sum_{i\in I_0\setminus I} \frac{\theta_i^2}{\sigma^2} + \Big(\frac{\beta}{2} + B\Big)|I_0|\log\Big(\frac{en}{|I_0|}\Big) - \frac{\beta}{2}|I|\log\Big(\frac{en}{|I|}\Big)\Big\} \\
&\quad = \Big(\frac{\lambda_I}{c_{\varkappa,n}}\Big)^{\beta} \exp\Big\{-\frac{\beta}{3} \sum_{i\in I'\setminus I} \frac{\theta_i^2}{\sigma^2} + \Big(\varkappa\beta + \frac{\beta}{2} + B\Big)\big|I \cup I'\big|\log\Big(\frac{en}{|I\cup I'|}\Big) - \frac{\beta}{2}|I|\log\Big(\frac{en}{|I|}\Big)\Big\} \\
&\quad \le \Big(\frac{\lambda_I}{c_{\varkappa,n}}\Big)^{\beta + \frac{\beta}{2\varkappa}} \exp\Big\{\Big(-\frac{\beta}{3}\bar{\tau} + \varkappa\beta + \frac{\beta}{2} + B\Big)\big|I \cup I'\big|\log\Big(\frac{en}{|I\cup I'|}\Big) - \frac{\beta}{3}M\Big\} \\
&\quad \le \Big(\frac{\lambda_I}{c_{\varkappa,n}}\Big)^{\beta + \frac{\beta}{2\varkappa}} e^{-\beta M/3}.
\end{aligned}
$$

Since $\varkappa > \frac{1}{\beta} - \frac{1}{2}$, by the same reasoning as in (6) we bound $\sum_I \big(\frac{\lambda_I}{c_{\varkappa,n}}\big)^{\beta(1+1/2\varkappa)} \le \big(1 - e^{1-\varkappa\beta-\beta/2}\big)^{-1} \triangleq H_1'$. This relation and the last display imply claim (ii): with $m_0' = \frac{\beta}{3}$,

$$
\mathrm{E}_\theta \hat{\pi}(I \in \mathcal{G}_2|X) = \sum_{I\in\mathcal{G}_2} \mathrm{E}_\theta \hat{\pi}(I|X) \le H_1' \exp\big\{-m_0' M\big\}. \tag{40}
$$

Let us derive the second claim of (ii). If $|I|\log(\frac{en}{|I|}) \le \varrho|\bar{I}_o|\log(\frac{en}{|\bar{I}_o|}) - M$, then $|I \cup \bar{I}_o| \times \log(\frac{en}{|I\cup\bar{I}_o|}) \le |I|\log(\frac{en}{|I|}) + |\bar{I}_o|\log(\frac{en}{|\bar{I}_o|}) \le (1+\varrho)|\bar{I}_o|\log(\frac{en}{|\bar{I}_o|}) - M$. Hence, $|\bar{I}_o|\log(\frac{en}{|\bar{I}_o|}) \ge \frac{1}{1+\varrho}|I \cup \bar{I}_o|\log(\frac{en}{|I\cup\bar{I}_o|}) + \frac{M}{1+\varrho}$, which, together with the definition of the τ-oracle, imply

$$
\begin{aligned}
\sum_{i\in\bar{I}_o\setminus I}\frac{\theta_i^2}{\sigma^2} &\ge \Big(\sum_{i\in I^c}\frac{\theta_i^2}{\sigma^2} - \sum_{i\in\bar{I}_o^c}\frac{\theta_i^2}{\sigma^2}\Big)\\
&\ge \tau_0\Big(|\bar{I}_o|\log\Big(\frac{en}{|\bar{I}_o|}\Big) - |I|\log\Big(\frac{en}{|I|}\Big)\Big)\\
&\ge \tau_0(1-\varrho)|\bar{I}_o|\log\Big(\frac{en}{|\bar{I}_o|}\Big) + \tau_0 M\\
&\ge \bar{\tau}|I\cup\bar{I}_o|\log\Big(\frac{en}{|I\cup\bar{I}_o|}\Big) + \frac{2\tau_0}{1+\varrho}M,
\end{aligned}
\tag{41}
$$

as $\frac{1-\varrho}{1+\varrho}\tau_0 \ge \bar{\tau}$ by the condition of the theorem. Thus, we obtain

$$
\begin{aligned}
&\mathrm{E}_\theta\hat{\pi}\Big(I: |I|\log\Big(\frac{en}{|I|}\Big) \le \varrho|\bar{I}_o|\log\Big(\frac{en}{|\bar{I}_o|}\Big) - M\big|X\Big)\\
&\le \mathrm{E}_\theta\hat{\pi}\Big(\sum_{i\in\bar{I}_o\setminus I}\frac{\theta_i^2}{\sigma^2} \ge \bar{\tau}|I\cup\bar{I}_o|\log\Big(\frac{en}{|I\cup\bar{I}_o|}\Big) + \frac{2\tau_0}{1+\varrho}M\big|X\Big).
\end{aligned}
$$

By this and (40) with $I' = \bar{I}_o$, the second claim of (ii) follows with $m_1' = \frac{2\tau_0 m_0'}{1+\varrho}$.

Finally, let us prove (iii). Denote $\mathcal{G}_3 = \mathcal{G}_3(\theta, M) = \big\{I : r^2(I,\theta) \ge c_3 r^2(\theta) + M\sigma^2\big\}$, where the constants $c_1 > 2$, c_2, c_3 are defined in Lemma 2. Applying Lemma 2 and using the fact (35), we complete the proof of (iii):

$$
\mathrm{E}_\theta\hat{\pi}(I\in\mathcal{G}_3|X) = \sum_{I\in\mathcal{G}_3}\mathrm{E}_\theta\hat{\pi}(I|X) \le e^{-c_2 M}\sum_{I\in\mathcal{I}}\Big(\frac{ne}{|I|}\Big)^{-c_1|I|} \le C_0 e^{-c_2 M}.
$$

□

Proof of Theorem 3. We first establish the coverage property. The constants M_1, H_1 and m_1 are defined in Theorem 1, the constant ϱ is from (22). Take some $M_2 \ge \frac{M_1}{\varrho}$, for example, $M_2 = \frac{M_1}{\varrho} + 1$. From (19) and (26), it follows that $r^2(\theta) \le r^2(\bar{I}_o,\theta) = (b(\theta)+1)\sigma^2|\bar{I}_o|\log(\frac{en}{|\bar{I}_o|}) + b(\theta)\sigma^2 \le (b(\theta)+1)\sigma^2\big(|\bar{I}_o|\log(\frac{en}{|\bar{I}_o|}) + 1\big)$. Combining this with claims (ii) from Theorems 1 and 2 and the definition (24) of $\hat{r}$ yields the coverage property:

$$
\begin{aligned}
&\mathrm{P}_\theta\Big(\theta\notin B\big(\hat{\theta}, \big[(b(\theta)+1)M_2\hat{r}^2 + (b(\theta)+2)M\sigma^2\big]^{1/2}\Big)\\
&\le \mathrm{P}_\theta\Big(\|\hat{\theta}-\theta\|^2 > (b(\theta)+1)M_2\hat{r}^2 + (b(\theta)+2)M\sigma^2, \hat{r}^2 \ge \varrho\sigma^2|\bar{I}_o|\log\Big(\frac{en}{|\bar{I}_o|}\Big) + \sigma^2 - \frac{M\sigma^2}{M_2}\Big)
\end{aligned}
$$

$$
\begin{aligned}
&+ \mathrm{P}_\theta\Big(\hat{r}^2 < \varrho\sigma^2|\bar{I}_o|\log\Big(\frac{en}{|\bar{I}_o|}\Big) + \sigma^2 - \frac{M\sigma^2}{M_2}\Big) \\
&\le \mathrm{P}_\theta\big(\|\hat{\theta}-\theta\|^2 > \varrho M_2 r^2(\theta) + M\sigma^2\big) + \mathrm{P}_\theta\Big(|\hat{I}|\log\Big(\frac{en}{|\hat{I}|}\Big) < \varrho|\bar{I}_o|\log\Big(\frac{en}{|\bar{I}_o|}\Big) - \frac{M}{M_2}\Big) \\
&\le H_1 e^{-m_1 M} + H_1' e^{-m_1'' M} \le H_2 e^{-m_2 M},
\end{aligned}
$$

where $m_1'' = m_1'/M_2$, $H_2 = H_1 + H_1'$, $m_2 = \min\{m_1, m_1''\}$; H_1', m_1' are defined in Theorem 2 and the constant ϱ is from (22). As $b(\theta) \le t$ for all $\theta \in \Theta_{\mathrm{eb}}(t)$, the coverage property follows.

The size property follows from the definition (24) of $\hat{r}$ and property (i) of Theorem 2. Indeed,

$$
\begin{aligned}
&\mathrm{P}_\theta\Big(\hat{r}^2 \ge \sigma^2 M_0'|I_o|\log\Big(\frac{en}{|I_o|}\Big) + (M+1)\sigma^2\Big) \\
&\quad = \mathrm{P}_\theta\Big(|\hat{I}|\log\Big(\frac{en}{|\hat{I}|}\Big) \ge M_0'|I_o|\log\Big(\frac{en}{|I_o|}\Big) + M\Big) \\
&\quad \le \mathrm{P}_\theta\Big(|\hat{I}|\log\Big(\frac{en}{|\hat{I}|}\Big) \ge M_0'|I\cap I_o|\log\Big(\frac{en}{|I\cap I_o|}\Big) + M\Big) \le H_0' e^{-M}. \qquad \square
\end{aligned}
$$

Supplementary Material

Supplement to "Needles and straw in a haystack: Robust confidence for possibly sparse sequences" (DOI: 10.3150/19-BEJ1122SUPP; .pdf). The elaboration on some points and some background information related to the paper is provided in Supplement [6].

References

[1] Abramovich, F., Grinshtein, V. and Pensky, M. (2007). On optimality of Bayesian testimation in the normal means problem. *Ann. Statist.* **35** 2261–2286. MR2363971 https://doi.org/10.1214/009053607000000226

[2] 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 https://doi.org/10.3103/S1066530710030026

[3] Baraud, Y. (2004). Confidence balls in Gaussian regression. *Ann. Statist.* **32** 528–551. MR2060168 https://doi.org/10.1214/009053604000000085

[4] Belitser, E. (2017). On coverage and local radial rates of credible sets. *Ann. Statist.* **45** 1124–1151. MR3662450 https://doi.org/10.1214/16-AOS1477

[5] Belitser, E. and Ghosal, S. (2018). Empirical Bayes oracle uncertainty quantification for regression. Preprint.

[6] Belitser, E. and Nurushev, N. (2020). Supplement to "Needles and straw in a haystack: Robust confidence for possibly sparse sequences." https://doi.org/10.3150/19-BEJ1122SUPP.

[7] Bhattacharya, A., Dunson, D.B., Pati, D. and Pillai, N.S. (2016). Sub-optimality of some continuous shrinkage priors. *Stochastic Process. Appl.* **126** 3828–3842. MR3565480 https://doi.org/10.1016/j.spa.2016.08.007

[8] Birgé, L. and Massart, P. (2001). Gaussian model selection. *J. Eur. Math. Soc.* (*JEMS*) **3** 203–268. MR1848946 https://doi.org/10.1007/s100970100031
[9] Bull, A.D. (2012). Honest adaptive confidence bands and self-similar functions. *Electron. J. Stat.* **6** 1490–1516. MR2988456 https://doi.org/10.1214/12-EJS720
[10] Bull, A.D. and Nickl, R. (2013). Adaptive confidence sets in L^2. *Probab. Theory Related Fields* **156** 889–919. MR3078289 https://doi.org/10.1007/s00440-012-0446-z
[11] Cai, T.T. and Low, M.G. (2004). An adaptation theory for nonparametric confidence intervals. *Ann. Statist.* **32** 1805–1840. MR2102494 https://doi.org/10.1214/009053604000000049
[12] Castillo, I., Schmidt-Hieber, J. and van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *Ann. Statist.* **43** 1986–2018. MR3375874 https://doi.org/10.1214/15-AOS1334
[13] Castillo, I. and Szabó, B. (2018). Spike and slab empirical Bayes sparse credible sets. Available at arXiv:1808.07721.
[14] 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 https://doi.org/10.1214/12-AOS1029
[15] Donoho, D.L. and Johnstone, I.M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455. MR1311089 https://doi.org/10.1093/biomet/81.3.425
[16] Donoho, D.L. and Johnstone, I.M. (1994). Minimax risk over l_p-balls for l_q-error. *Probab. Theory Related Fields* **99** 277–303. MR1278886 https://doi.org/10.1007/BF01199026
[17] Donoho, D.L., Johnstone, I.M., Hoch, J.C. and Stern, A.S. (1992). Maximum entropy and the nearly black object (with discussion). *J. Roy. Statist. Soc. Ser. B* **54** 41–81. MR1157714
[18] Johnstone, I. (2017). Gaussian estimation: Sequence and wavelet models. Book draft.
[19] 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 https://doi.org/10.1214/009053604000000030
[20] Li, K.-C. (1989). Honest confidence regions for nonparametric regression. *Ann. Statist.* **17** 1001–1008. MR1015135 https://doi.org/10.1214/aos/1176347253
[21] Martin, R. and Walker, S.G. (2014). Asymptotically minimax empirical Bayes estimation of a sparse normal mean vector. *Electron. J. Stat.* **8** 2188–2206. MR3273623 https://doi.org/10.1214/14-EJS949
[22] Nickl, R. and van de Geer, S. (2013). Confidence sets in sparse regression. *Ann. Statist.* **41** 2852–2876. MR3161450 https://doi.org/10.1214/13-AOS1170
[23] Picard, D. and Tribouley, K. (2000). Adaptive confidence interval for pointwise curve estimation. *Ann. Statist.* **28** 298–335. MR1762913 https://doi.org/10.1214/aos/1016120374
[24] Robins, J. and van der Vaart, A. (2006). Adaptive nonparametric confidence sets. *Ann. Statist.* **34** 229–253. MR2275241 https://doi.org/10.1214/009053605000000877
[25] Ročková, V. (2018). Bayesian estimation of sparse signals with a continuous spike-and-slab prior. *Ann. Statist.* **46** 401–437. MR3766957 https://doi.org/10.1214/17-AOS1554
[26] Rousseau, J. and Szabó, B. (2016). Asymptotic frequentist coverage properties of Bayesian credible sets for sieve priors. Available at arXiv:1609.05067.
[27] Rousseau, J. and Szabo, B. (2017). Asymptotic behaviour of the empirical Bayes posteriors associated to maximum marginal likelihood estimator. *Ann. Statist.* **45** 833–865. MR3650402 https://doi.org/10.1214/16-AOS1469
[28] Szabó, B., van der Vaart, A.W. and van Zanten, J.H. (2015). Frequentist coverage of adaptive nonparametric Bayesian credible sets. *Ann. Statist.* **43** 1391–1428. MR3357861 https://doi.org/10.1214/14-AOS1270
[29] van der Pas, S., Szabó, B. and van der Vaart, A. (2017). Uncertainty quantification for the horseshoe (with discussion). *Bayesian Anal.* **12** 1221–1274. MR3724985 https://doi.org/10.1214/17-BA1065

[30] van der Pas, S.L., Kleijn, B.J.K. and van der Vaart, A.W. (2014). The horseshoe estimator: Posterior concentration around nearly black vectors. *Electron*. *J*. *Stat*. **8** 2585–2618. MR3285877 https://doi.org/10.1214/14-EJS962

Received March 2018 and revised January 2019