The Annals of Statistics
2019, Vol. 47, No. 1, 93–126
https://doi.org/10.1214/17-AOS1680

ADAPTIVE ESTIMATION OF THE SPARSITY IN THE GAUSSIAN VECTOR MODEL

BY ALEXANDRA CARPENTIER[1] AND NICOLAS VERZELEN

OvGU Magdeburg and INRA

Consider the Gaussian vector model with mean value θ. We study the twin problems of estimating the number $\|\theta\|_0$ of nonzero components of θ and testing whether $\|\theta\|_0$ is smaller than some value. For testing, we establish the minimax separation distances for this model and introduce a minimax adaptive test. Extensions to the case of unknown variance are also discussed. Rewriting the estimation of $\|\theta\|_0$ as a multiple testing problem of all hypotheses $\{\|\theta\|_0 \leq q\}$, we both derive a new way of assessing the optimality of a sparsity estimator and we exhibit such an optimal procedure. This general approach provides a roadmap for estimating the complexity of the signal in various statistical models.

1. Introduction. Many estimation methods in high or infinite-dimensional statistics rely on the assumption that the parameter of interest belongs to some smaller parameter space. Depending on the problem at hand, the assumptions on the structure of the unknown parameter take various forms. In high-dimensional linear regression, it is usually assumed that the regression parameter is sparse [4]. In matrix completion, the underlying matrix may be supposed to be low-rank [30]. In density estimation, many nonparametric methods are based on the assumption that the function satisfies some smoothness properties [21]. Many model-based clustering methods require the data to follow a mixture distribution with several Gaussian components [22]. In practice, the exact complexity of the parameter (e.g., the rank of the matrix, the smoothness of the function) is unknown. Although a lot of work has been devoted to the construction of statistical procedures performing as well as if the model complexity was known (e.g., [4, 20, 35]), the literature on the estimation of the complexity of the parameter is scarcer.

In this paper, we are interested in the twin problems of estimating the complexity of the parameter and testing whether the parameter belongs to some complexity

Received March 2017; revised September 2017.

[1] Supported in part by the Deutsche Forschungsgemeinschaft (DFG) Emmy Noether grant MuSyAD (CA 1488/1-1) and by the Deutsche Forschungsgemeinschaft (DFG) through grant CRC 1294 “Data Assimilation—The seamless integration of data and models,” Project A03 “Sequential and adaptive learning under dependence and nonstandard objective functions,” by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—314838170, GRK 2297 MathCoRe.

MSC2010 subject classifications. Primary 62C20, 62G10; secondary 62B10.

Key words and phrases. Sparsity estimation and testing, composite-composite testing problems, minimax separation distance in testing problems.

class. There are several motivations for these problems. First, complexity estimation allows to assess the relevance of specific parameter estimation approaches. For instance, inferring the smoothness of a function allows to justify the use of regularity-based procedures. Second, the construction of adaptive confidence regions is related to the model testing problem since the size of an adaptive confidence region should depend on the complexity of the unknown parameter [23]. Finally, in some practical applications, the primary objective is rather to evaluate the complexity of the parameter than the parameter itself. This is for instance the case in some heritability studies where the goal is to decipher whether a trait is multigenic or "highly polygenic" which amounts to inferring whether a high-dimensional regression parameter is sparse or dense [34, 41].

In this paper, we focus on a comparatively simple, yet emblematic setting, namely the Gaussian vector model, that we define as follows:

$$Y_i = \theta_i + \varepsilon_i, \qquad i = 1, \dots, n, \tag{1}$$

where $\theta = (\theta_i) \in \mathbb{R}^n$ is unknown and the noise components ε_i are independent and follow a centered normal distributions with variance σ^2. We are interested in (i) estimating the number $\|\theta\|_0$ of nonzero components of θ and (ii) given some nonnegative integer k_0, testing whether $\|\theta\|_0 \le k_0$ or $\|\theta\|_0 > k_0$. The former problem is called sparsity estimation and the latter sparsity testing.

1.1. *Sparsity testing and separation distances.* As the sparsity testing problem is easier to formalize than the sparsity estimation problem, let us first be more specific about it. Given a nonnegative integer $k_0 \in [0, n]$, we write

$$\mathbb{B}_0[k_0] := \{\theta \in \mathbb{R}^n : \|\theta\|_0 \le k_0\}, \tag{2}$$

for the set of k_0-sparse vectors θ, that is to say the set of vectors θ with less than k_0 nonzero coefficients. Our goal is to test whether θ belongs to $\mathbb{B}_0[k_0]$ or not. Before describing our results and the literature, we shall first define the notion of minimax separation distance of a test.

Let $\|\cdot\|_2$ stand for the Euclidean norm in $\mathbb{R}^n$. For any $\theta \in \mathbb{R}^n$, we write $d_2(\theta, \mathbb{B}_0[k_0]) := \inf_{u \in \mathbb{B}_0[k_0]} \|\theta - u\|_2$ for the distance of θ to the set of k_0-sparse vectors. Intuitively, any α-level test T of the null hypothesis $\{\theta \in \mathbb{B}_0[k_0]\}$ cannot reject the null with high probability when the true parameter θ is arbitrarily close (in the $d_2(\theta, \mathbb{B}_0[k_0])$ sense) to $\mathbb{B}_0[k_0]$. Conversely, any reasonable test should reject the null hypothesis with high probability for parameters θ that are really distant to $\mathbb{B}_0[k_0]$. In order to quantify the performances of a given test T, it is then classical [2, 25] to rely on the notion of separation distance. Given positive integers $k_1 > k_0$ and a real number $\rho > 0$, define

$$\mathbb{B}_0[k_1, k_0, \rho] := \{\theta \in \mathbb{B}_0[k_1] : d_2(\theta, \mathbb{B}_0[k_0]) \ge \rho\}, \tag{3}$$

as the set of k_1-sparse vectors that lie at distance larger than ρ from the null. Then, for a fixed $\Delta > 0$ and $\rho > 0$, we consider the testing problem

$$H_{k_0} : \theta \in \mathbb{B}_0[k_0] \quad \text{versus} \quad H_{\Delta, k_0, \rho} : \theta \in \mathbb{B}_0[k_0 + \Delta, k_0, \rho]. \tag{4}$$

The purpose of this definition is to remove from the alternative hypothesis parameters θ that are too close to the null hypothesis. Given a test T, its risk $R(T; k_0, \Delta, \rho)$ for the above problem (4) is defined as the sum of the type I and type II error probabilities:

$$R(T; k_0, \Delta, \rho) := \sup_{\theta \in \mathbb{B}_0[k_0]} \mathbb{P}_{\theta,\sigma}[T = 1] + \sup_{\theta \in \mathbb{B}_0[k_0+\Delta, k_0, \rho]} \mathbb{P}_{\theta,\sigma}[T = 0]. \tag{5}$$

Here, $\mathbb{P}_{\theta,\sigma}$ stands for the distribution of Y. The function $\rho \mapsto R(T; k_0, \Delta, \rho)$ is nonincreasing and equals at least one for $\rho = 0$. For a fixed $\gamma \in (0, 1)$, the separation distance $\rho_\gamma(T; k_0, \Delta)$ is the largest ρ such that the hypotheses

$$\rho_\gamma(T; k_0, \Delta) := \sup\{\rho > 0 | R(T; k_0, \Delta, \rho) > \gamma\}. \tag{6}$$

The separation distance of a good test T should be the smallest possible. Finally, the minimax separation distance is

$$\rho^*_\gamma[k_0, \Delta] := \inf_T \rho_\gamma(T; k_0, \Delta), \tag{7}$$

where the infimum is taken over all tests T. In other words, $\rho^*_\gamma[k_0, \Delta]$ is the minimal distance to $\mathbb{B}_0[k_0]$ such that some test is able to reliably distinguish parameters in $\mathbb{B}_0[k_0]$ from parameters in $\mathbb{B}_0[k_0 + \Delta, k_0, \rho]$. Hence, it characterizes the difficulty of the testing problem. A test T whose separation distance $\rho_\gamma(T; k_0, \Delta)$ is (up to a multiplicative constant) smaller than $\rho^*_\gamma[k_0, \Delta]$ is said to be minimax.

1.2. *Our contribution.* Our contribution is threefold:

(i) When σ is known, we first establish the minimax separation distance $\rho^*_\gamma[k_0, \Delta]$ for all integers k_0 and all $\Delta > 0$. Besides, we introduce a new test which is minimax adaptive for all Δ.

(ii) In the more realistic setting where the noise level σ is unknown, the minimax separation distance $\rho^*_{\gamma,\mathrm{var}}[k_0, \Delta]$ (defined in Section 3) is established and minimax adaptive tests are exhibited. Interestingly, it is proved that the sparsity testing problem under unknown noise level is no more difficult than under known noise level for small Δ. For large Δ, the knowledge of σ plays an important role.

(iii) We reformulate the sparsity estimation problem as a multiple testing problem where we simultaneously consider all nested hypotheses H_q for $q \in [0, n]$. Introducing a multiple testing procedure which is simultaneously optimal over all q, we derive an estimator $\widehat{k}$ which is less than or equal to $\|\theta\|_0$ with high probability and is also closest to $\|\theta\|_0$ in a minimax sense. Interestingly, this property will be valid for all possible $\theta \in \mathbb{R}^n$ and avoid us to rely on any particular assumption on the parameter. More generally, this perspective also provides a general roadmap to handle the problem of complexity estimation using simultaneous separation distances.

In order to discuss more specifically our contribution, we need to contrast them with earlier results for related problems such as signal detection. This is why we review the related literature before describing in depth our results.

1.3. *Related literature.* Although the twin problems of sparsity testing and sparsity estimation are closely connected, we start by discussing the literature mostly related to the testing version of our problem and then turn to the estimation version.

Signal detection. The signal detection problem which amounts to testing whether $\theta = 0$ is a special instance of the sparsity testing problem (corresponding to $k_0 = 0$). Signal detection in the Gaussian vector model has been extensively studied [2, 15, 17, 25] in the last fifteen years and is now well understood. For instance, it has been established in [15] that the minimax separation distance $\rho^*_\gamma[0, \Delta]$ satisfies

$$\rho^{*2}_\gamma[0, \Delta] \asymp_\gamma \sigma^2 \Delta \log\left(1 + \frac{\sqrt{n}}{\Delta}\right),$$

where $f(\Delta, n) \asymp_\gamma g(\Delta, n)$ means that there exist positive constants c_γ and c'_γ (possibly depending on γ) such that $f(\Delta, n) \leq c_\gamma g(\Delta, n) \leq c'_\gamma f(\Delta, n)$. Besides, some tests are able to simultaneously achieve the above separation distances for all positive Δ.

Looking more closely at the above equation, one can distinguish two main regimes for this problem depending on the sparsity Δ of the alternative: the sparse case ($\Delta \leq \sqrt{n}$) and the dense case ($\Delta > \sqrt{n}$). In the sparse case, $\rho^{*2}_\gamma[0, \Delta]$ is of order $\Delta \log(1 + \sqrt{n}/\Delta)$. This entails that it is possible to detect sparse vectors θ whose nonzero values are of order $\sqrt{\log(n^{1/2}/\Delta)}$. Known optimal tests such as the Higher Criticism test [17] or the one proposed in [15] amount to counting the number of values $|Y_i|$ that are larger than some threshold t and to compare this number to what is expected under the null hypothesis. Doing this simultaneously for a wide range of t leads to near-optimal performances simultaneously for all $\Delta \in [1, \sqrt{n}]$. In the dense case ($\Delta \geq \sqrt{n}$), the situation is qualitatively different as the square minimax separation distance $\rho^{*2}_\gamma[0, \Delta]$ is of order $\sqrt{n}$. A near-optimal test proposed, for example, in [2], is based on the statistic $\|Y\|_2^2/\sigma^2$, which, under the null, follows a χ^2 distribution with n degrees of freedom and, under the alternative, follows a noncentral χ^2 distribution with noncentrality parameter $\|\theta\|_2^2$.

Composite-composite testing problems and functional estimation. For the signal detection problem ($k_0 = 0$), the null hypothesis is simple whereas for the general case $k_0 > 0$, the null hypothesis is composite, thereby making the analysis of the problem more challenging. Although we are not aware of any general treatment of this kind of problem in the literature, some partial results and methods may be derived in our setting from prior work on related problems.

Minimax analysis of composite-composite testing problems has, up to our knowledge, been tackled per se in a few work [3, 12, 16, 28]. As the minimax analysis of confidence regions and functional estimation problems is related to such testing problems, some work in these two fields (e.g., [5, 6, 9, 10, 23, 38] and [7, 8, 11, 19, 32]) have also lead to progress in the understanding of composite-composite problems.

To be more specific on the challenge of composite-composite problems, let us describe a simple strategy called ”infimum testing” [20]. Consider a signal detection test based on the statistic $S(\cdot)$, that rejects the null hypothesis H_{k_0} with $k_0 = 0$ for large values of $S(Y)$. The corresponding infimum test for the composite-composite testing problem (4), is a test rejecting H_{k_0} for large values of $S_{\inf} := \inf_{u\in\mathbb{B}_0[k_0]} S(Y-u)$. Indeed, there exists, under the null hypothesis H_{k_0}, some u such that the expectation of $Y-u$ is zero. As one may expect, considering this infimum over all possible parameters in the null hypothesis is not priceless and the separation distance $\rho_\gamma[T_{k_0}; k_0, \Delta]$ of the corresponding infimum test T_{k_0} may depend on the complexity of the null hypothesis. Conversely, simple inclusion arguments that will be recalled in our proofs entail that the composite problem is at least as difficult as the signal detection problem, that is, $\rho^*_\gamma[k_0, \Delta]$ is at least of the order of $\rho^*_\gamma[0, \Delta]$. The main challenge is therefore to decipher whether $\rho^*_\gamma[k_0, \Delta]$ is indeed of order $\rho^*_\gamma[0, \Delta]$ or if it is larger than that and really depends on k_0. In other words, we seek to understand how the complexity of the null hypothesis influences the difficulty of the testing problem.

Sparsity estimation. Closer to our setting, Cai, Jin and Low [8] study the problem of estimating $\|\theta\|_0$ for sparse vectors θ such that $\|\theta\|_0 \leq \sqrt{n}$. They consider a Bayesian framework, where each component θ_i is drawn independently from a two points mixture distribution $(1-\eta)\delta_0 + \eta\delta_a$ for some unknown $a > 0$ (δ_x denotes the Dirac measure at x). The goal is then to estimate $\eta = \mathbb{E}[\|\theta\|_0]/n$. Relying on the tail distribution of Y, they introduce an estimator $\widehat{\eta}$ that satisfies $\widehat{\eta} \leq \eta$ with high probability and such that the risk $\mathbb{E}[|1-\widehat{\eta}/\eta|]$ is as small as possible. In [26], Jin introduced a class of estimators of θ based on the empirical characteristic function of Y to handle the denser case $\|\theta\|_0 \geq \sqrt{n}$. Later, these procedures have been extended [7, 27] to allow for unknown noise level σ and even unknown mean in the more general model $Y_i = u + \theta_i + \varepsilon_i$, where u is unknown. Again, in a Bayesian framework where all θ_i's follow the same mixture distribution $(1-\eta)\delta_0 + \eta\pi$ for some smooth density π, their estimator $\widehat{\eta}$ is proved to achieve an optimal minimax rate.

In multiple testing, estimating the number of false hypotheses has a longer history. Rephrased in the Gaussian vector model, multiple hypotheses testing amounts to testing simultaneously whether each θ_i is zero or not. Hence, estimating the number of false hypotheses is equivalent to sparsity estimation. Nevertheless, most work on this field (e.g., [14, 31, 36, 39, 40]) consider a more general setting where each Y_i follows a mixture of a normal distribution and some unknown distribution that stochastically dominates the normal distribution. Hence, the methods and results are not directly comparable to ours.

1.4. *Further description of our results.* We now discuss in more details our three contributions (known variance sparsity testing, unknown variance sparsity testing and sparsity estimation) mentioned in Section 1.2.

TABLE 1
Square minimax separation distances (*in the* $\asymp_\gamma$ *sense*) *when the noise level* σ *is known for all* $k_0 \in [0, n-1]$ *and* $\Delta \in [1, n-k_0]$

$\boldsymbol{k_0}$	$\boldsymbol{\Delta}$	$\boldsymbol{\rho_\gamma^{*2}[k_0, \Delta]/\sigma^2}$
$k_0 \leq \sqrt{n}$	$1 \leq \Delta \leq \sqrt{n}$	$\Delta \log(1 + \frac{\sqrt{n}}{\Delta})$
	$\sqrt{n} < \Delta \leq n - k_0$	$\sqrt{n}$
$k_0 > \sqrt{n}$	$1 \leq \Delta \leq \sqrt{n^{1/2}k_0}$	$\Delta \log(1 + \frac{k_0}{\Delta})$
	$\sqrt{n^{1/2}k_0} \leq \Delta \leq k_0$	$\Delta \frac{\log^2(1+\frac{k_0}{\Delta})}{\log(1+\frac{k_0}{\sqrt{n}})}$
	$k_0 \leq \Delta \leq n - k_0$	$\frac{k_0}{\log(1+\frac{k_0}{\sqrt{n}})}$

Sparsity testing for known σ. Table 1 summarizes the squared minimax separation distances $\rho_\gamma^{*2}[k_0, \Delta]$. Interestingly, for $k_0 \leq \sqrt{n}$, the minimax separation distance is the same as for signal detection ($k_0 = 0$). In contrast, for more complex null hypotheses ($k_0 \geq \sqrt{n}$), the complexity of the null hypothesis comes into play. For instance, when $\Delta \geq k_0 \geq \sqrt{n}$, then $\rho_\gamma^{*2}[k_0, \Delta]$ is of order $k_0/[\log(1 + \frac{k_0}{\sqrt{n}})]$. This is smaller by a polylog multiplicative term than what can be obtained by infimum tests and we have to rely on really different statistics. Our minimax adaptive procedure is a combination of three tests. The first one is an adaptation of the Higher Criticism test introduced in [17]. The second one relies on the empirical characteristic function of Y and borrows ideas from [26]. The third statistic is novel and relies on deconvolution ideas. As for the lower bounds of the minimax separation distances for large k_0, the proof ideas are more involved than for signal detection [2] and make use of the moment matching techniques introduced in [32] and later refined in [11, 28].

Sparsity testing for unknown σ. The results discussed above hold under the restrictive assumption that the noise level σ is known. For unknown σ, the situation is qualitatively different (see Table 2). As a first step, we study the signal detection problem ($k_0 = 0$) for which only partial results had been established. For sparse alternatives ($\Delta \leq \sqrt{n}$), one can plug an estimator of σ in the signal detection statistic so that the minimax separation distance $\rho_{\gamma,\text{var}}^*(0, \Delta)$ for unknown variance [defined in (30)] is the same as $\rho_\gamma^*(0, \Delta)$. However, for Δ larger than $\sqrt{n}$ and much smaller than n, one cannot simply plug a variance estimator and new test statistics are required. The squared separation distance $\rho_{\gamma,\text{var}}^{*2}(0, \Delta)$ is of order $\sqrt{\Delta n^{1/2}}$ whereas $\rho_\gamma^{*2}(0, \Delta)$ is only of order $\sqrt{n}$. In the really dense case where Δ is proportional to n, we establish that the separation distance $\rho_{\gamma,\text{var}}^{*2}(0, \Delta)$ is even larger. Turning to the general case $k_0 > 0$, we establish that $\rho_{\gamma,\text{var}}^*(k_0, \Delta)$ is larger than its counterpart for known σ for all $\Delta \geq \max(\sqrt{n}, k_0)$. In comparison to the

TABLE 2
Square minimax separation distance $\rho^{*2}_{\gamma,\mathrm{var}}[k_0,\Delta]$ [*as defined in equation* (30)] *when the noise level* σ *is unknown but belongs to some known fixed interval* $[\sigma_-,\sigma_+]$. *Here*, $c \in (0,1)$ *is some fixed universal constant and* $\xi \in (0,1)$ *can be chosen arbitrarily small*

$\boldsymbol{k_0}$	$\boldsymbol{\Delta}$	$\boldsymbol{\rho^{*2}_{\gamma,\mathrm{var}}[k_0,\Delta]/\sigma_+^2}$
$0 \le k_0 \le \sqrt{n}$	$1 \le \Delta \le \sqrt{n}$	$\Delta\log(1+\frac{\sqrt{n}}{\Delta})$
	$\sqrt{n} < \Delta \le cn$	$\sqrt{\Delta n^{1/2}}$
$\sqrt{n} \le k_0 \le n^{1-\xi}$	$1 \le \Delta \le \sqrt{k_0 n^{1/2}}$	$\Delta\log(1+\frac{k_0}{\Delta})$
	$\sqrt{k_0 n^{1/2}} < \Delta \le k_0$	$\Delta\frac{\log^2(1+\frac{k_0}{\Delta})}{\log(1+\frac{k_0}{\sqrt{n}})}$
	$k_0 < \Delta \le cn$	$\frac{\sqrt{\Delta k_0}}{\log(1+\frac{k_0}{\sqrt{n}})}$

known variance case, one cannot simply accommodate the adaptive test by estimating the noise level. In fact, the minimax adaptive test in this new setting is based on quite different statistics. Differences of minimax rates between known and unknown noise level have already been observed in other statistical problems such as signal detection [24, 42] and confidence intervals [6] in the high-dimensional linear model.

Sparsity estimation. Let us first verbalize the desirable properties of a good estimator of $\|\theta\|_0$. The functional $\|\theta\|_0$ is not continuous with respect to θ. Consider a one-sparse vector θ (with one large nonzero component) and a perturbation θ' of θ whose components are all nonzero but are arbitrarily small. As the distribution $\mathbb{P}_{\theta,\sigma}$ is close to $\mathbb{P}_{\theta',\sigma}$, the estimator $\widehat{k}$ will follow almost the same distribution for both parameters. It is obviously preferable for $\widehat{k}$ to be concentrated around one under $\mathbb{P}_{\theta',\sigma}$ than around n under $\mathbb{P}_{\theta,\sigma}$. In other words, a good estimator $\widehat{k}$ should have a small overestimation probability. Besides, a good estimator $\widehat{k}$ should be larger than any fixed q, as soon as the distance of θ to the collection $\mathbb{B}_0[q]$ is large enough.

To formalize the above intuition, let us consider the multiple testing problems with all hypotheses (H_q), for $q = 0,\dots,n$ where H_q is defined in (4). Then the set of true hypotheses is exactly $\{H_q, q \ge \|\theta\|_0\}$. Similarly, an estimator $\widehat{k}$ of $\|\theta\|_0$ can be interpreted as a multiple testing procedure rejecting all hypotheses H_q with $q < \widehat{k}$ and accepting all hypotheses H_q with $q \ge \widehat{k}$. Conversely, one can build an estimator of $\|\theta\|_0$ from any multiple testing procedure. Building on this correspondence between complexity tests and complexity estimation, we first construct a multiple sparsity testing procedure. Although the minimax optimality of multiple testing procedures is difficult to assess (but see [18]), we are able to prove that our procedure is simultaneously minimax for all single hypotheses H_q. Then the corresponding estimator $\widehat{k}$ satisfies, with high probability, the three following properties:

(a) $\widehat{k} \leq \|\theta\|_0$, which is equivalent to $\theta_{(\widehat{k})} \neq 0$ (here $\theta_{(i)}$ stands for the ith largest entry of θ in absolute value[2] with the convention $\theta_{(0)} = +\infty$).

(b) For all $q = 1, \dots, n - \widehat{k}$, $|\theta_{(\widehat{k}+q)}| \leq c\psi_{\widehat{k},q}$, where c is a numerical constant and the function $\psi_{\widehat{k},q}$ is defined in (24). In other words, we can certify, that even if $\widehat{k}$ is possibly smaller than $\|\theta\|_0$, each of its remaining $(\|\theta\|_0 - \widehat{k})$ nonzero components are small enough.

(c) $d_2(\theta, \mathbb{B}_0[\widehat{k}]) \leq c'\rho^*_\gamma[\widehat{k}, \|\theta\|_0 - \widehat{k}]$, where c' is a numerical constant and γ is fixed. In other words, θ is close in l_2 distance to the collection of $\widehat{k}$-sparse vectors.

Note that both properties (a) and (b) produce data-driven certificates for all $\theta_{(\widehat{k}+q)}$, $q \geq 0$ in the sense that corresponding bounds are explicit. Besides, the three above properties are valid for *all* $\theta \in \mathbb{R}^n$, whereas previous work [7, 8, 27] only considered specific classes θ by assuming for instance that the θ_i's are sampled according to a mixture of a Dirac at 0 and a smooth distribution. For a given θ, one can invert the inequalities in conditions (b) and (c) to obtain a bound for $|\widehat{k} - \|\theta\|_0|$. Finally, both conditions (b) and (c) are optimal from a minimax perspective defined in Section 4.

1.5. *Notation and organization of the paper.*

Although some of the notation have already been introduced, we gather them here to ease the reading. Given a vector $u \in \mathbb{R}^n$ and $p \geq 1$, we denote $\|u\|_p^p := (\sum_i |u_i|^p)^{1/p}$ its l_p norm. Also, $\|u\|_\infty := \max_i |u_i|$ stands for its l_∞ norm and $\|u\|_0 = \sum_i \mathbf{1}_{u_i \neq 0}$ its l_0 function. In the sequel, $\phi(\cdot)$ stands for the density of a standard normal variable, and $\Phi(\cdot)$ for its survival function. Also $\mathcal{N}(x, \sigma^2)$ stands for the normal distribution with mean x and variance σ^2. Given $x \in \mathbb{R}$, we write as usual $\lfloor x \rfloor$ for the integer part of x and $\lceil x \rceil$ for the rounding to the upper integer, and $(x)_+ := \max(x, 0)$. Given $x, y \in \mathbb{R}$, $x \vee y$ (resp., $x \wedge y$) corresponds to the maximum (resp., minimum) of x and y. Also $[n]$ is short for the set $\{1, \dots, n\}$. For any $i \in [n]$, $\theta_{(i)}$ stands for the ith largest entry of θ in absolute value. In other words, one has $|\theta_{(1)}| \geq |\theta_{(2)}| \geq \cdots \geq |\theta_{(n)}|$.

In the sequel, c, c_1, ... denote positive universal constants that may change from line to line. We also denote c_α, c'_β, ..., positive constants whose values may depend on α or β.

When Y is distributed according to the model (1), we write $\mathbb{P}_{\theta,\sigma}$ for the distribution of Y. As σ is fixed and supposed to be known in Sections 2 and 4, we drop the dependency on σ in these two sections and simply write $\mathbb{P}_\theta$.

In Section 2, we describe our model testing results when the variance of the noise is known, presenting both upper and lower bounds. Section 3 is devoted to the unknown variance case. In Section 4, we detail how these testing results can be applied to the relevant problem of sparsity estimation. Finally, remaining results and all the proofs are postponed the Supplementary Material [13].

[2] Consequently, we have $|\theta_{(1)}| \geq |\theta_{(2)}| \geq \cdots \geq |\theta_{(n)}|$.

2. Sparsity testing with known variance.

2.1. *Minimax lower bound.* In this section, we consider the sparsity testing problem (4) in a setting when the noise variance σ^2 is known. The following theorem states a lower bound on the minimax separation distance $\rho^*_\gamma[k_0, \Delta]$.

THEOREM 1. *There exists a numerical constant $c > 0$ such that the following holds. Consider any $\gamma \leq 0.5$. For any $k_0 \leq \sqrt{n}$ and $\Delta \leq n - k_0$, we have*

$$\rho^{*2}_\gamma[k_0, \Delta] \geq \sigma^2 \Delta \log\left[1 + \frac{\sqrt{n}}{8\Delta}\right]. \tag{8}$$

For any $k_0 > \sqrt{n}$, we have

$$\rho^{*2}_\gamma[k_0, \Delta] \geq c\sigma^2 \begin{cases} \Delta\left[\dfrac{\log^2[1+\frac{k_0}{\Delta}]}{\log[1+\frac{k_0}{\sqrt{n}}]} \wedge \log\left[1+\dfrac{k_0}{\Delta}\right]\right] \\ \qquad \textit{if } \Delta \leq k_0 \wedge (n-k_0), \\ \dfrac{k_0}{\log[1+\frac{k_0}{\sqrt{n}}]} \\ \qquad \textit{if } k_0 < \Delta \leq n - k_0. \end{cases} \tag{9}$$

As proved in the next subsection, this lower bound turns out to be sharp. We shall precisely discuss these quantities later. Before this, we only give a glimpse of the different regimes unveiled by the above theorem.

Whenever $k_0 \leq \sqrt{n}$, the lower bound on the minimax separation distance is the same as the signal detection minimax separation distance $\rho^*_\gamma[0, \Delta]$; see [2, 15]. In this regime, the size k_0 of the null hypothesis does not play a role in the separation distance. In fact, the proof of (8) is a consequence of known results for the signal detection problem. More precisely, we follow Le Cam's method and choose a particular $\theta_0 \in \mathbb{B}_0[k_0]$ and a prior distribution ν on the collection $\mathbb{B}_0[k_0 + \Delta, k_0, \rho]$. Let us write $\mathbb{Q}_1 := \int \mathbb{P}_\theta \nu(d\theta)$ the marginal distribution of Y when θ is sampled according to ν. Then the risk $R(T; k_0, \Delta, \rho)$ (5) of any test T is larger than $1 - \|\mathbb{P}_{\theta_0} - \mathbb{Q}_1\|_{\mathrm{TV}}$ ($\|\cdot\|_{\mathrm{TV}}$ is the total variation distance) and we bound the total variation distance by the χ^2 distance (see the proof for more details).

For k_0 much larger than $\sqrt{n}$ and for $\Delta \geq k_0$, the lower bound (9) is of order $k_0/\log[\frac{k_0}{\sqrt{n}}]$—which is significantly larger than the signal detection rate $\rho^*_\gamma[0, \Delta]$. In this regime, the complexity of the null hypothesis H_{k_0} has to be taken into account to obtain the right lower bound. Following an approach pioneered in [32], we build two product prior distributions $\mu_0^{\otimes n}$ and $\mu_1^{\otimes n}$ (almost) supported by $\mathbb{B}_0[k_0]$ and $\mathbb{B}_0[k_0 + \Delta, k_0, \rho]$ in such a way that the first moments of μ_0 and μ_1 are matching. Writing $\mathbb{Q}_0 := \int \mathbb{P}_\theta \mu_0^{\otimes n}(d\theta)$ and $\mathbb{Q}_1 := \int \mathbb{P}_\theta \mu_1^{\otimes n}(d\theta)$, we need to upper bound the χ^2 distance between $\mathbb{Q}_0$ and $\mathbb{Q}_1$. It turns out that matching the moments of μ_0 and μ_1 enforces the χ^2 distribution between $\mathbb{Q}_0$ and $\mathbb{Q}_1$ to be small enough. The main technical hurdle in the proof is the construction of the two measures μ_0 and μ_1 that maximize the number of matching moments, while being supported respectively on the null and alternative hypothesis with ρ as large as possible.

Finally, consider the first equation in (9) that corresponds to the regime $k_0 \geq \sqrt{n}$ and $\Delta \leq k_0$. When $\Delta \leq \sqrt{n^{1/2}k_0}$ (sparse alternative), the lower bound is of the order of $\Delta \log[1 + k_0/\Delta]$. For $\sqrt{n^{1/2}k_0} \leq \Delta \leq k_0$, the lower bound (9) is of order $\Delta \log^2[k_0/\Delta]/\log[1 + k_0/\sqrt{n}]$. In this intermediary regime, the logarithmic terms vary smoothly from the sparse regime ($\log[1 + k_0/\Delta]$) to the dense regime ($\log^{-1}[1 + k_0/\sqrt{n}]$).

2.2. *Minimax upper bound.* In this subsection, we construct three tests that are most effective in three different situations: the Higher Criticism regime (large but few nonzero components), the bulk regime (many but small nonzero components) and the intermediary regime. Then a combination of these three procedures is proved to achieve the minimax lower bounds of Theorem 1 and is even adaptive to the sparsity Δ. Each of the statistics introduced in this section are of the form $S = \sum_{i=1}^n h(Y_i)$ for some function h. Denoting g the function such that $\mathbb{E}_\theta[h(Y_i)] = g(\theta_i)$, we shall choose functions h in such a way that g approximates well the indicator function $1_{x\neq 0}$, so that $\mathbb{E}_\theta[S]$ approximates $\|\theta\|_0$. All the statistics considered in this subsection correspond to symmetric functions g such that $g(0) = 0$ and $g(x)$ converge to 1 for $x \to \infty$. However, these statistics differ in the way g approximates the indicator function in the vicinity of 0 (this corresponds to a bias term) and in the size of their variance. As explained below, a large bias and a small variance are tailored to detect a few large coefficients (Higher Criticism statistic) whereas a small bias and a higher variance are mostly suited to detect many small coefficients (bulk statistic). Throughout this subsection, we consider some fixed α and β in $(0, 1)$.

2.2.1. *Higher criticism statistic.* Let us adapt the Higher Criticism statistic introduced in [17] for signal detection. Recall that, for $t > 0$, $\Phi(t)$ is the survival function of the standard normal distribution For any $t > 0$, define

$$N_t := \#\{i, |Y_i| \geq t\}, \tag{10}$$

the number of components larger (in absolute value) than t, $t_{*,\alpha}^{\mathrm{HC}} := \lceil\sqrt{2\log[4n/\alpha]}\rceil$ and the collection $\mathcal{T}_\alpha := \{1, \dots, t_{*,\alpha}^{\mathrm{HC}}\}$. Then the test $T_{\alpha,k_0}^{\mathrm{HC}}$ rejects the null hypothesis H_{k_0}, if either $N_{\sigma t_{*,\alpha}^{\mathrm{HC}}} \geq k_0 + 1$ or for some $t \in \mathcal{T}_\alpha$,

$$N_{\sigma t} \geq k_0 + 2(n - k_0)\Phi(t) + u_{t,\alpha}^{\mathrm{HC}}, \tag{11}$$

where

$$u_{t,\alpha}^{\mathrm{HC}} := 2\sqrt{n\Phi(t)\log\left(\frac{t^2\pi^2}{3\alpha}\right)} + \frac{2}{3}\log\left(\frac{t^2\pi^2}{3\alpha}\right). \tag{12}$$

Under the null hypothesis H_{k_0}, θ contains at most k_0 nonzero coefficients and $N_{\sigma t} - k_0$ is therefore stochastically dominated by a Binomial random variable with parameters $(n - k_0, 2\Phi(t))$. It then follows from Chebychev inequality that $N_{\sigma t} \leq k_0 + 2(n - k_0)\Phi(t) + O_p(\sqrt{(n - k_0)\Phi(t)})$. Note that, for large t, $N_{\sigma t}$ has a smaller variance but it does manage to select coordinates θ_i that are close to zero. For small t, the variance is higher but smaller nonzero components at taken into account. This is why the test $T^{\mathrm{HC}}_{\alpha,k_0}$ is an aggregation of such statistics for a wide scope of values of t. The specific choice of the term $u^{\mathrm{HC}}_{t,\alpha}$ allows to handle the multiplicity of the tests. For $k_0 = 0$ (signal detection), $T^{\mathrm{HC}}_{\alpha,k_0}$ is somewhat analogous to the vanilla Higher Criticism test [17], but there are some differences that we discuss below.

PROPOSITION 1. *There exists a positive constant $c_{\alpha,\beta}$ such that the following holds. The size of the test $T^{\mathrm{HC}}_{\alpha,k_0}$ does not exceed α. Besides, any $\theta \in \mathbb{R}^n$ such that*

$$|\theta_{(k_0+q)}| \geq c_{\alpha,\beta}\sigma\sqrt{\log\left(2 + \frac{\sqrt{n} \vee k_0}{q}\right)} \qquad \text{for some } q \in [1, n - k_0] \tag{13}$$

belongs to the high probability rejection region of $T^{\mathrm{HC}}_{\alpha,k_0}$, that is $\mathbb{P}_\theta[T^{\mathrm{HC}}_{\alpha,k_0} = 1] \geq 1 - \beta$.

In the specific case $k_0 = 0$, we recover the known behavior or the Higher Criticism statistic in the signal detection setting. The test $T^{\mathrm{HC}}_{\alpha,k_0}$ is powerful when, for a given integer q, there are least $(k_0 + q)$ coefficients larger than some threshold depending on q. For $q = 1$, the threshold is of order $\sigma\sqrt{\log(n)}$, whereas for $q \geq \sqrt{n} \vee k_0$, the threshold is of order one. More generally, Proposition 1 provides an upper bound matching to the minimax lower bound in Theorem 1 whenever $\Delta \leq \sqrt{n}$ for $k_0 \leq \sqrt{n}$ and $\Delta \leq \sqrt{(k_0 n^{1/2})}$ for $k_0 \geq \sqrt{n}$.

REMARK. In signal detection, alternatives to the Higher Criticism have been considered. In particular, Li and Siegmund [33] and Moscovich et al. [37] advocate for the use of Berk–Jones-type procedures. It has been numerically demonstrated that these methods exhibit slightly better finite sample performances than the Higher Criticism while achieving the same asymptotic power as the Higher Criticism [17]. In our context, the original Higher Criticism would amount to considering a supremum of normalized statistics of the form

$$\sup_{t \in \mathcal{T}_\alpha} \frac{N_{\sigma t} - k_0 - 2(n - k_0)\Phi(t)}{\sqrt{(n - k_0)2\Phi(t)(1 - 2\Phi(t))}},$$

whereas the exact Berk–Jones [37] would correspond to a minimum of p-values for each individual tests

$$\inf_{t \in \mathcal{T}_\alpha} g_{n-k_0,2\Phi(t)}\big[(N_{\sigma t} - k_0)_+\big], \tag{14}$$

where $g_{n,p}(k)$ stands the probability a Binomial random variable with parameters (n, p) is larger than or equal to k. In (11), we use an intermediary approach between computing the exact p-values (Berk–Jones) and an upper bound based on empirical normalization (Higher Criticism) by taking the Bernstein's upper bound of the quantiles. Instead of $T^{\mathrm{HC}}_{\alpha,k_0}$, we could have studied a statistic based on (14). We did not pursue this strategy, as the analysis seems to be more technical and $T^{\mathrm{HC}}_{\alpha,k_0}$ is already minimax optimal for sparse alternatives.

2.2.2. *Detecting the signal in the bulk distribution.* When there are many small coefficients, we rely on the empirical characteristic functions of Y following an approach introduced in [26]. Given $s > 0$, define the function

$$\kappa_s(x) := \int_{-1}^{1} (1 - |\xi|) e^{s^2\xi^2/2} \cos(s\xi x)\, d\xi, \tag{15}$$

and the test statistic $Z(s)$

$$Z(s) := \sum_{i=1}^{n} \bigl(1 - \kappa_s(Y_i/\sigma)\bigr). \tag{16}$$

Let us describe the intuition behind this statistic using a population approach. Denoting $\overline{\varphi}_n(s)$ the empirical characteristic function and $\overline{\varphi}(s)$ its expectation

$$\overline{\varphi}_n(s) := n^{-1} \sum_{i=1}^{n} \cos(sY_i), \qquad \overline{\varphi}(s) := n^{-1} \sum_{i\le n} \cos(s\theta_i) e^{-\frac{s^2\sigma^2}{2}}, \tag{17}$$

one can derive the expectation of $Z(s)$

$$\mathbb{E}_\theta\bigl[Z(s)\bigr] = \sum_{i=1}^{n} \Bigl[1 - \int_{-1}^{1} (1 - |\xi|) \cos(s\xi\theta_i/\sigma)\, d\xi\Bigr] = \sum_{i=1}^{n} \Bigl[1 - 2\frac{1 - \cos(s\theta_i/\sigma)}{(s\theta_i/\sigma)^2}\Bigr],$$

with the convention $(1 - \cos(x))/x^2 = 1/2$ for $x = 0$. Denoting $g(x) = 1 - 2(1 - \cos(x))/x^2$, one may easily show that $g(x) \in [0, 1]$, $g(x) \ge 1 - c/x^2$ for large x and $g(x) = x^2/12 + o(x^2)$ around 0. As a consequence, $\mathbb{E}_\theta(Z(s)) = \sum_{i=1}^{n} g(s\theta_i/\sigma)$ approximates $\|\theta\|_0$ and is able to partially take into account even small values of θ_i. If θ contains many coefficients that are large in front of σ/s or if there are so many small coefficients $|\theta_i|$ that the corresponding sum $\sum_i \theta_i^2 s^2/\sigma^2$ is large in front of k_0, then at least in expectation, $Z(s)$ is larger than under the null. Proposition 2 below makes this informal argument rigorous.

REMARK. Rewriting the statistic $Z(s)/n = 1 - \int_{-1}^{1}(1 - |\xi|)e^{s^2\xi^2/2}\overline{\varphi}_n(s\xi/\sigma)\, d\xi$, one observes that the empirical characteristic function is multiplied by the function $(1 - |\xi|)$ before integration. In [26], Jin also suggests other statistics such as $\int_{-1}^{1} e^{s^2\xi^2/2}\overline{\varphi}_n(s\xi/\sigma)\, d\xi$ or the deconvolution estimator $e^{s^2/2}\overline{\varphi}_n(s/\sigma)$. However, these two statistics turn out to be suboptimal in our setting.

In practice, we set $s_{k_0} := \sqrt{\log(ek_0^2/n)} \vee 1$ and we define the test T^B_{α,k_0} rejecting the null hypothesis when

$$Z(s_{k_0}) \geq k_0 + u^B_{k_0,\alpha} \qquad \text{where } u^B_{k_0,\alpha} := \frac{e^{s_{k_0}^2/2}}{s_{k_0}}\sqrt{8n\log(2/\alpha)}. \tag{18}$$

PROPOSITION 2. *There exist three positive constants $c_{\alpha,\beta}$, $c'_{\alpha,\beta}$, $c''_{\alpha,\beta}$ such that the following holds. The type I error probability of T^B_{α,k_0} does not exceed α. Besides, any $\theta \in \mathbb{R}^n$ satisfying any of the two following conditions*:

$$|\theta_{(k_0+q)}| \geq c_{\alpha,\beta}\sigma\sqrt{\frac{k_0}{q\log(1+\frac{k_0}{\sqrt{n}})}} \qquad \textit{for some } q \geq \frac{c'_{\alpha,\beta}k_0}{\sqrt{\log(1+\frac{k_0^2}{n})}}, \tag{19}$$

$$\sum_{i=1}^n [\theta_i^2 \wedge s_{k_0}^{-2}] \geq c''_{\alpha,\beta}\sigma^2\frac{k_0}{\log(1+k_0/\sqrt{n})}, \tag{20}$$

belongs to the high probability rejection region of T^B_{α,k_0}, that is, $\mathbb{P}_\theta[T^B_{\alpha,k_0} = 1] \geq 1-\beta$.

The above proposition provides two sufficient conditions for T^B_{α,k_0} to be powerful. The second condition (20) formalizes the above discussion for the population version of the statistic: when the squared l_2 norm of the restriction of θ to its small coefficients is larger than $\sigma\frac{k_0}{\log(1+k_0/\sqrt{n})}$, then the test is powerful. Condition (19) ensures that the test is also powerful when there are more than k_0+q coefficients larger than some threshold depending on q. In comparison to the Higher Criticism test, Condition (19) is effective for large q (many nonzero coefficients), but these coefficients can be much smaller than one.

As justified in the proof of Corollary 1, the test T^B_{α,k_0} together with the previous test $T^{\mathrm{HC}}_{\alpha,k_0}$ match the minimax lower bound whenever $\Delta \geq k_0 \vee \sqrt{n}$.

Practical implementation. In order to compute $Z(s)$, one can approximate the integral by its Riemann sum. Given some large $M > 0$ large, we approximate $\int_{-1}^1 (1-|\xi|)\exp(s^2\xi^2/2)\cos(s\xi x)\,d\xi$, by

$$\frac{2}{M}\sum_{i=1}^M\left(1-\frac{i}{M}\right)\exp\left[\left(\frac{i}{M}\right)^2\frac{s^2}{2}\right]\cos\left(s\frac{i}{M}\right)d\xi.$$

As the function in the integral has higher variations near $|\xi| = 1$, one could also take finer steps in vicinity of 1.

2.2.3. *Intermediary regimes*. A combination of the two previous tests covers the extreme regimes for the sparsity testing problem: a few large coefficients (Higher Criticism) and many small coefficients (bulk). Unfortunately, they turn out to be suboptimal in intermediate regimes, that is, for any parameters in between. This is why we have to devise a third test. In this subsection, we aim at discovering intermediary signals whose signature is neither in the bulk of the empirical distribution of (Y_i) nor in its extreme values. This strategy is relevant for large k_0 only and we assume henceforth that $k_0 \geq 20\sqrt{n}$.

Given two tuning parameters r and l, define the function

$$\eta_{r,w}(x) := \frac{r}{(1-2\Phi(r))} \int_{-1}^{1} \frac{e^{-r^2\xi^2/2}}{\sqrt{2\pi}} e^{\xi^2 w^2/2} \cos(\xi w x)\, d\xi \tag{21}$$

and the statistic

$$V(r,w) := \sum_{i=1}^{n} [1 - \eta_{r,w}(Y_i/\sigma)].$$

In order to get a grasp this statistic, let us consider the expectation of $\eta_{r,w}(X)$ for $X \sim \mathcal{N}(x,1)$. Simple computations [see (41) in the proof of Proposition 3] lead to

$$\mathbb{E}[1-\eta_{r,w}(X)] = 1 - \frac{1}{1-2\Phi(r)} \int_{-r}^{r} \phi(\xi) \cos\left(\xi x \frac{w}{r}\right) d\xi,$$

which, for large r, is of order $1 - \int_{\mathbb{R}} \phi(\xi) \cos(\xi x \frac{w}{r})\, d\xi = 1 - \exp(-x^2 \frac{w^2}{2r^2})$. As a consequence, $\mathbb{E}_\theta[V(r,w)]$ approximates the function $\|\theta\|_0$ at an exponential rate. In contrast, the population version of the statistic $Z(s)$ [defined in (16)] only approximates the function $\|\theta\|_0$ at a quadratic rate. Therefore, the statistic $V(r,w)$ better handles coefficients θ_i that are large compared to $\sigma r/w$ than the statistic $Z(s)$. Unfortunately, the variance $V(r,w)$ is quite large which prevents us from taking w/r as large as the tuning parameter s_{k_0} in the previous test. This is why this statistic is tailored to intermediary values of $|\theta_i|$. As the optimal choice of w/r depends on θ, we consider below a collection of such statistics and use an aggregated test.

The test T^I_{α,k_0} is an aggregation of multiple tests based on the statistics $V(r,w)$ for different tuning parameters r and w. Define $l_{k_0} := \lceil (k_0\sqrt{n})^{1/2} \rceil$ and the dyadic collection $\mathcal{L}_{k_0} = \{l_{k_0}, 2l_{k_0}, 4l_{k_0}, \dots, l_{\max}\}$ where $l_{\max} := 2^{\lfloor \log_2(k_0/l_{k_0}) \rfloor} l_{k_0}/4 \leq k_0/4$ where $\log_2$ is the binary logarithm. Note that $\mathcal{L}_{k_0}$ is not empty if $k_0 \geq 20\sqrt{n}$ and n is large enough. Given any $l \in \mathcal{L}_{k_0}$, define

$$r_{k_0,l} := \sqrt{2\log\left(\frac{k_0}{l}\right)}, \qquad w_l := \sqrt{\log\left(\frac{l}{\sqrt{n}}\right)}. \tag{22}$$

Then the test T^I_{α,k_0} rejects the null hypothesis if, for some $l \in \mathcal{L}_{k_0}$,

$$\begin{aligned} &V(r_{k_0,l}, w_l) \geq k_0 + l + u^I_{k_0,l,\alpha} \\ &\qquad \text{where } u^I_{k_0,l,\alpha} := \sqrt{2ln^{1/2} \log\left(\frac{\pi^2[1+\log_2(l/l_{k_0})]^2}{6\alpha}\right)}. \end{aligned} \tag{23}$$

PROPOSITION 3. *There exists four positive constants $c, c_{\alpha,\beta}, c'_{\alpha,\beta}, c''_{\alpha,\beta}$ such that the following holds. Assume that $k_0 \geq 20\sqrt{n}$ and $n \geq c$. The type I error probability of T^I_{α,k_0} does not exceed α. If $k_0 \geq c_{\alpha,\beta}\sqrt{n}$, any $\theta \in \mathbb{R}^n$ satisfying*

$$|\theta_{(k_0+q)}| \geq c'_{\alpha,\beta}\sigma \frac{1+\log(1+\frac{k_0}{q\wedge k_0})}{\sqrt{\log(1+\frac{k_0}{\sqrt{n}})}} \qquad \textit{for some } q \geq c''_{\alpha,\beta}\sqrt{k_0 n^{1/2}},$$

belongs to the high probability rejection region of T^I_{α,k_0}, that is, $\mathbb{P}_\theta[T^I_{\alpha,k_0} = 1] \geq 1-\beta$.

In view of the minimax lower bound in Theorem 1, Proposition 3 turns out to be mostly relevant for $k_0 \geq \sqrt{n}$ and $\sqrt{n^{1/2}k_0} \leq \Delta \leq k_0$ [the logarithmic terms in Proposition 3 match those in (9)]. As justified in the next subsection, T^I_{α,k_0} together with $T^{\mathrm{HC}}_{\alpha,k_0}$ achieves a minimax separation distance in this regime.

2.3. *Combination of the tests.* For any integer $q \in [n-k_0]$, define $\psi_{k_0,q} > 0$ by

$$\psi^2_{k_0,q} := \begin{cases} \log\left[1+\dfrac{\sqrt{n}}{q}\right] & \text{if } k_0 \leq \sqrt{n}, \\ \dfrac{\log^2(1+\frac{k_0}{q})}{\log(1+\frac{k_0}{\sqrt{n}})} \wedge \log\left(1+\dfrac{k_0}{q}\right) & \text{if } k_0 > \sqrt{n} \text{ and } q \leq k_0, \\ \dfrac{k_0}{q\log(1+\frac{k_0}{\sqrt{n}})} & \text{if } k_0 > \sqrt{n} \text{ and } q > k_0. \end{cases} \tag{24}$$

Let T^C_{α,k_0} denote the aggregation of the three previous tests. We take $T^C_{\alpha,k_0} := \max(T^{\mathrm{HC}}_{\alpha/3,k_0}, T^B_{\alpha/3,k_0}, T^I_{\alpha/3,k_0})$, if $k_0 \geq 20\sqrt{n}$ and $T^C_{\alpha,k_0} := \max(T^{\mathrm{HC}}_{\alpha/2,k_0}, T^B_{\alpha/2,k_0})$ else. The following result holds.

COROLLARY 1. *There exist three constants c, $c_{\alpha,\beta}$ and $c'_{\alpha,\beta}$ such that the following holds for $n \geq c$. The type I error probability of T^C_{α,k_0} does not exceed α. Besides, $\mathbb{P}_\theta[T^C_{\alpha,k_0} = 1] \geq 1-\beta$ for any vector θ such that*

$$|\theta_{(k_0+q)}| \geq c_{\alpha,\beta}\sigma\psi_{k_0,q} \qquad \textit{for some } q \in [n-k_0]. \tag{25}$$

Also, $\mathbb{P}_\theta[T^C_{\alpha,k_0} = 1] \geq 1-\beta$ for any vector θ satisfying,

$$\begin{aligned} &\theta \in \mathbb{B}_0(k_0+\Delta) \quad \textit{and} \\ &d^2[\theta, \mathbb{B}_0(k_0)] \geq c'_{\alpha,\beta}\sigma^2\Delta\psi^2_{k_0,\Delta} \qquad \textit{for some } \Delta \in [n-k_0]. \end{aligned} \tag{26}$$

In view of Theorem 1 and (26) in Corollary 1, it holds that $\rho^*_{\alpha+\beta}[k_0,\Delta] \asymp_\gamma \sigma^2\Delta\psi^2_{k_0,\delta}$. Besides, the test T^C_{α,k_0} simultaneously achieves (up to multiplicative constants) these minimax separation distances over all $\Delta \in [n-k_0]$. Condition (25) provides a complementary characterization of T^C_{α,k_0} power function. This bound will be central for sparsity estimation in the next section.

To conclude this section, we summarize the results on the testing separation distance $\rho^{*2}_\gamma[k_0,\Delta]$ as depicted in Table 1 in the Introduction. For $k_0 \leq \sqrt{n}$, $\rho^*_\gamma[k_0,\Delta]$ is of same order as the signal detection separation distance $\rho^*_\gamma[0,\Delta]$. For $k_0 > \sqrt{n}$, the minimax-optimal separation distance $\rho^*_\gamma[k_0,\Delta]$ becomes significantly larger than the signal detection separation distance. The complexity of the null hypothesis plays an important role in $\rho^*_\gamma[k_0,\Delta]$. For instance, when $k_0 = n^\zeta$ with $\zeta > 1/2$ and for $\Delta \geq k_0$, $\rho^{*2}_\gamma[k_0,\Delta]$ is of order $k_0/\log(n)$. Besides, for Δ between $\sqrt{n^{1/2}k_0}$ and k_0, there is smooth transition from squared separation distances of order $\Delta\log(n)$ to $\Delta/\log(n)$.

3. Sparsity testing with unknown variance. In this part, we consider the problem of testing the sparsity of θ when the noise level σ is unknown. For the sake of simplicity, it is assumed that σ belongs to some fixed interval $[\sigma_-,\sigma_+]$ where $0 < \sigma_- < \sigma_+$ are known. This assumption is not restrictive since, in most interesting settings, one may build a data-driven interval $[\widehat{\sigma}_-,\widehat{\sigma}_+]$ containing σ with large probability and such that the ratio $\widehat{\sigma}_+/\widehat{\sigma}_-$ remains bounded. See Section A for further explanations.

In this section and in the corresponding proofs, we denote $\mathbb{P}_{\theta,\sigma}$ the distribution of Y. Given two integers $k_0 \geq 0$ and $\Delta > 0$, we consider the sparsity testing problem with unknown variance

$$\begin{aligned} &H_{k_0,\mathrm{var}}: \quad \theta \in \mathbb{B}_0[k_0], \qquad \sigma \in [\sigma_-,\sigma_+] \quad \text{versus} \\ &H_{\Delta,k_0,\rho,\mathrm{var}}: \quad \theta \in \mathbb{B}_0[k_0+\Delta,k_0,\rho], \qquad \sigma \in [\sigma_-,\sigma_+]. \end{aligned} \tag{27}$$

Given a test T, let us define its risk $R_{\mathrm{var}}(T;k_0,\Delta,\rho)$ for the problem (27) by

$$\begin{aligned} R_{\mathrm{var}}(T;k_0,\Delta,\rho) := &\sup_{\theta\in\mathbb{B}_0[k_0],\sigma\in[\sigma_-,\sigma_+]} \mathbb{P}_{\theta,\sigma}[T=1] \\ &+ \sup_{\theta\in\mathbb{B}_0[k_0+\Delta,k_0,\rho],\sigma\in[\sigma_-,\sigma_+]} \mathbb{P}_{\theta,\sigma}[T=0], \end{aligned} \tag{28}$$

and its γ-separation distance $\rho_{\gamma,\mathrm{var}}(T)$ by

$$\rho_{\gamma,\mathrm{var}}(T;k_0,\Delta) := \sup\{\rho > 0 : R_{\mathrm{var}}(T;k_0,\Delta,\rho) > \gamma\}. \tag{29}$$

Finally, the minimax separation distance for the problem with unknown variance is defined by

$$\rho^*_{\gamma,\mathrm{var}}[k_0,\Delta] := \inf_T \rho_{\gamma,\mathrm{var}}(T;k_0,\Delta). \tag{30}$$

3.1. *Detection problem* ($k_0 = 0$). Before turning to the general case, let us first restrict ourselves to the signal detection problem. To the best of our knowledge, the minimax separation distances for unknown variance have not been derived yet. Besides, this provides an introduction to the general case. Obviously, the problem with unknown variance is at least as difficult as the initial problem (4) so that, for all Δ, $\rho^*_{\gamma,\mathrm{var}}[k_0, \Delta] \geq \rho^*_\gamma[k_0, \Delta]$ (where $\rho^{*2}_\gamma[k_0, \Delta]$ is defined for known $\sigma = \sigma_+$). Our purpose is to pinpoint the range of Δ such that $\rho^*_{\gamma,\mathrm{var}}[k_0, \Delta]$ is of order $\rho^*_\gamma[k_0, \Delta]$ so that the knowledge of the variance is not critical and the range of Δ such that $\rho^*_{\gamma,\mathrm{var}}[k_0, \Delta]$ is much larger than $\rho^*_\gamma[k_0, \Delta]$ so that the knowledge of the variance effectively makes the testing problem easier.

PROPOSITION 4. *Fix any $\gamma < 0.25$. There exist two positive constants c_γ and c'_γ such that the following holds For any $\Delta \leq \sqrt{n}$, we have*

$$c_\gamma \sigma_+^2 \Delta \log\left(1 + \frac{\sqrt{n}}{\Delta}\right) \leq \rho^{*2}_{\gamma,\mathrm{var}}[0, \Delta] \leq c'_\gamma \sigma_+^2 \Delta \log\left(1 + \frac{\sqrt{n}}{\Delta}\right). \tag{31}$$

For any $\eta < 1/3$ and any $\Delta \in [\sqrt{n}, (\frac{1}{3} - \eta)n]$,

$$c_\gamma \sigma_+^2 \sqrt{\Delta n^{1/2}} \leq \rho^{*2}_{\gamma,\mathrm{var}}[0, \Delta] \leq c'_{\gamma,\eta} \sigma_+^2 \sqrt{\Delta n^{1/2}}, \tag{32}$$

where the constant $c_{\gamma,\eta}$ and $c'_{\gamma,\eta}$ only depend on γ and η.

For $\Delta \leq \sqrt{n}$, the minimax separation distance is the same as for known variance. This can be achieved, for instance, by a generalization of the Higher Criticism to the unknown variance setting as explained in Section 3.3.

For Δ between $\sqrt{n}$ and $n/3$, $\rho^{*2}_{\gamma,\mathrm{var}}[0, \Delta]$ is of order $\sqrt{\Delta n^{1/2}}$ which is much larger than the squared separation distance $\sqrt{n}$ for known variance. When σ is known, a near optimal test amounts to rejecting the null hypothesis when $S_2 = \|Y\|_2^2/\sigma^2 - n$ is large compared $\sqrt{n}$. Under the null, $S_2 + n$ follows a χ^2 distribution with n degrees of freedom whereas, under the alternative, $S_2 + n$ follows a noncentral χ^2 distribution with noncentrality parameter $\|\theta\|_2^2/\sigma^2$ so that the test is powerful when $\|\theta\|_2^2$ is large compared to $\sigma^2\sqrt{n}$. When σ is unknown, one cannot simply rely on the second moment of Y and higher order moments are needed. For instance, a test achieving the separation distance (32) is based on the statistic

$$S_4 = \frac{n\|Y\|_4^4}{\|Y\|_2^4} - 3. \tag{33}$$

Under the null, it follows from Chebychev inequality that $S_4 = O_P(n^{-1/2})$. Under the alternative, $\mathbb{E}_{\theta,\sigma}[\|Y\|_2^2] = \|\theta\|_2^2 + n\sigma^2$ and $\mathbb{E}_{\theta,\sigma}[\|Y\|_4^4] = \|\theta\|_4^4 + 6\sigma^2\|\theta\|_2^2 + 3n\sigma^2$ so that, one may expect that S_4 is of order

$$\frac{n\|\theta\|_4^4 - 3\|\theta\|_2^4}{(\|\theta\|_2^2 + n\sigma^2)^2} \geq (n - 3\|\theta\|_0)\frac{\|\theta\|_4^4}{(\|\theta\|_2^2 + n\sigma^2)^2},$$

by Cauchy–Schwarz inequality. As a consequence, S_4 takes significantly larger values when $(n - 3\|\theta\|_0)\|\theta\|_4^4$ is large compared to $n^{3/2}$. When $n - 3\Delta$ is of order n, this occurs when $\|\theta\|_2^2$ is larger than $\sqrt{\Delta n^{1/2}}$. See the proof of Proposition 4 for further details.

Conversely, the proof of the minimax lower bound (32) also proceeds from moments arguments. For known variance $\sigma = 1$, one builds a prior probability measure ν on θ supported by $\mathbb{B}_0[\Delta]$ such that the expectation of $\sum_{i=1}^n Y_i$ is the same under $\int \mathbb{P}_{\theta,\sigma}\nu(d\theta)$ and $\mathbb{P}_{0,\sigma}$. When the variance is unknown, one may choose the prior ν and $\sigma_1 \neq \sigma_0$ such that all expectations of $\sum_{i=1}^n Y_i^q$ for $q = 1, 2, 3$ are matching under $\int \mathbb{P}_{\theta,\sigma_1}\nu(d\theta)$ and $\mathbb{P}_{0,\sigma_0}$. As explained in the proof of Theorem 2, these moment matching properties translate into a smaller total variation between $\int \mathbb{P}_{\theta,\sigma_1}\nu(d\theta)$ and $\mathbb{P}_{0,\sigma_0}$ which in turn implies that the separation distance $\rho^*_{\gamma,\mathrm{var}}[0, \Delta]$ is large.

Proposition 4 above characterizes the signal detection separation distance for all Δ small compared to $n/3$. For $\Delta = cn$ with $c < 1/3$, $\rho^{*2}_{\gamma,\mathrm{var}}[0, \Delta]$ is of order $n^{3/4}$. One may then wonder if $\rho^{*2}_{\gamma,\mathrm{var}}[0, \Delta]$ remains of order $n^{3/4}$ for all $\Delta \in (n/3, n]$. This turns out to be false. In fact, $\rho^{*2}_{\gamma,\mathrm{var}}[0, n]$ is of order $(\sigma_+^2 - \sigma_-^2)n$. Indeed, let ν denote the centered normal distribution with variance $(\sigma_+^2 - \sigma_-^2)I_n$. When θ is sampled according to ν and for $\sigma = \sigma_-$, the marginal distribution of Y is $\mathbb{P}_{0,\sigma^+}$. As a consequence, it is impossible to distinguish $\theta = 0$ from $\theta \sim \nu$ for which $\|\theta\|_2^2$ is of order $(\sigma_+^2 - \sigma_-^2)n$. This entails that $\rho^{*2}_{\gamma,\mathrm{var}}[0, n]$ is at least of order $(\sigma_+^2 - \sigma_-^2)n$.

In fact, the squared minimax separation distance $\rho^{*2}_{\gamma,\mathrm{var}}[0, \Delta]$ jumps above $n^{3/4}$ well before $\Delta = n$ as stated by the next proposition.

PROPOSITION 5. *Consider any* $0 \leq \gamma \leq 0.25$. *Fix any* $\eta \in (0, 2/3)$ *and take* $\Delta = \lfloor(\frac{1}{3} + \eta)n\rfloor$. *For n large enough, we have*

$$\rho^{*2}_{\gamma,\mathrm{var}}[k_0, \Delta] \geq c_\eta \sigma_+^2 n^{5/6},$$

for some constant $c_\eta > 0$ *only depending on* η.

As a consequence, the detection problem become much more difficult when Δ is above $n/3$ and the condition on Δ in Proposition 4 is tight. In comparison to the proof of the lower bound (32), for Δ larger than $n/3$, it is possible to define a prior measure ν supported on $\mathbb{B}_0[\Delta]$, σ_0 and σ_1 such that all expectations $\sum_{i=1}^n Y_i^q$ for $q = 1, \dots, 5$ are matching under $\int \mathbb{P}_{\theta,\sigma_1}\nu(d\theta)$ and $\mathbb{P}_{0,\sigma_0}$. Matching these five moments then allows to recover the $n^{5/6}$ rate. See the proof of Proposition 5 for details.

To summarize, for $\Delta \leq \sqrt{n}$ the minimax detection distance is the same as for known variance. For $\Delta \in [\sqrt{n}, cn]$ with $c < 1/3$, the square minimax detection distance is of order $\sqrt{\Delta n^{1/2}}$ which is larger than its counterpart for known variance. For $\Delta > cn$ with $c > 1/3$, the difficulty of the testing problem greatly increases.

In view of this phenomenon, we shall restrict ourselves, for the general sparsity testing problems, to values (k_0, Δ) such that $k_0 + \Delta \le cn$ where c is some constant small enough.

3.2. *Lower bounds.* For $\Delta \le \sqrt{n} \vee k_0$, we simply use the lower bound $\rho^{*2}_{\gamma,\mathrm{var}}[k_0, \Delta] \ge \rho^{*2}_{\gamma}[k_0, \Delta]$ (where $\rho^{*2}_{\gamma}[k_0, \Delta]$ is defined for known $\sigma = \sigma_+$). The following corollary is then a direct consequence of Theorem 1.

COROLLARY 2. *Consider any $\gamma \le 0.5$. For any $k_0 \le \sqrt{n}$ and $\Delta \le n - k_0$, we have*

$$\rho^{*2}_{\gamma,\mathrm{var}}[k_0, \Delta] \ge \sigma_+ \Delta \log\left[1 + \frac{\sqrt{n}}{8\Delta}\right]. \tag{34}$$

There exists a numerical constant $c > 0$ such that the following holds. For any $k_0 > \sqrt{n}$ and $\Delta \le k_0 \wedge (n - k_0)$, we have

$$\rho^{*2}_{\gamma,\mathrm{var}}[k_0, \Delta] \ge c\sigma_+ \Delta\left[\frac{\log^2[1 + \frac{k_0}{\Delta}]}{\log[1 + \frac{k_0}{\sqrt{n}}]} \wedge \log\left[1 + \frac{k_0}{\Delta}\right]\right]. \tag{35}$$

Additional work is needed to pinpoint the minimax separation distance $\rho^{*}_{\gamma,\mathrm{var}}[k_0, \Delta]$ for $\Delta \ge \sqrt{n} \vee k_0$. As for known variance, there are two different regimes depending whether $k_0 \le \sqrt{n}$ or $k_0 > \sqrt{n}$.

THEOREM 2. *Consider any $0 \le \gamma \le 0.25$. For any $0 \le k_0 \le \sqrt{n}$ and* $\max(\sqrt{n}, 48) \le \Delta \le n - k_0$, *we have*

$$\rho^{*2}_{\gamma,\mathrm{var}}[k_0, \Delta] \ge c\sigma_+^2 \sqrt{\Delta n^{1/2}},$$

where c is a numerical constant.

For $k_0 \le \sqrt{n}$ and $\Delta \ge \sqrt{n}$, the separation distance $\rho^{*2}_{\gamma,\mathrm{var}}[k_0, \Delta]$ is the same as in the signal detection setting $\rho^{*2}_{\gamma,\mathrm{var}}[0, \Delta]$. In comparison to $\rho^{*2}_{\gamma}[k_0, \Delta]$, the squared distance $\sqrt{n}$ has increased up to $\sqrt{\Delta n^{1/2}}$. The intuition behind Theorem 2 has been already described below Proposition 4.

THEOREM 3. *Consider any $0 \le \gamma \le 0.25$. There exist three positive constants c_1, c_2 and c_3 such that the following holds. Assume that $n/c_1 \ge \Delta \ge c_1 k_0 \ge c_1\sqrt{n}$ and that $n \ge c_2$. Then we have*

$$\rho^{*2}_{\gamma,\mathrm{var}}[k_0, \Delta] \ge c_3\sigma_+^2 \frac{\sqrt{\Delta k_0}}{\log(1 + k_0/\sqrt{n})}.$$

In the known variance setting, the squared separation distance is of order $\frac{k_0}{\log(1+k_0/\sqrt{n})}$. The price to pay for not knowing the variance is a multiplicative factor of order $\sqrt{\Delta/k_0}$.

Contrary to the proof of Theorem 1 for known variance, it is difficult to follow here a moment matching approach. Given two suitable prior distributions $\mu_0^{\otimes n}$ and $\mu_1^{\otimes n}$ on θ and variances σ_0^2 and σ_1^2 in such a way that $\mu_0^{\otimes n}$ is almost supported in $\mathbb{B}_0[k_0]$ and $\mu_1^{\otimes n}$ is almost supported in $\mathbb{B}_0[k_0+\Delta, k_0, \rho]$, the goal is to prove that the two marginal distribution of Y, $\int \mathbb{P}_{\theta,\sigma_0}\mu_0^{\otimes n}(d\theta)$ and $\int \mathbb{P}_{\theta,\sigma_1}\mu_1^{\otimes n}(d\theta)$ are close to each other in total variation distance. Since the two last measures are product measures, this is equivalent to proving that the densities $\pi_0(x) := \int \phi(\frac{t-x}{\sigma_0})\mu_0(dx)$ and $\pi_1(x) := \int \phi(\frac{t-x}{\sigma_1})\mu_1(dx)$ are close in l_1 distance [recall that $\phi(\cdot)$ denotes the density of the standard normal distribution]. It is difficult to obtain an analytic expression of the l_1 distance between two mixture distribution, and hence one cannot directly choose the measure μ_0 and μ_1 minimizing this l_1 distance. As performed earlier in, for example, [11, 29], we choose instead μ_0 and μ_1 in such a way that the Fourier transforms $\widehat{\pi}_0$ and $\widehat{\pi}_1$ are matching for all frequencies small enough. Afterwards, we prove that this particular choice of μ_0 and μ_1 makes the l_1 distance between π_0 and π_1 small. Although the general approach is not new, the control of the l_1 distance is more delicate than in previous work, especially in the regime where k_0 is close to $\sqrt{n}$. In the proof, our implicit construction of the prior distributions μ_0 may be of independent interest.

3.3. *Upper bounds.* In this subsection, we build matching upper bounds for all (k_0, Δ) such that $k_0 + \Delta \le cn$ where c a numerical constant small enough. Indeed, when Δ is of order n, it has been proved in Proposition 5 that the detection problem becomes much more difficult, so that there is no hope to find tests matching Theorems 2 and 3 when $k_0 + \Delta$ is too large. Note that, in the regime $k_0 + \Delta \le cn$, one may construct a data-driven confidence interval of σ so that the knowledge of the fixed interval $[\sigma_+, \sigma_-]$ is not really critical. In Appendix A, we provide such a confidence interval and we briefly explain how to how to extend the testing procedures to completely unknown variances $\sigma \in \mathbb{R}^+$.

Throughout this subsection, we consider some fixed α and β in $(0, 1)$.

3.3.1. *Adaptive higher criticism statistic.* The principle underlying the Higher Criticism is to compare the number N_t of components of Y larger than t in absolute value to an upper bound of their expectation under the null, namely $k_0 + (n - k_0)\Phi(t/\sigma)$. This is why we adapt this test by plugging a suitable estimator of σ and adding some correcting terms accounting for the variance estimation error. Let

$$(36)\quad \widehat{\sigma}^2 = \widehat{\sigma}^2(v) := -\frac{2}{v^2}\log[\overline{\varphi}_n(v)] \qquad \text{where } v^2 := \frac{2}{\sigma_+^2}\left[\log\left(1+\frac{k_0}{\sqrt{n}}\right) \vee 1\right],$$

where we recall that $\overline{\varphi}_n$ is the empirical characteristic function (17) of Y. Let us briefly explain the idea behind this definition by replacing $\overline{\varphi}_n(v)$ by its expectation $\overline{\varphi}(v)$ (17). Intuitively, $\widehat{\sigma}^2$ is expected to be close to

$$-\frac{2}{v^2}\log\biggl[e^{-v^2\sigma^2/2}\frac{1}{n}\sum_i\cos(v\theta_i)\biggr]=\sigma^2-\frac{2}{v^2}\log\biggl[\frac{1}{n}\sum_i\cos(v\theta_i)\biggr], \tag{37}$$

so that when $\frac{1}{n}\sum_i\cos(v\theta_i)$ is close to one, $\widehat{\sigma}^2$ should be close to σ^2. Estimation of σ based on the empirical characteristic function has been first tackled by Cai and Jin [7, 27]. Nevertheless, our estimator (36) differs from theirs, as we do not assume that the nonzero components of θ are sampled from a smooth distribution.

Defining $t_{*,\alpha}^{\mathrm{HC,var}}:=\lceil 2\sqrt{2\log(\frac{4n}{\alpha})}\rceil$, we consider the test $T_{\alpha,k_0}^{\mathrm{HC,var}}$ that rejects the null hypothesis, if either $N_{\sigma_+t_{*,\alpha}^{\mathrm{HC,var}}}\geq k_0+1$ or if for some integer $t\geq 1$,

$$N_{\sigma_+t}\geq k_0+2(n-k_0)\Phi\biggl(\frac{t\sigma_+}{\hat{\sigma}}\biggr)+u_{t,\alpha}^{\mathrm{HC,var}}, \tag{38}$$

where

$$\begin{aligned}u_{t,\alpha}^{\mathrm{HC,var}}&:=\sqrt{4n\Phi(t)\log\biggl(\frac{t^2\pi^2}{\alpha}\biggr)}+\frac{2}{3}\log\biggl(\frac{t^2\pi^2}{\alpha}\biggr)\\&\quad+8t\frac{\sigma_+^3}{\sigma_-^3}\frac{k_0}{\log(1+\frac{k_0}{\sqrt{n}})}\phi(t)\sqrt{\log\biggl(\frac{6}{\alpha}\biggr)}.\end{aligned} \tag{39}$$

In comparison to the original calibration parameter $u_{t,\alpha}^{\mathrm{HC}}$, the third term is new and accounts for the estimation error of σ^2.

THEOREM 4. *Let C be any constant larger than* 1. *There exist constants c, c'_α, $c''_{\beta,\sigma_+/\sigma_-,C}$ and $c'''_{\alpha,\beta}$ such that the following holds. If $n\geq c'_\alpha$ and $k_0\leq cn$, the type I error probability of $T_{\alpha,k_0}^{B,\mathrm{var}}$ does not exceed α, that is,*

$$\mathbb{P}_{\theta,\sigma}[T_{\alpha,k_0}^{\mathrm{HC,var}}=1]\leq\alpha\qquad\forall\theta\in\mathbb{B}_0[k_0].$$

Now assume that $n\geq c''_{\beta,\sigma_+/\sigma_-,C}$. Any $\theta\in\mathbb{R}^n$ satisfying $\|\theta\|_0\leq cn$,

$$|\theta_{(k_0+q)}|\geq c'''_{\alpha,\beta}\sigma_+\biggl[\sqrt{\log(C)}+\sqrt{\log\biggl(\frac{\sigma_+}{\sigma_-}\biggr)}+\sqrt{\log\biggl(2+\frac{k_0\vee\sqrt{n}}{q}\biggr)_+}\biggr], \tag{40}$$

for some $q\in[1,n-k_0]$ and

$$\sum_{i=1}^n[(v\theta_i)^4\wedge 1]\leq C(k_0\vee\sqrt{n}), \tag{41}$$

belongs to the high probability rejection region of $T_{\alpha,k_0}^{B,\mathrm{var}}$, that is, $\mathbb{P}_{\theta,\sigma}[T_{\alpha,k_0}^{\mathrm{HC,var}}=0]\leq\beta$.

Condition (41) aside, the behavior of $T_{\alpha,k_0}^{\mathrm{HC},\mathrm{var}}$ is similar to the one of $T_{\alpha,k_0}^{\mathrm{HC}}$ as stated in Proposition C.2. In fact, Condition (41) allows to bound the term $\frac{1}{n}\sum_i \cos(v\theta_i)$ in (37) and ensures that $|\widehat{\sigma}^2 - \sigma^2|$ is, with high probability, at most of order $\frac{k_0}{n\log(1+k_0/\sqrt{n})}$. When this condition (41) is not met, we are unable to control the behavior of the adaptive Higher Criticism test. Nevertheless, it turns out that parameters θ not satisfying (41) belong to the high-probability rejection region of the test $T_{\alpha,k_0}^{B,\mathrm{var}}$ described below so that a combination of $T_{\alpha,k_0}^{\mathrm{HC},\mathrm{var}}$ and $T_{\alpha,k_0}^{B,\mathrm{var}}$ achieves similar performances to the original Higher Criticism test $T_{\alpha,k_0}^{\mathrm{HC},\mathrm{var}}$. At the end of the section, the constant C in Theorem 4 is carefully chosen to put the three tests $T^{\mathrm{HC},\mathrm{var}}$, $T^{B,\mathrm{var}}$ and $T^{I,\mathrm{var}}$ together.

3.3.2. *Detecting the signal in the bulk distribution.* Analogously to the above extension of the Higher Criticism test, it would be natural to plug a variance estimator $\widehat{\sigma}^2$ in the statistic $Z(s)$ (16) and then to build a test based on this data-driven statistic. Unfortunately, it turns out that the estimation error for such $\widehat{\sigma}$ is not negligible in the dense setting. Such a phenomenon is not unexpected as we have proved in Theorem 3 that no test in the unknown variance setting can perform as well as T_{α,k_0}^{B} for known σ.

This is why we define a new statistic which is almost invariant with respect to the noise variance. Denoting P_B the linear polynom $P_B(\xi) := 4\xi - 3$, we define, for $s > 0$, the statistic $Z^{\mathrm{var}}(s)$

$$Z^{\mathrm{var}}(s) := n\int_0^1 P_B(\xi)\log\Big[\Big(\overline{\varphi}_n\Big(\frac{s\xi}{\sigma_+}\Big)\Big)_+\Big]\,d\xi. \tag{42}$$

The polynom P_B has been defined in such a way that $\int_0^1 P_B(\xi)\xi^2\,d\xi = 0$. To understand the rationale behind $Z^{\mathrm{var}}(s)$, let us assume that $\overline{\varphi}_n(s\xi)$ is close to its expectation $\overline{\varphi}(s\xi)$. Since for x close to 1, $\log(x)$ is approximately $x - 1$, we obtain

$$\begin{aligned} Z^{\mathrm{var}}(s) &\approx n\int_0^1 P_B(\xi)\Big[-\frac{\xi^2 s^2\sigma^2}{2\sigma_+^2} + \log\Big(\frac{1}{n}\sum_{i=1}^n \cos\Big(\frac{s\xi\theta_i}{\sigma_+}\Big)\Big)\Big]\,d\xi \\ &\approx \sum_{i=1}^n \int_0^1 P_B(\xi)\Big(\cos\Big(\frac{s\xi\theta_i}{\sigma_+}\Big) - 1\Big)\,d\xi = \sum_{i=1}^n g\Big(\frac{s\theta_i}{\sigma_+}\Big), \end{aligned}$$

where $g(x) = \int_0^1 P_B(\xi)(\cos(\xi x) - 1)\,d\xi$. For small x, a Taylor expansion of the cos function enforces that $g(x) \approx \int_0^1 P_B(\xi)[-\xi^2\frac{x^2}{2} + \xi^4\frac{x^4}{12}]\,d\xi = x^4\int_0^1 P_B(\xi) \times \frac{\xi^4}{12}\,d\xi > 0$. For larger x (in absolute value), one can prove that $g(x)$ is positive and bounded away from zero. As a consequence, $\sum_{i=1}^n g(s\theta_i/\sigma_+)$ behaves like $\sum_{i=1}^n[(s\theta_i/\sigma_+)^4 \wedge 1]$ and approximates $\|\theta\|_0$. This informal discussion is made rigorous in the proof of Theorem 5 below. In practice, we set

$$s_{k_0}^{\mathrm{var}} = \left[\sqrt{1 + \log\left(\frac{k_0}{n^{1/2}}\right)} \vee 1\right], \tag{43}$$

and we define $T_{\alpha,k_0}^{B,\mathrm{var}}$ as the test rejecting the null hypothesis for large values of $Z^{\mathrm{var}}(s_{k_0}^{\mathrm{var}})$, that is when

$$Z^{\mathrm{var}}(s_{k_0}^{\mathrm{var}}) \geq 1.09k_0 + 16\frac{k_0^2}{n} + 4\sqrt{e}\left(\sqrt{k_0 n^{1/2}} \vee \sqrt{n}\right)\sqrt{\log(2/\alpha)}. \tag{44}$$

THEOREM 5. *There exist numerical constants c, c' and $c''_{\alpha,\beta}$ such that the following holds. Assume that $n \geq c$ and that $k_0 \leq c'n$. For any k_0-sparse vector θ, the type I error probability of $T_{\alpha,k_0}^{B,\mathrm{var}}$ is small, that is,*

$$\mathbb{P}_{\theta,\sigma}\left[T_{\alpha,k_0}^{B,\mathrm{var}} = 1\right] \leq \alpha + \frac{2(\|\theta\|_1/\sigma_+ + n)}{n^4}. \tag{45}$$

Any $\theta \in \mathbb{R}^n$ such that $\|\theta\|_0 \leq c'n$, and

$$\sum_{i=k_0+1}^{n} \left[\left(\frac{s_{k_0}^{\mathrm{var}}\theta_{(i)}}{\sigma_+}\right)^4 \wedge 1\right] \geq c''{}_{\alpha,\beta}(k_0 \vee \sqrt{n}) \tag{46}$$

belongs to the high probability rejection region of $T_{\alpha,k_0}^{B,\mathrm{var}}$, that is,

$$\mathbb{P}_{\theta,\sigma}\left[T_{\alpha,k_0}^{B,\mathrm{var}} = 0\right] \leq \beta + \frac{2(\|\theta\|_1/\sigma_+ + n)}{n^4}.$$

The sufficient condition (46) for $T_{\alpha,k_0}^{B,\mathrm{var}} = 1$ to be powerful corresponds to the heuristics described above. This condition will be the main ingredient towards matching the $\sigma_+^2 \frac{\sqrt{\Delta k_0}}{\log(1+k_0/\sqrt{n})}$ separation distance of Theorem 3.

The main downside to the above theorem is the presence of the small term $\|\theta\|_1/(\sigma_+ n^4)$ in the type I and type II error probabilities. Although in most relevant case this term will be negligible, this makes the supremum of the type I error bound (45) over all $\theta \in \mathbb{B}_0[k_0]$ infinite. In Section 3.3.4, we sketch a trimming approach which amounts to first discarding components large components Y and then applying the test to the trimmed vector $\tilde{Y}$.

3.3.3. *Intermediary regimes.* As for T_{α,k_0}^{B}, one cannot easily adapt T_{α,k_0}^{I} by plugging an estimator of σ. Following the same approach as above, we modify the statistic by considering the logarithm of the empirical characteristic function and multiplying it by some suitable polynom.

As the following test aims at discovering intermediary signals whose signature is neither in the bulk of the empirical distribution of (Y_i) nor in its extreme values, we restrict ourselves to the case $k_0 \geq 20\sqrt{n}$ (as for T^I_{α,k_0}). Consider the dyadic collection $\mathcal{L}_{k_0}$ defined in Section 2.2.3. For $l \in \mathcal{L}_{k_0}$, let

$$r_{k_0,l} := \sqrt{16\log\left(\frac{k_0}{l}\right)}, \qquad w_l := \sqrt{\log\left(\frac{l}{\sqrt{n}}\right)}. \tag{47}$$

Note that, if w_l is defined as in (22) for T^I_{α,k_0}, the definition of $r_{k_0,l}$ is slightly different. Equipped with this notation, we consider the statistic

$$V^{\text{var}}(r_{k_0,l}, w_l) := nr_{k_0,l}\int_{-1}^{1} P_l(r_{k_0,l}\xi)\phi(r_{k_0,l}\xi)\log\left[\overline{\varphi}_n\left(\frac{w_l\xi}{\sigma_+}\right)_+\right]d\xi, \tag{48}$$

where $P_l(t) = \gamma_l[\zeta_l t^2 - \kappa_l]$ with

$$\begin{aligned} \kappa_l &:= -2r^3_{k_0,l}\phi(r_{k_0,l}) - 6r\phi(r_{k_0,l}) + 3\big(1 - 2\Phi(r_{k_0,l})\big),\\ \zeta_l &:= -2r_{k_0,l}\phi(r_{k_0,l}) + 1 - 2\Phi(r_{k_0,l}), \\ \gamma_l &:= [\kappa_l - \zeta_l]^{-1} \quad \text{and} \quad \delta_l := 4\gamma_l\big(r_{k_0,l} + 4r^{-1}_{k_0,l}\big)\phi(r_{k_0,l}). \end{aligned} \tag{49}$$

The purpose of this polynom P_l is to cancel the term $\int_{-1}^1 P_l(r_{k_0,l}\xi)\phi(r_{k_0,l}\xi)\xi^2\,d\xi$. Heuristically, $\log[\overline{\varphi}_n(w_l\xi/\sigma_+)_+]$ should be close to

$$\begin{aligned} \log\left[\overline{\varphi}\left(\frac{w_l\xi}{\sigma_+}\right)_+\right] &= -\frac{\sigma^2 w_l^2\xi^2}{2\sigma_+^2} + \log\left[\frac{1}{n}\sum_i \cos\left(\frac{w_l\xi\theta_i}{\sigma_+}\right)\right] \\ &\approx -\frac{\sigma^2 w_l^2\xi^2}{2\sigma_+^2} + \frac{1}{n}\sum_i\left[\cos\left(\frac{w_l\xi\theta_i}{\sigma_+}\right) - 1\right]. \end{aligned}$$

Since $P_l(r_{k_0,l}\xi)\phi(r_{k_0,l}\xi)$ is orthogonal to ξ^2, we expect that

$$V^{\text{var}}(r_{k_0,l}, w_l) \approx \sum_{i=1}^n r_{k_0,l}\int_{-1}^1 P_l(r_{k_0,l}\xi)\phi(r_{k_0,l}\xi)\left[\cos\left(\frac{w_l\xi\theta_i}{\sigma_+}\right) - 1\right]d\xi.$$

Each term of this sum is zero for $\theta_i = 0$. More generally, we show in the proof of Theorem 6 that, when θ does not contain too many large coefficients, this sum approximates the number of coefficient larger than $r^2_{k_0,l}/w_l$.

Finally, let $T^{I,\text{var}}_{\alpha,k_0}$ be the test rejecting the null hypothesis, if for some $l \in \mathcal{L}_{k_0}$, $V^{\text{var}}(r_{k_0,l}, w_l)$ is large enough, that is,

$$V^{\text{var}}(r_{k_0,l}, w_l) \geq k_0(1 + \delta_l) + 32\frac{k_0^2}{n} + 8\sqrt{ln^{1/2}\log\left(\frac{\pi^2[1 + \log_2(l/l_0)]^2}{3\alpha}\right)}. \tag{50}$$

THEOREM 6. *There exist numerical constants c, c', $c''_{\alpha,\beta}$ and $c'''_{\alpha,\beta}$ such that, for any $C > 2$, the following holds. Assume that $n \geq c$ and that $k_0 \leq c'n$. For any k_0-sparse vector θ, the type I error probability of $T^{I,\text{var}}_{\alpha,k_0}$ is small, that is,*

$$\mathbb{P}_{\theta,\sigma}\big[T^{I,\text{var}}_{\alpha,k_0} = 1\big] \leq \alpha + \frac{2(\|\theta\|_1/\sigma_+ + n)}{n^4}.$$

Recall $s^{\mathrm{var}}_{k_0}$ defined in (43). *Any parameter $\theta \in \mathbb{R}^n$ satisfying $\|\theta\|_0 \le c'n$ and the two following properties*:

$$\sum_{i=1}^{n} \mathbf{1}_{s^{\mathrm{var}}_{k_0}|\theta_i|\ge\sigma_+} \le Ck_0, \tag{51}$$

$$|\theta_{(k_0+q)}| \ge c''_{\alpha,\beta}\log(C)\sigma_+ \frac{1+\log(1+\frac{k_0}{q\wedge k_0})}{\sqrt{\log(1+\frac{k_0}{\sqrt{n}})}} \quad \textit{for some } q \ge c'''_{\alpha,\beta}C^2\biggl[\sqrt{k_0 n^{1/2}} \vee \frac{k_0^2}{n}\biggr], \tag{52}$$

belongs to the high probability rejection region of $T^{I,\mathrm{var}}_{\alpha,k_0}$, that is,

$$\mathbb{P}_{\theta,\sigma}\bigl[T^{I,\mathrm{var}}_{\alpha,k_0}=0\bigr] \le \beta + \frac{2(\|\theta\|_1/\sigma_+ + n)}{n^4}.$$

Condition (52) for $T^{I,\mathrm{var}}_{\alpha,k_0}$ to be powerful is analogous to Condition (38) in the Supplementary Material for T^{I}_{α,k_0} in the known variance setting except that q is now restricted to be larger than k_0^2/n. This restriction will turn out to be benign except when k_0 is too close to n. Also, contrary to Proposition C.4, θ is assumed to contain less than Ck_0 coefficients larger than $\sigma_+/s^{\mathrm{var}}_{k_0}$ [which is of order $\sigma_+\log^{-1/2}(k_0/\sqrt{n})$]. Again, this restriction is not a serious issue as $T^{B,\mathrm{var}}_{\alpha,k_0}$ is powerful for such θ not satisfying this assumption.

3.3.4. *Combination of the tests.* For any integers $k_0 \ge 0$ and $q > 0$, define $\psi^{\mathrm{var}}_{k_0,q} > 0$ by

$$(\psi^{\mathrm{var}}_{k_0,q})^2 := \begin{cases} \sigma_+^2 \log\biggl[1+\dfrac{\sqrt{n}}{q}\biggr] & \text{if } k_0 \le \sqrt{n} \text{ and } q \le \sqrt{n}, \\ \sigma_+^2\biggl(\dfrac{\sqrt{n}}{q}\biggr)^{1/2} & \text{if } k_0 \le \sqrt{n} \text{ and } q > \sqrt{n}, \\ \sigma_+^2\biggl(\dfrac{\log^2(1+\frac{k_0}{q})}{\log(1+\frac{k_0}{\sqrt{n}})} \wedge \log\biggl[1+\dfrac{k_0}{q}\biggr]\biggr) & \\ & \text{if } k_0 > \sqrt{n} \text{ and } q \le k_0, \\ \sigma_+^2 \dfrac{k_0^{1/2}}{q^{1/2}\log(1+\frac{k_0}{\sqrt{n}})} & \text{if } k_0 > \sqrt{n} \text{ and } q > k_0. \end{cases} \tag{53}$$

Let $T^{C,\mathrm{var}}_{\alpha,k_0}$ denote the aggregation of the three previous tests, that is,

$$T^{C,\mathrm{var}}_{\alpha,k_0} := \max\bigl(T^{\mathrm{HC,var}}_{\alpha/3,k_0}, T^{B,\mathrm{var}}_{\alpha/3,k_0}, T^{I,\mathrm{var}}_{\alpha/3,k_0}\bigr) \qquad \text{if } k_0 \ge 20\sqrt{n}$$

and

$$T^{C,\mathrm{var}}_{\alpha,k_0} := \max\bigl(T^{\mathrm{HC},\mathrm{var}}_{\alpha/2,k_0}, T^{B,\mathrm{var}}_{\alpha/2,k_0}\bigr) \qquad \text{else.}$$

As pointed out above, it is not possible to control uniformly the type I error probability of this test as such probabilities depend on the l_1 norm of θ. This is why introduce a trimmed version of this test by removing large components of Y. Given $z > 0$ and $V \in \mathbb{R}^n$, let $\mathcal{S}(z; V) = \{i \in [n], |V_i| > (z+1)\sigma_+ n^2\}$. Let $U \sim \mathcal{U}[0,1]$ be an uniformly distributed random variable independent of Y. We write $\mathcal{S}(U,Y) = \mathcal{S}[(U+1)\sigma_+ n^2; Y]$ for the coordinates i such that $|Y_i| > (U+1)\sigma_+ n^2$. Let $\widetilde{Y}(\mathcal{S}(U,Y)) := (Y_i), i \in ([n] \setminus \mathcal{S}(U,Y))$ be the subvector of Y of size $n - |\mathcal{S}(U,Y)|$. Finally, we define the trimmed test $\overline{T}^{C,\mathrm{var}}_{\alpha,k_0}$ rejecting the null hypothesis if either $k_0 - |\mathcal{S}(U,Y)|$ is negative or if the test $T^{C,\mathrm{var}}_{\alpha,k_0-|\mathcal{S}(U,Y)|}$ applied to the size $n - |\mathcal{S}(U,Y)|$ vector $\widetilde{Y}(\mathcal{S}(U,Y))$ rejects the null hypothesis.

We use a random threshold $(U+1)\sigma_+ n^2$ instead of a deterministic one to make the subset $\mathcal{S}$ of trimmed variable almost independent of Y, which facilitate the analysis of the two-step procedure $\overline{T}^{C,\mathrm{var}}_{\alpha,k_0}$.

COROLLARY 3. *Fix any $\xi \in (0,1)$. There exist positive constants c, c', $c''_{\alpha,\beta,\xi}$ and $c'''_{\alpha,\beta,\xi}$ such that the following holds. Consider any $k_0 \le n^{1-\xi}$ and $n \ge c$. Then, for any $\theta \in \mathbb{B}_0[k_0]$, one has*

$$\mathbb{P}_{\theta,\sigma}\bigl[\overline{T}^{C,\mathrm{var}}_{\alpha,k_0} = 1\bigr] \le \alpha + \frac{c'\log(n)}{n}.$$

Moreover, $\mathbb{P}_{\theta,\sigma}[\overline{T}^{C,\mathrm{var}}_{\alpha,k_0} = 1] \ge 1 - \beta - \frac{c'\log(n)}{n}$ for any vector θ satisfying $\|\theta\|_0 \le c'n$ and

$$(54) \qquad |\theta_{(k_0+q)}| \ge c''_{\alpha,\beta,\xi}\sigma_+\psi^{\mathrm{var}}_{k_0,q} \qquad \textit{for some } q \in [1, n-k_0].$$

Also, $\mathbb{P}_{\theta,\sigma}[\overline{T}^{C,\mathrm{var}}_{\alpha,k_0} = 1] \ge 1 - \beta - \frac{c'\log(n)}{n}$ for any vector θ satisfying

$$(55) \qquad \begin{aligned} &\theta \in \mathbb{B}_0(k_0+\Delta) \quad \textit{and} \\ &d^2[\theta, \mathbb{B}_0(k_0)] \ge c'''_{\alpha,\beta,\xi}\sigma_+^2\Delta\bigl(\psi^{\mathrm{var}}_{k_0,\Delta}\bigr)^2 \qquad \textit{for some } \Delta \in [1, c'n - k_0]. \end{aligned}$$

As a consequence, for $k_0 \le n^{1-\xi}$ [and ξ is an arbitrary constant in $(0,1)$], $\overline{T}^{C}_{\alpha,k_0}$ simultaneously achieves the minimax separation distance for all Δ such that $k_0 + \Delta \le cn$ where c is constant small enough.

Building on the statistics introduced in this section, one can then construct an adaptive estimator of the sparsity for unknown variance in the spirit of what will be done in Section 4. For reasons of space, we do not pursue this direction.

4. Sparsity estimation.

Given an observation Y, our goal is now to estimate the number $\|\theta\|_0$ of nonzero components of θ. As explained in the Introduction, this estimation problem can be rephrased as a multiple testing problem. Let $\mathcal{H} = (H_k)_{k=0,\dots,n}$ denote the nested collection of all hypotheses H_k (4). For a parameter θ, the set of true hypotheses $\mathcal{T}(\theta)$ is the collection $\{H_k, k \geq \|\theta\|_0\}$ and the set of false hypotheses $\mathcal{R}(\theta)$ is the collection $\{H_k, k < \|\theta\|_0\}$. A multiple hypothesis test is a measurable collection $\widehat{\mathcal{R}} \subset \mathcal{R}$.

Let us make explicit the connection between these two problems. Given an estimator $\widehat{k}$ of $\|\theta\|_0$, taking $\widehat{\mathcal{R}} = \{H_k, k < \widehat{k}\}$ defines a multiple test. Conversely, consider a multiple test $\widehat{\mathcal{R}}$. Then one may define the estimator $\widehat{k} = 1 + \max\{k : H_k \in \widehat{\mathcal{R}}\}$. In our framework, a closed test $\widehat{\mathcal{R}}$ is a test that satisfies the property "$H' \subset H$ and $H \subset \widehat{\mathcal{R}}$ implies $H' \subset \widehat{\mathcal{R}}$" (see, e.g., [18]). It follows from the above construction that sparsity estimators $\widehat{k}$ are in one to one correspondence with closed testing procedures.

The above correspondence leads us (i) to build estimators $\widehat{k}$ that rely on the test statistics introduced in the Section 2 and (ii) to evaluate the performances of $\widehat{k}$ in terms of separation distances of a multiple testing procedure.

4.1. *From single tests to multiple tests.*

Fix some $\alpha \in (0, 1)$. We introduce an estimator $\widehat{k}$

$$\widehat{k} := \lceil \widehat{k}_{\mathrm{HC}} \rceil \vee \lceil \widehat{k}_B \rceil \vee \lceil \widehat{k}_I \rceil, \tag{56}$$

that combines statistics for the three regimes (Higher Criticism, Bulk, Intermediary) unveiled in Section 2.

Construction of $\widehat{k}_{\mathrm{HC}}$. Let $t_* := t^{\mathrm{HC}}_{*,\alpha/3}$ where $t^{\mathrm{HC}}_{*,\alpha/3}$ is defined in Section 2.2.1 and write $\mathcal{T} = \{1, \dots, t_*\}$. Define the Higher Criticism estimator of $\|\theta\|_0$ by

$$\widehat{k}_{\mathrm{HC}} := N_{\sigma t_*} \vee \sup_{t \in \mathcal{T}} \frac{N_{\sigma t} - 2n\Phi(t) - u^{\mathrm{HC}}_{t,\alpha/3}}{1 - 2\Phi(t)}, \tag{57}$$

where N_t and $u^{\mathrm{HC}}_{t,\alpha}$ are introduced in Section 2.2.1. Note that $\widehat{k}_{\mathrm{HC}}$ is reminiscent of the estimators of Meinshausen and Rice [36] and Li and Siegmund [33] developed for mixtures. Let us explain the rationale behind this estimator. Recall that $N_{\sigma t_*}$ is the number of coordinates of Y larger than t_* (in absolute value). With high probability, all coordinates Y_i larger than t_* have a nonzero mean, which implies $N_{\sigma t_*} \leq \|\theta\|_0$. For $t \in \mathcal{T}$, it follows from Bernstein inequality that with high probability, among the coordinates Y_i larger than σt, at most $2(n - \|\theta\|_0)\Phi(t) + u^{\mathrm{HC}}_{t,\alpha/3}$ of them correspond to null coordinates. As a consequence, $(N_{\sigma t} - 2n\Phi(t) - u^{\mathrm{HC}}_{t,\alpha/3})/(1 - 2\Phi(t))$, is with high probability less than $\|\theta\|_0$. Since the optimal choice of the tuning parameter t depends on θ, we simply pick the largest of all these estimators. See the proof of Theorem 7 for more details.

Construction of $\widehat{k}_B$ *and* $\widehat{k}_I$. Following the intuition explained in the Introduction, it would be tempting to define $\widehat{k}_B - 1$ as the largest $q \in [n]$ such that the test $T^B_{\alpha_q,q}$ (with some suitable tuning parameters α_q) rejects the null. However, this simple strategy leads to a logarithmic loss in comparison to the optimal testing separation rate. As explained in Sections 2.2.2 and 2.2.3, the statistics $Z(s)$ and $V(r,w)$ involved in the tests T^B_{α,k_0} and T^I_{α,k_0} can be interpreted as (possibly biased) estimators of $\|\theta\|_0$. The bias and the variance of these estimators depends on choice of the tuning parameters s, r and w. For instance, for a large value of s, the variance $Z(s)$ is higher but $\mathbb{E}_\theta[Z(s)]$ is close to $\|\theta\|_0$ (see Section 2.2.2). As the optimal value of these tuning parameters depends on θ, we shall compute these statistics for a collection of tuning parameters and take the largest estimator.

Define $k_{\min} := \lceil\sqrt{n}\rceil$ and consider the dyadic collection $\mathcal{K}_0 := \{k_{\min}, 2k_{\min}, \dots, k_{\max}\}$, where $k_{\max} \in (n/2; n]$. In order to calibrate this large collection of statistics, we have to adjust the thresholds $u^B_{k_0,\alpha}$ and $u^I_{k_0,l,\alpha}$ of the statistics. For any $k_0 \in \mathcal{K}_0$, denote $\alpha_{k_0} := 2\alpha([1+\log_2(\frac{k_0}{k_{\min}})]^2\pi^2)^{-1}$ so that $\sum_{k_0\in\mathcal{K}_0}\alpha_{k_0} \le \alpha/3$. Equipped with this notation, we define the Bulk and Intermediary estimators of $\|\theta\|_0$ as follows:

$$\widehat{k}_B := \sup_{k_0\in\mathcal{K}_0} Z(s_{k_0}) - u^B_{k_0,\alpha_{k_0}}, \tag{58}$$

$$\widehat{k}_I := \sup_{k_0\in\mathcal{K}_0, k_0\ge 20\sqrt{n}} \sup_{l\in\mathcal{L}_{k_0}} \frac{V(r_{k_0,l}, w_l) - u^I_{k_0,l,\alpha_{k_0}}}{1 + l/k_0}, \tag{59}$$

where $Z(s)$, $V(r,w)$, $u^B_{k_0,\alpha}$ and $u^I_{k_0,l,\alpha}$ are introduced in Sections 2.2.2 and 2.2.3.

REMARK. The number of statistics required to compute $\widehat{k}$ is of order $\log^2(n)$.

4.2. *Optimal sparsity estimation rates.*

THEOREM 7. *Fix any* $\beta \in (0,1)$. *There exist two positive constants* $c_{\alpha,\beta}$ *and* $c'_{\alpha,\beta}$ *such that the following holds for any* $\theta \in \mathbb{R}^n$. *With probability higher than* $1-\alpha$, $\widehat{k}$ *does not overestimate the number of nonzero components*,

$$\mathbb{P}_\theta[\widehat{k} > \|\theta\|_0] \le \alpha. \tag{60}$$

With probability higher than $1-\beta$, *the vector* θ *contains no more than* $\widehat{k}$ *large coefficients in the sense that*

$$|\theta_{(\widehat{k}+q)}| \le c_{\alpha,\beta}\sigma\psi_{\hat{k},q} \qquad \forall q = 1,\dots,n-\widehat{k} \tag{61}$$

and

$$d^2[\theta, \mathbb{B}_0(\widehat{k})] \le c'_{\alpha,\beta}\sigma^2[\|\theta\|_0 - \hat{k}]_+ \psi^2_{\hat{k},(\|\theta\|_0-\hat{k})_+}, \tag{62}$$

where the sequence ψ *is defined in equation* (24).

As a consequence, outside an event of probability less than $\alpha + \beta$, we have $\widehat{k} \leq \|\theta\|_0$ and θ is so close to $\mathbb{B}_0[\widehat{k}]$ that it is impossible to reliably decipher whether $\theta \in \mathbb{B}_0[\widehat{k}]$ or not. Alternatively, Theorem 7 provides the following data-driven certificate: with high probability and simultaneously for all $q \geq 1$, there are no more than $\widehat{k} + q$ coefficients larger (up to constants) than $\psi_{\widehat{k},q}$. In Section B, we restate Theorem 7 in terms of separation distances for multiple testing procedures.

For a given θ, we can easily "invert" the conditions (61) and (62) to control the error $|\widehat{k} - \|\theta\|_0|$.

COROLLARY 4. *There exists a positive constant $c_{\alpha,\beta}$ such that the following holds. For any $\theta \in \mathbb{R}^n$, the sparsity estimator satisfies the three following properties*:

$$\widehat{k} \leq \|\theta\|_0, \tag{63}$$

$$\big(\|\theta\|_0 - \widehat{k}\big)_+ < \min\{q, s.t.\ d_2^2\big(\theta, \mathbb{B}_0[\|\theta\|_0 - q]\big) \geq c_{\alpha,\beta}\sigma^2 q \psi^2_{\|\theta\|_0-q,q}\}, \tag{64}$$

$$\widehat{k} \geq 1 + \max\{r, s.t.\ \exists q \in [n-r], |\theta_{(r+q)}| \geq c_{\alpha,\beta}\sigma\psi_{r,q}\}, \tag{65}$$

outside an event of probability less than $\alpha + \beta$. In the above equations, we choose the convention $\min\{\emptyset\} = \infty$ *and* $\max\{\emptyset\} = -\infty$.

Conversely, it is not possible to improve the bounds (64) and (65).

PROPOSITION 6. *There exists a positive constant $c'_{\alpha,\beta}$ such that the following holds. Fix any integers $q > 0$ and $k > 0$ such that $k + q \leq n$. No estimator $\tilde{k}$ can satisfy simultaneously $\inf_{\theta \in \mathbb{B}_0[k]} \mathbb{P}_\theta[\tilde{k} \leq k] \geq 1 - \alpha$ and at least one of the two following properties*:

$$\inf_{\theta \in \mathbb{B}_0[k+q,k,c'_{\alpha,\beta}\sigma\sqrt{q}\psi_{k,q}]} \mathbb{P}_\theta\big[\tilde{k} \geq \|\theta\|_0 - q\big] \geq 1 - \beta, \tag{66}$$

$$\inf_{\theta \in \mathbb{R}^n, |\theta_{(k+q)}| \geq c'_{\alpha,\beta}\sigma\psi_{k,q}} \mathbb{P}_\theta[\tilde{k} > k] \geq 1 - \beta. \tag{67}$$

For any fixed (r, q), if we replace $\psi^2_{r,q}$ in (64) by $\frac{c'_{\alpha,\beta}}{c_{\alpha,\beta}}\psi^2_{r,q}$, then (63) cannot hold together with (64) on an event of large probability. The same optimality results holds for (65).

To better grasp the implication of (64), let us consider a toy example with $\|\theta\|_0 = n^\gamma$ for some $\gamma \in (0, 1)$. Given $\Delta \in [1, \ldots, \|\theta\|_0]$, we define $m_\Delta^2 = \frac{1}{\Delta}\sum_{j=1}^{\Delta} \theta^2_{(\|\theta\|_0+1-j)}$ the mean square of the Δ smallest nonzero values of θ. Note that m_Δ is a nondecreasing function of Δ. It corresponds to the typical value of the Δ smallest nonzero components of θ. Depending on the behavior of m_Δ we may bound the error of the estimator of $\|\theta\|_0$. First, if m_1 is large compared to $\sqrt{\log(n)}$, then $\widehat{k} = \|\theta\|_1$ with high probability. Then we analysis is divided into two subcases depending on $\|\theta\|_0$:

(i) $\gamma \in (0, 1/2)$ (sparse vector). Take $\Delta = n^{\zeta}$ with $\zeta \in (0, \gamma]$.

$$\text{If } m_{\Delta} \ge c_{\alpha,\beta}\sigma\sqrt{(1/2-\zeta)\log(n)} \qquad \text{then } \frac{\|\theta\|_0 - \widehat{k}}{\|\theta\|_0} \le n^{\zeta-\gamma}.$$

Conversely, if $m_{\|\theta\|_0} \le c'_{\alpha,\beta}\sigma\sqrt{(1/2-\gamma)\log(n)}$, then it is impossible to distinguish θ from 0. As a consequence, the relative estimation precision is mainly driven by the proportion of nonzero components that are large compared to $\sigma\sqrt{\log(n)}$.

(ii) $\gamma \in (1/2, 1)$ (sparse vector). Here, the situation is more intricate:

(a) $\Delta = n^{\zeta}$ with $\zeta \in (0, \gamma)$.

$$\text{If } m_{\Delta} \ge c_{\alpha,\beta}\sigma\left[\sqrt{2(\gamma-\zeta)} \wedge \frac{2(\gamma-\zeta)}{\sqrt{\gamma-1/2}}\right]\sqrt{\log(n)}$$

$$\text{then } \frac{\|\theta\|_0 - \widehat{k}}{\|\theta\|_0} \le n^{\zeta-\gamma}.$$

In that case, all nonzero components of θ except a polynomially small proportion of them are larger than $\sigma\sqrt{\log(n)}$ and the relative estimation error $\frac{|\|\theta\|_0-\widehat{k}|}{\|\theta\|_0}$ converges polynomially fast to zero.

(b) $\Delta = \frac{\|\theta\|_0}{u_n}$ with $u_n \to \infty$ and $u_n n^{-\zeta} \to 0$ for all $\zeta > 0$.

$$\text{If } m_{\Delta} \ge c_{\alpha,\beta}\sigma\frac{\log(u_n)}{\sqrt{(\gamma-1/2)\log(n)}} \qquad \text{then } \frac{\|\theta\|_0 - \widehat{k}}{\|\theta\|_0} \le \frac{1}{u_n}.$$

For concreteness, fix $u_n = \log^{\zeta}(n)$ with $\zeta > 0$. the relative convergence rate is of order $\log^{-\zeta}(n)$ if all nonzero components of θ except a proportion u_n^{-1} of them are larger than $\sigma\zeta\frac{\log\log(n)}{\sqrt{\log(n)}}$.

(c) $\Delta = \zeta\|\theta\|_0$ with some $\zeta \in (0, 1)$. If $m_{\Delta} \ge c_{\alpha,\beta}\sigma\frac{\log(1/\zeta)}{\sqrt{\gamma\log(n)}}$, then $\frac{\|\theta\|_0-\widehat{k}}{\|\theta\|_0} \le (1-\zeta)$. In that setting, a fixed proportion of nonzero coefficients are larger than $\sigma\frac{1}{\sqrt{\log(n)}}$. One is able to estimate $\|\theta\|_0$ up to a constant multiplicative factor.

(d) More generally, consider $\Delta = \|\theta\|_0(1-\frac{1}{u_n})$ with $u_n \to \infty$.

$$\text{If } m_{\Delta} \ge c_{\alpha,\beta}\sigma\frac{1}{\sqrt{u_n\log(1+\frac{\|\theta\|_0}{u_n\sqrt{n}})}} \qquad \text{then } \widehat{k} \ge \frac{\|\theta\|_0}{u_n}.$$

For instance, take $u_n = n^{\zeta}$ for $\zeta \in (0, \gamma)$. Even if most nonzero components of θ, are polynomially small, it is still possible to distinguish θ from zero, but it is just possible to estimate $\log(\|\theta\|_0)$ up to a multiplicative constant.

Finally, let us emphasize that all these convergence rates are optimal in the sense of Corollary 4 and Proposition 6.

Comparison with the literature. In [8], Cai et al. consider an asymptotic framework where $\|\theta\|_0 = n^\gamma$ with $\gamma \in (0, 1/2)$ and θ only takes the values 0 and $\sigma\sqrt{2r\log(n)}$ for some $r > 0$. These authors obtain convergence rates similar to Case (i) above but with explicit optimal constant $c(\alpha, \beta)$. In [7], Cai and Jin consider an asymptotic framework where the nonzero components of θ are sampled according to a fixed distribution with a smooth density h in the sense that its characteristic function decays at rate not slower than $t^{-\alpha}$ for some $\alpha > 2$. Their estimator $\widetilde{k}$ [7], Section 3.1, achieves a relative convergence rate of order $\log^{-\alpha/2}(n)$. However, if h does not satisfy an uniform smoothness assumption, then $\widetilde{k}$ can be inconsistent. According to Case (ii)(b), when h is continuous at 0, the relative convergence rate of our estimator $\widehat{k}$ is of order $\frac{\log\log(n)}{\sqrt{\log(n)}}$. This rate is slightly slower than that of Cai and Jin when h is highly smooth, but our estimator is not tailored to vectors θ that are sampled according to a smooth distribution and is valid for all θ.

5. Discussion.

5.1. *Other noise distributions*. Some of our testing procedures heavily rely on the assumption that the noise distribution is Gaussian. For instance, the behavior of the Bulk and Intermediary statistics depends on the exact form of the characteristic function of the noise. The radical change in the rates between the known variance case, and the unknown variance case, is already eloquent on the importance of knowing the exact shape of the noise distribution—even a slight deformation of the noise distribution by changing the variance has a strong effect on the minimax separation distances. We may consider two different extensions to non-Gaussian noises:

1. The noise distribution is not Gaussian but is explicitly known. For the sake of discussion, let us also assume that it is symmetric. In that case, one could adapt the Higher Criticism statistic by replacing $\Phi(\cdot)$ by the survival function of this distribution. Also, both the Bulk and Intermediary statistics could be accommodated by replacing $\exp(-\xi^2 w^2/2)$ in (15) by the characteristic function of the noise distribution. Nevertheless, some additional work would be needed to adapt the lower bounds
2. Only an upper bound of the tail distribution of the noise is known. For instance, the noise is only assumed to be sub-Gaussian with a bounded sub-Gaussian norm. In that situation, one cannot rely anymore on its characteristic function. Nevertheless, one could adapt some signal detection tests [2] to build "infimum test" [19, 38] such as those described in the Introduction. From rough calculations, it seems that the corresponding test would achieve the optimal separation distances up to polylogarithmic multiplicative terms. It remains an open problem to understand whether this polylog loss is intrinsic or not.

5.2. *Approximate sparsity.* Fix some $r \in (0, 2)$ and define the (pseudo)-norm $\|\theta\|_r = (\sum_i \theta_i^r)^{1/r}$ of θ. Instead of estimating or testing the value of the exact sparsity $\|\theta\|_0$, one may try to evaluate an approximate sparsity. Thus amount to estimating $\|\theta\|_r$ or testing whether $\|\theta\|_r$ is less than some given value $B > 0$. If the case $r = 2$ have been thoroughly investigated (see [15] and references therein), the literature on smaller r is scarcer. Cai and Low [11] have carefully considered the case $r = 1$ whereas Lepski et al. [32] provide some minimax rates (up to polylogarithmic multiplicative terms) for more general $r \in [1, 2]$. As a surrogate for the sparsity, the case $r \in (0, 1)$ is more relevant and it would be of interest to pinpoint sharp optimal rates of estimation and testing.

5.3. *Other models.* The same general roadmap of first deriving optimal separation rates for a single test and then rephrasing parameter estimation as a multiple testing problem can be adapted in other discrete functional estimation problems, including rank estimation in matrix regression and matrix completion models, smoothness estimation in the density framework, number of clusters estimation in model-based clustering, etc. A prominent example is sparsity estimation in the high-dimensional linear regression model. Let $Y \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n \times p}$ be such that

$$Y = \mathbf{X}\theta + \varepsilon,$$

where the parameter $\theta \in \mathbb{R}^p$ is unknown and $\varepsilon = (\varepsilon_i)$ is made of centered independent normal distributions with variance σ^2. In the specific case where $n = p$ and $\mathbf{X}$ is the identity matrix, it is equivalent to Gaussian vector model (1). Estimation of θ under sparsity assumptions has received a lot of attention in the last decade [4]. When the entries of $\mathbf{X}$ are independently sampled according to the standard normal distribution, the minimax separation distances for the detection problem has been derived in [1, 24]. For the purpose of building adaptive confidence intervals, Nickl and van de Geer [38] have introduced and analyzed sparsity testing procedures. However, the optimal separation distances for the sparsity testing problem remain unknown (except in some specific regimes). Further work is therefore needed to establish the minimax separation distances and to construct adaptive sparsity tests and sparsity estimators. This setting is more challenging than the one considered in this paper as the high-dimensional linear model also includes difficulties related to inverse problems.

Acknowledgements. The authors thank Christophe Giraud and anonymous referees for careful reading and insightful suggestions on the presentation of the results.

SUPPLEMENTARY MATERIAL

Supplement to “Adaptive estimation of the sparsity in the Gaussian vector model” (DOI: 10.1214/17-AOS1680SUPP; .pdf). Proofs of the results.

REFERENCES

[1] ARIAS-CASTRO, E., CANDÈS, E. J. and PLAN, Y. (2011). Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism. *Ann. Statist.* **39** 2533–2556. MR2906877

[2] BARAUD, Y. (2002). Non-asymptotic minimax rates of testing in signal detection. *Bernoulli* **8** 577–606. MR1935648

[3] BARAUD, Y., HUET, S. and LAURENT, B. (2005). Testing convex hypotheses on the mean of a Gaussian vector. Application to testing qualitative hypotheses on a regression function. *Ann. Statist.* **33** 214–257. MR2157802

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

[5] CAI, T. T. and GUO, Z. (2016). Accuracy assessment for high-dimensional linear regression. Preprint. Available at arXiv:1603.03474.

[6] CAI, T. T. and GUO, Z. (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *Ann. Statist.* **45** 615–646. MR3650395

[7] CAI, T. T. and JIN, J. (2010). Optimal rates of convergence for estimating the null density and proportion of nonnull effects in large-scale multiple testing. *Ann. Statist.* **38** 100–145. MR2589318

[8] CAI, T. T., JIN, J. and LOW, M. G. (2007). Estimation and confidence sets for sparse normal mixtures. *Ann. Statist.* **35** 2421–2449. MR2382653

[9] CAI, T. T. and LOW, M. G. (2004). An adaptation theory for nonparametric confidence intervals. *Ann. Statist.* **32** 1805–1840. MR2102494

[10] CAI, T. T. and LOW, M. G. (2006). Adaptive confidence balls. *Ann. Statist.* **34** 202–228. MR2275240

[11] CAI, T. T. and LOW, M. G. (2011). Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional. *Ann. Statist.* **39** 1012–1041. MR2816346

[12] CARPENTIER, A. (2015). Testing the regularity of a smooth signal. *Bernoulli* **21** 465–488.

[13] CARPENTIER, A. and VERZELEN, N. (2019). Supplement to "Adaptive estimation of the sparsity in the Gaussian vector model." DOI:10.1214/17-AOS1680SUPP.

[14] CELISSE, A. and ROBIN, S. (2010). A cross-validation based estimation of the proportion of true null hypotheses. *J. Statist. Plann. Inference* **140** 3132–3147.

[15] COLLIER, O., COMMINGES, L. and TSYBAKOV, A. B. (2017). Minimax estimation of linear and quadratic functionals on sparsity classes. *Ann. Statist.* **45** 923–958. MR3662444

[16] COMMINGES, L. and DALALYAN, A. S. (2013). Minimax testing of a composite null hypothesis defined via a quadratic functional in the model of regression. *Electron. J. Stat.* **7** 146–190. MR3020417

[17] DONOHO, D. and JIN, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32** 962–994.

[18] FROMONT, M., LERASLE, M. and REYNAUD-BOURET, P. (2016). Family-wise separation rates for multiple testing. *Ann. Statist.* **44** 2533–2563. MR3576553

[19] GAYRAUD, G. and POUET, C. (2005). Adaptive minimax testing in the discrete regression scheme. *Probab. Theory Related Fields* **133** 531–558.

[20] GINÉ, E. and NICKL, R. (2015). *Mathematical Foundations of Infinite-Dimensional Statistical Models* **40**. Cambridge Univ. Press, Cambridge.

[21] HÄRDLE, W., KERKYACHARIAN, G., PICARD, D. and TSYBAKOV, A. (2012). *Wavelets, Approximation, and Statistical Applications* **129**. Springer, New York. MR1618204

[22] HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning*. Springer, New York.

[23] HOFFMANN, M. and NICKL, R. (2011). On adaptive inference and confidence bands. *Ann. Statist.* **39** 2383–2409. MR2906872

[24] INGSTER, Y., TSYBAKOV, A. and VERZELEN, N. (2010). Detection boundary in sparse regression. *Electron. J. Stat.* **4** 1476–1526.

[25] INGSTER, YU. and SUSLINA, I. A. (2003). *Nonparametric Goodness-of-Fit Testing Under Gaussian Models. Lecture Notes in Statistics* **169**. Springer, New York. MR1991446

[26] JIN, J. (2008). Proportion of non-zero normal means: Universal oracle equivalences and uniformly consistent estimators. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 461–493.

[27] JIN, J. and TONY CAI, T. (2007). Estimating the null and the proportional of nonnull effects in large-scale multiple comparisons. *J. Amer. Statist. Assoc.* **102** 495–506.

[28] JUDITSKY, A. and NEMIROVSKI, A. (2002). On nonparametric tests of positivity/monotonicity/convexity. *Ann. Statist.* **30** 498–527. MR1902897

[29] KALAI, A. T., MOITRA, A. and VALIANT, G. (2012). Disentangling Gaussians. *Commun. ACM* **55** 113–120.

[30] KESHAVAN, R. H., MONTANARI, A. and OH, S. (2010). Matrix completion from a few entries. *IEEE Trans. Inform. Theory* **56** 2980–2998. MR2683452

[31] LANGAAS, M., LINDQVIST, B. H. and FERKINGSTAD, E. (2005). Estimating the proportion of true null hypotheses, with application to DNA microarray data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 555–572.

[32] LEPSKI, O., NEMIROVSKI, A. and SPOKOINY, V. (1999). On estimation of the L_r norm of a regression function. *Probab. Theory Related Fields* **113** 221–253.

[33] LI, J. and SIEGMUND, D. (2015). Higher criticism: p-values and criticism. *Ann. Statist.* **43** 1323–1350. MR3346705

[34] MAHER, B. (2008). Personal genomes: The case of the missing heritability. *Nature* **456** 18–21.

[35] MASSART, P. (2007). *Concentration Inequalities and Model Selection. Lecture Notes in Math.* **1896**. Springer, Berlin. MR2319879

[36] MEINSHAUSEN, N. and RICE, J. (2006). Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *Ann. Statist.* **34** 373–393. MR2275246

[37] MOSCOVICH, A., NADLER, B. and SPIEGELMAN, C. (2016). On the exact Berk–Jones statistics and their p-value calculation. *Electron. J. Stat.* **10** 2329–2354. MR3544289

[38] NICKL, R. and VAN DE GEER, S. (2013). Confidence sets in sparse regression. *Ann. Statist.* **41** 2852–2876. MR3161450

[39] PATRA, R. K. and SEN, B. (2015). Estimation of a two-component mixture model with applications to multiple testing. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 869–893. MR3534354

[40] STOREY, J. D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 479–498.

[41] TORO, R. et al. (2015). Genomic architecture of human neuroanatomical diversity. *Mol. Psychiatry* **20** 1011–1016.

[42] VERZELEN, N. (2012). Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electron. J. Stat.* **6** 38–90.

INSTITUT FÜR MATHEMATISCHE STOCHASTIK
OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG
UNIVERSITÄTSPLATZ 2
39106, MAGDEBURG
GERMANY
E-MAIL: alexandra.carpentier@ovgu.de

INRA, MONTPELLIER SUPAGRO, MISTEA
UNIVERSITÉ DE MONTPELLIER
MONTPELLIER
FRANCE
E-MAIL: nicolas.verzelen@inra.fr