The Annals of Statistics
2017, Vol. 45, No. 5, 1920–1950
DOI: 10.1214/16-AOS1513

DETECTION AND FEATURE SELECTION IN SPARSE MIXTURE MODELS[1]

BY NICOLAS VERZELEN AND ERY ARIAS-CASTRO

INRA and University of California, San Diego

We consider Gaussian mixture models in high dimensions, focusing on the twin tasks of detection and feature selection. Under sparsity assumptions on the difference in means, we derive minimax rates for the problems of testing and of variable selection. We find these rates to depend crucially on the knowledge of the covariance matrices and on whether the mixture is symmetric or not. We establish the performance of various procedures, including the top sparse eigenvalue of the sample covariance matrix (popular in the context of Sparse PCA), as well as new tests inspired by the normality tests of Malkovich and Afifi [*J. Amer. Statist. Assoc.* **68** (1973) 176–179].

1. Introduction. Variable (aka feature) selection is a fundamental aspect of regression analysis and classification, particularly in high-dimensional settings where the number of variables exceeds the number of observations. The corresponding literature is vast, from the early proposals based on penalizing the number of variables (i.e., the ℓ_0 norm) [Akaike (1974), Mallows (1973), Schwarz (1978)], to the more recent variants based on convex relaxations (e.g., the ℓ_1 norm) [Tibshirani (1996), Candes and Tao (2007), Zhu and Hastie (2004), Chen, Donoho and Saunders (1998)] and a wide array of alternative approaches, including nonconvex relaxations [Fan and Peng (2004)], greedy methods [Mallat and Zhang (1993), Tropp (2004)] and methods based on multiple testing [Ji and Jin (2012), Jin (2009), Ingster, Pouet and Tsybakov (2009), Donoho and Jin (2009)]. We refer the reader to Massart (2007) and Hastie, Tibshirani and Friedman (2009), Chapters 3, 7, 18, for additional pointers.

In contrast, variable selection in the context of clustering is at a comparatively infant stage of development, even though clustering is routinely used in high-dimensional settings. Also, according to Hastie, Tibshirani and Friedman (2009):

> Specifying an appropriate dissimilarity measure is far more important in obtaining success with clustering than choice of clustering algorithm.

And, of course, choosing a dissimilarity measure is intimately related to weighting the variables, or combinations of variables, according to their importance in

Received May 2014; revised December 2015.

[1] Supported in part by the US Office of Naval Research (N00014-13-1-0257) and the French Agence Nationale de la Recherche (ANR 2011 BS01 010 01 projet Calibration).

MSC2010 subject classifications. 62H30, 62H15.

Key words and phrases. Gaussian mixture models, detection of mixtures, feature selection for mixtures, sparse mixture models, the sparse eigenvalue problem, projection tests based on moments.

clustering the observations. The literature on variable selection for clustering is indeed much more recent, scarce and ad hoc. Chang (1983) concludes empirically that performing principal component analysis as a preprocessing step to clustering a Gaussian mixture is not necessarily useful. Raftery and Dean (2006) and Maugis and Michel (2011) propose a model selection approach, while penalized methods are suggested in Pan and Shen (2007), Xie, Pan and Shen (2008), Wang and Zhu (2008), Friedman and Meulman (2004), Witten and Tibshirani (2010).

We focus here on the emblematic setting of a mixture of two Gaussians in high-dimensions. Working under the crucial assumption that the difference in means is sparse, we study the cousin problems of mixture detection (i.e., testing whether the difference in means is zero or not) and variable selection (i.e., estimating the support of the difference in means), both when the covariance matrix is known and when it is unknown. We obtain minimax lower bounds and propose a number of methods which are able to match these bounds.

1.1. *Detection problem.* The first problem that we consider is that of *detection* of mixing, specifically, we test the null hypothesis that there is only one component, versus the alternative hypothesis that there are two components, in a sample assumed to come from a Gaussian mixture model. We assume throughout that the group covariance matrices are identical, and we consider the case where it is known and the case where it is unknown. Formally, in the case where it is unknown, we observe $X_1, \dots, X_n \in \mathbb{R}^p$ and consider the general testing problem

$$\begin{gathered} H_0 : X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \mathbf{\Sigma}), \\ \text{for some } \mu \in \mathbb{R}^p \text{ and some } \mathbf{\Sigma} \in \mathbb{R}^{p \times p} \text{ p.s.d.;} \end{gathered} \tag{1}$$

versus

$$\begin{gathered} H_1 : X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \nu \mathcal{N}(\mu_0, \mathbf{\Sigma}) + (1-\nu)\mathcal{N}(\mu_1, \mathbf{\Sigma}), \\ \text{for some } \mathbf{\Sigma} \in \mathbb{R}^{p \times p} \text{ p.s.d., some } \mu_0 \neq \mu_1 \in \mathbb{R}^p \text{ and some } \nu \in (0, 1). \end{gathered} \tag{2}$$

(As usual, "p.s.d." stands for "positive semidefinite.") We are specifically interested in settings where the difference in means is sparse:

$$\Delta\mu := \mu_1 - \mu_0 \text{ is } s\text{-sparse,} \tag{3}$$

where $1 \leq s \leq p$ and ν belongs to $(0, 1)$. (We say that a vector is s-sparse if it has at most s nonzero entries.) In the sequel, we denote $\theta = (\nu, \mu_0, \mu_1, \mathbf{\Sigma})$ the set of parameters with the convention that under the null hypothesis $\mu_1 = \mu_0$, so that $\Delta\mu = 0$, and $\nu \in (0, 1)$ is arbitrary. We then write $\mathbb{P}_\theta$ for the probability distribution of $X_1, \dots, X_n$.

We note that the model (1)–(2) can be written as

$$\begin{gathered} X_i = \mu_0 + (1 - \eta_i)\Delta\mu + \mathbf{\Sigma}^{1/2} Z_i, \\ \text{where } \eta_1, \dots, \eta_n \overset{\text{i.i.d.}}{\sim} \text{Bern}(\nu) \text{ and independent of } Z_1, \dots, Z_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I}), \end{gathered} \tag{4}$$

where $\mathbf{I}$ denotes the identity matrix (here in dimension p), and the hypothesis testing problem then reads

$$H_0 : \Delta\mu = 0 \quad \text{versus} \quad H_{1,s}^{\nu} : \Delta\mu \neq 0 \text{ is } s\text{-sparse}. \tag{5}$$

For simplicity of exposition:

- We assume that the sparsity s is known. This is a rather mild assumption (at least in theory) as discussed in Section 5.
- We assume the parameter ν is unknown and bounded away from 0 and 1. When ν approaches 0 or 1, the problem becomes that of testing for contamination. Although the two settings are intimately related, treating both would burden the presentation.

We consider the testing problem (1) versus (2) in a high-dimensional large-sample context where all the parameters $(p, s, \Delta\mu, \mathbf{\Sigma})$ may depend on n. Unless specified otherwise, all the limits are taken when the sample size increases to infinity, $n \to \infty$. We adopt a minimax perspective, which consists of quantifying the performance of tests in the worst case sense.

As various testing problems are studied in this manuscript, the notion of minimax detection rates is first introduced in an abstract way. Consider $H_0 : \theta \in \Omega_0^n$ versus $H_1 : \theta \in \Omega_1^n$ based on a sample from a distribution belonging to some family $\{\mathbb{P}_\theta : \theta \in \Omega\}$ and define a nonnegative function R that satisfies $R(\theta) = 0$ for all $\theta \in \Omega_0^n$ and $R(\theta) > 0$ for all $\theta \in \Omega_1^n$. Henceforth, $R(\cdot)$ is called the signal-to-noise ratio. In our Gaussian mixture framework, think of $R(\theta)$ as some (pseudo-)norm of $\Delta\mu$. Given some number $r_n > 0$, define $\Omega_1^n(R, r_n) := \{\theta \in \Omega_1^n : R(\theta) \geq r_n\}$, the set of parameters in the alternative that are r_n-separated from the null hypothesis. Then the worst-case risk of a test ϕ for testing $\theta \in \Omega_0^n$ versus $\theta \in \Omega_1^n(R, r_n)$ is the sum of its probabilities of type I and type II errors, maximized over the null set Ω_0^n and alternative distributions $\mathbb{P}_\theta$ whose signal-to-noise ratio is larger than r_n, or in formula

$$\gamma\big(\phi; \Omega_0^n, \Omega_1^n(R, r_n)\big) := \sup_{\theta \in \Omega_0^n} \mathbb{P}_\theta(\phi = 1) + \sup_{\theta \in \Omega_1^n(R, r_n)} \mathbb{P}_\theta(\phi = 0).$$

The rationale behind the introduction of r_n is that in testing problems such as (1)–(2) some distributions in the alternative are arbitrarily close to the null hypothesis so that $\gamma(\phi; \Omega_0^n, \Omega_1^n; R, 0) = 1$ for any test ϕ. This is why the probability type II error is maximized over alternatives that are sufficiently separated from the null distribution, which is here quantified as $R(\theta) \geq r_n$. Then the minimax risk for this testing problem is defined as

$$\gamma^*\big(\Omega_0^n, \Omega_1^n(R, r_n)\big) := \inf_\phi \gamma\big(\phi; \Omega_0^n, \Omega_1^n(R, r_n)\big),$$

where the infimum is over all possible tests for H_0 versus H_1. Formally speaking, we consider a sequence of hypotheses indexed by the sample size n and, correspondingly, consider sequences of tests, also indexed by n. Understood as such,

$\liminf_{n\to\infty} \gamma^*[\Omega_0^n, \Omega_1^n(R, r_n)] = 1$ is equivalent to saying that, in the large-sample limit, no test does better than random guessing. When a sequence of tests ϕ_n satisfies $\gamma(\phi_n; \Omega_0^n, \Omega_1^n(R, r_n)) \to 0$, it is said to be asymptotically powerful. A real sequence r_n^* is said to be a *minimax separation rate* of H_0 versus H_1 if for any sequence r_n satisfying $r_n \ll r_n^*$, $\gamma^*[\Omega_0^n, \Omega_1^n(R, r_n)] \to 1$, while for any sequence r_n satisfying $r_n \gg r_n^*$, $\gamma^*[\Omega_0^n, \Omega_1^n(R, r_n)] \to 0$. As we shall see in concrete situations, the minimax separation rate r_n^* characterizes the minimal distance between the mixture means to enable reliable mixture detection. As is customary, we leave the dependency on n implicit in the sequel.

Contribution. We distinguish between the cases where $\mathbf{\Sigma}$ is known or unknown. The case where $\mathbf{\Sigma}$ is diagonal will play a special role, due to the fact that it combines well with the assumption that the mean difference vector $\Delta\mu$ is assumed sparse in the canonical basis of $\mathbb{R}^p$. We also distinguish between the symmetric setting, where $\nu = 1/2$, and the asymmetric setting, where $\nu \neq 1/2$.

For each situation, we introduce an appropriate signal-to-noise ratio function R and derive the minimax detection rate with an explicit dependency in the sample size n, the ambient dimension p, the sparsity s of the difference in means $\Delta\mu$, the mixture weight ν. We also propose some tests—some of them new—which are shown to be minimax rate optimal.

- When the covariance matrix $\mathbf{\Sigma}$ is known, the test based on the top eigenvalue of the normalized sample covariance matrix is competitive when s is relatively large; while the test based on the top sparse (in the eigen-basis of $\mathbf{\Sigma}$) eigenvalue of the normalized sample covariance matrix is competitive when s is relatively small.
- When the covariance matrix $\mathbf{\Sigma}$ is unknown, we propose some new projection tests based on moments à la Malkovich and Afifi (1973), which are shown to achieve the minimax rate. The detection rates that we obtain for the projection skewness and kurtosis statistics proposed in Malkovich and Afifi (1973) are suboptimal.

Our results are summarized in Tables 1 and 2. Note that when $\mathbf{\Sigma}$ is known, the signal-to-noise ratio is measured in terms of the Mahalanobis distance of $\Delta\mu$ from 0 (see Table 1) while a different measure is used when $\mathbf{\Sigma}$ is unknown; see Table 2. We show that using the Mahalanobis distance in the latter setting leads to exponential minimax bounds. This is detailed in Section 3.1.4.

1.2. *Variable selection.* The second problem that we consider is that of *variable selection*, where the goal is to estimate the support of $\Delta\mu$ in (3) under the mixture model (2). The support of a vector $v = (v_j)$ is $\{j : v_j \neq 0\}$. A problem of particular interest when s is small compared to p—meaning $s = o(p)$—is that of estimating the support of $\Delta\mu$, which corresponds to the variables that are responsible for separating the population into two groups. This is what we mean by

TABLE 1

Minimax detection rates and near-optimal tests as a function of s when $\mathbf{\Sigma}$ is known and $p \geq n$. The minimax detection rates are expressed in terms of the signal-to-noise ratio $R_0 = \Delta\mu^\top \mathbf{\Sigma}^{-1}\Delta\mu$. Here, γ denotes any arbitrary constant in $(0, 1)$

Sparsity regimes	**Minimax detection rates**	**Near-optimal test**
$s \leq \frac{n}{\log(ep/n)}$	$[\frac{s\log(ep/s)}{n}]^{1/2}$	Top sparse eigenvalue (15)
$\frac{n}{\log(ep/n)} \leq s \leq (np)^{\gamma/2}$	$\frac{s\log(ep/s)}{n}$	Top sparse eigenvalue (15)
$s \geq \sqrt{np}$	$\sqrt{p/n}$	Top eigenvalue (14)

variable selection, and in a setting where the hypothesis testing problem is parameterized by the sample size n, we say that a certain estimator $\hat{J}_n$ is consistent for $J := \{j : \Delta\mu_j \neq 0\}$ (which may depend on n) if

$$\frac{|\hat{J}_n \triangle J|}{|J|} \to 0, \qquad n \to \infty. \tag{6}$$

The dependency on n will often be left implicit.

For the problem of variable selection, we work under the assumption that the effective dynamic range of $\Delta\mu$ and the $2s$-sparse Riesz constant of $\mathbf{\Sigma}$ are both bounded. We define the effective dynamic range of a set of real numbers $\{x_j\}$ (possibly organized as a vector) as $\sup_j |x_j| / \inf_{j\in J} |x_j|$, assuming $J := \{j : x_j \neq 0\} \neq \varnothing$. Given a $p \times p$ positive semidefinite matrix $\mathbf{\Sigma} \neq 0$ and an integer $1 \leq s \leq p$, we define the largest s-sparse eigenvalue of $\mathbf{\Sigma}$ as $\lambda_s^{\max}(\mathbf{\Sigma}) = \max_u u^\top \mathbf{\Sigma} u$, where the maximum is over s-sparse unit vectors $u \in \mathbb{R}^p$. The smallest s-sparse eigenvalue of $\mathbf{\Sigma}$ is defined analogously, replacing "max" with "min". The s-sparse Riesz constant of $\mathbf{\Sigma}$ is simply $\lambda_s^{\max}(\mathbf{\Sigma})/\lambda_s^{\min}(\mathbf{\Sigma})$. Equivalently, it is the supremum of $u^\top \mathbf{\Sigma} u / v^\top \mathbf{\Sigma} v$ over all pairs of unit s-sparse vectors u and v.

Contribution. Since in each case our testing procedure in the sparse setting $[s = o(p)]$ is based on maximizing some form of moment over direction vectors

TABLE 2

Minimax detection rates and near-optimal tests when $\mathbf{\Sigma}$ is unknown. The minimax detection rates are expressed in terms of the signal-to-noise $R_1 = \|\Delta\mu\|^4/\Delta\mu^\top \mathbf{\Sigma}\Delta\mu$. In this summary, we assume that $s\log(ep/s) = o(n)$. If this is not the case and $\mathbf{\Sigma}$ is unknown, our lower bounds show that the problem is extremely hard

	Minimax detection rates	**Near-optimal test**
Symmetric ($\nu = 1/2$)	$[\frac{s}{n}\log(\frac{ep}{s})]^{1/4}$	Projection 1st moment (27)
Asymmetric ($\nu \neq 1/2$)	$[\frac{s}{n}\log(\frac{ep}{s})]^{1/3}$	Projection 2nd signed moment (38)

which are sparse (in some way made explicit later on), it is natural to use the support of the maximizing direction as an estimator for the support of $\Delta\mu$:

- When $\mathbf{\Sigma}$ is known, we show that this estimator is indeed consistent in the sense of (6) at (essentially) the minimax rate for detection.
- When $\mathbf{\Sigma}$ is unknown, surprisingly, this estimator may be suboptimal. This leads us to propose nontrivial variants in (31) (symmetric setting) and (40) (asymmetric setting). We are able to show that the support estimator (31) is consistent at (essentially) the minimax rate for detection.

1.3. *Consequences for clustering.* We see the problems of detection and variable selection as complementary to the problem of clustering. We could imagine a work flow where detection is performed first, then variable selection if the test is significant, and then clustering based on the selected variables. To keep this paper concise, we do not provide here an analysis of these multi-step clustering algorithms. See Azizyan, Singh and Wasserman (2013, 2015), Jin and Wang (2014) for recent results in this direction.

The motivation for performing detection and variable selection first is meaningful because these can be successfully accomplished with a much smaller separation between the components than clustering. Indeed, consider a Gaussian mixture of the form $\frac{1}{2}\mathcal{N}(0, \mathbf{\Sigma}) + \frac{1}{2}\mathcal{N}(\Delta\mu, \mathbf{\Sigma})$. Even if $\mathbf{\Sigma}$ and $\Delta\mu$ are known—in which case the best clustering method is the rule $\{x^\top \mathbf{\Sigma}^{-1}\Delta\mu > \Delta\mu^\top \mathbf{\Sigma}^{-1}\Delta\mu/2\}$—the expected clustering error is at least $\mathbb{P}(\mathcal{N}(0,1) > \|\mathbf{\Sigma}^{-1/2}\Delta\mu\|/2)$, which converges to 0 only if $\|\mathbf{\Sigma}^{-1/2}\Delta\mu\| \to \infty$.

1.4. *Methodology, computational issues and mathematical technique.*

Methodology. Most of the tests that we propose are novel. While the first test in Table 1 is very natural, the second test is new. It is a close cousin of the sparse eigenvalue (19), considered in the sparse PCA literature (see Section 1.5). However, the latter appears suboptimal so that our variant brings a nontrivial improvement. The tests in Table 2 are new. They compete with the projection kurtosis (25) and skewness (36) that we adapted from the normality tests of Malkovich and Afifi (1973). The motivation for introducing new tests is our inability to prove that these kurtosis and skewness tests achieve the minimax rate. This is because they are based on higher-order moments, which we found harder to control under the null.

Computational issues. We emphasize that except for the top eigenvalue, the other test statistics in Tables 1 and 2 are very hard to compute even for moderate p. We conjecture that no testing procedure with polynomial computational complexity is able to achieve the minimax rates of detection. When the covariance $\mathbf{\Sigma}$ is known, our testing problem shares many similarities with the sparse PCA detection problem for which a gap between optimal and computationally amenable procedures has been established [Berthet and Rigollet (2013a)].

TABLE 3
Detection rates achieved by some computationally feasible tests when $\mathbf{\Sigma}$ is known. See Section 4 for precise statements and assumptions. Compared to Table 1, the rates are at most $\sqrt{s}$ slower than the optimal rates

Sparsity regimes	**Detection rates**	**Test**
$s \leq \sqrt{p/\log(p)}$	$[s^2 \frac{\log(p)}{n}]^{1/2}$	Maximal canonical variance (47)
$s \geq \sqrt{p/\log(p)}$	$\sqrt{p/n}$	Top eigenvalue (14)

Another contribution of this paper is to propose computationally feasible tests:

- We study coordinatewise methods based on moments.
- We study existing convex relaxations to the sparse eigenvalue problem.

See Tables 3 and 4. The tests in Table 4 are new and are the coordinatewise equivalents of the tests appearing in Table 2.

A note on the mathematical technique. Regarding the technical arguments, the derivation of the information lower bounds for the detection problem is typical: we reduce the set of null hypotheses to the standard normal distribution and put a prior on the set of alternatives, and then bound the variance of the resulting likelihood ratio under the null. The latter amounts to bounding the chi-squared divergence between the reduced null and alternative distributions; see Tsybakov (2009), Theorem 2.2. That said, in the details, the calculations are both complicated and tedious. The test statistics that we study are based on sample moments of Gaussian random variables of degree up to 4. To control these statistics under the null, we use a combination of chaining à la Dudley [van der Vaart and Wellner (1996)] and concentration bounds that we derive based on approximations of Gaussian random variables by sums of Rademacher random variables together with concentration bounds for these obtained by Boucheron et al. (2005).

TABLE 4
Detection rates achieved by some computationally feasible tests when $\mathbf{\Sigma}$ is unknown. See Section 4 for precise statements and assumptions. Compared to Table 2, the rates are respectively at most $s^{1/4}$ and $s^{1/3}$ slower than the optimal rates. In this summary, we assume that $\log(p) = o(n)$

	Detection rates	**Test**
Symmetric ($\nu = 1/2$)	$[\frac{s^4}{n}\log(\frac{ep}{s})]^{1/4}$	Coordinatewise 1st moment (50)
Asymmetric ($\nu \neq 1/2$)	$[\frac{s^3}{n}\log(\frac{ep}{s})]^{1/3}$	Coordinatewise 2nd signed moment (51)

1.5. *Closely related literature.* We already cited a number of publications proposing various methods for variable selection in the context of high-dimensional clustering. None of these papers offers any real theoretical insights on the difficulty of this problem. In fact, very few mathematical results are available in this area.

Most of them are on the estimation of Gaussian mixture parameters. Recent papers in this line of work include [Belkin and Sinha (2010), Kalai, Moitra and Valiant (2012), Hsu and Kakade (2013), Brubaker and Vempala (2008)], and references therein. These papers focus on designing polynomial time algorithms that work when there is sufficient parameter identifiability, which is often not optimized. An exception to that is Chaudhuri, Dasgupta and Vattani (1999), where a multistage variant of k-means is analyzed in the canonical setting of a symmetric mixture of two Gaussians with identity covariance, and showed to match an information-theoretic bound when the centers are at a distance exceeding 1. We note that there is no assumption of sparsity made in this literature.

Related to our proposal of coordinatewise methods presented in Section 4.1, Chan and Hall (2010) test each coordinate for unimodality and prove variable selection consistency in a nonparametric setting. Similar in spirit, Jin and Wang (2014) propose[2] the selection of features based on coordinatewise Kolmogorov–Smirnov goodness-of-fit testing. Their setting is slightly different from ours as the number of mixtures in their paper is allowed to be larger than 2 but the covariance matrix is restricted to be diagonal and the distributions are supposed to be asymmetric. Nevertheless, when specialized to a common framework (two components, diagonal unknown covariance matrix, $\nu \neq 1/2$), their detection rates and ours are the same. Azizyan, Singh and Wasserman (2013) consider the task of clustering a sparse symmetric mixture of two Gaussians in high-dimensions with identity covariance matrix. They prove a minimax lower bound for some clustering error, but do not exhibit any method that matches that lower bound. Instead, they propose a coordinatewise approach which is almost identical to one of the methods considered by Amini and Wainwright (2009) (see below) and is very similar to what we do in Section 4.1. This work is closely related to what we obtain in Section 2 (specialized to $\mathbf{\Sigma} = \mathbf{I}$) and in Section 4.1. The same authors propose[2] in Azizyan, Singh and Wasserman (2015) to first learn the parameters of the Gaussian mixture model using Hardt and Price (2015) and then apply sparse linear discriminant analysis. Their results are not directly comparable to ours as they assume that $\mathbf{\Sigma}^{-1}\Delta\mu$ (instead of $\Delta\mu$) is sparse.

Close to our work is the recent literature on sparse principal component analysis, in view of the following expression for the covariance matrix:

$$\mathrm{Cov}(X) = \nu(1-\nu)\Delta\mu\Delta\mu^\top + \mathbf{\Sigma}. \tag{7}$$

[2] This work appeared after the initial version of the present paper was made publicly available.

The difference is that, in this line of work, $X_1, \dots, X_n$ are i.i.d. centered normal with covariance matrix of the form (7). We note that most of the work considers the case where $\mathbf{\Sigma}$ is known and isotropic. The most closely related is the work of Berthet and Rigollet (2013a) on testing for a leading sparse principal direction. From them, we drew the idea of using the SDP relaxation of d'Aspremont et al. (2007) for the sparse eigenvalue problem; see Section 4.2. Also closely related is Amini and Wainwright (2009), where the authors tackle the problem of variable selection in the same context. They propose a coordinatewise approach which selects the coordinates corresponding to the top s largest variances, identical to a preprocessing step in Johnstone and Lu (2009). They also study the SDP method of d'Aspremont et al. (2007), but under very strong constraints—in particular, they assume that $s = O(\log p)$. The estimation of the leading principal component(s), which concerns, for example, Johnstone and Lu (2009), Cai, Ma and Wu (2013, 2015), Birnbaum et al. (2013), Vu and Lei (2012, 2013) is also closely related.

REMARK. We note that most of the references in the sparse PCA literature assume that $\mathbf{\Sigma} = \mathbf{I}$ in (7). This can easily be extended to the case of a diagonal covariance matrix, which is also an important case in our work. That said, it is important to realize that, even when more general covariance structures are considered—as in Vu and Lei (2012, 2013)—the parallel with our work is essentially restricted to the case where the covariance matrix is known. Indeed, once the covariance matrix is unknown, looking for unusually large eigenvalues in the (sample) covariance matrix becomes meaningless in the context of clustering.

1.6. *Organization and notation.*

The paper is organized as follows. In Section 2, we consider the case where the covariance is known. In Section 3, we treat the case where the covariance is unknown, including the special case where it is known to be diagonal. In Section 4, we suggest and study coordinatewise methods and some relaxations. We then compare some of them in small numerical experiments. We discuss extensions and important issues in Section 5, such as the case of unknown sparsity, the case of mixture models with different covariances, the case of mixtures with more than two components, and more. The proofs are gathered in the online supplement [Verzelen and Arias-Castro (2016)].

Notation. For an integer p, $[p] = \{1, \dots, p\}$. For a matrix $A \in \mathbb{R}^{p \times p}$ and a subset $S \subset [p]$, A_S denotes the principal submatrix of A indexed by S. For a finite set S, $|S|$ denotes its size. For two vectors $u = (u_j)$ and $v = (v_j)$ in a Euclidean space, $\|u\|$ denotes the Euclidean norm, $\langle u, v \rangle$ the inner product, $\|u\|_\infty = \max_j |u_j|$ the supnorm, and $\|u\|_0$ the cardinality of the support $\operatorname{supp}(u) := \{j : u_j \neq 0\}$. Finally, C, C_1, C_2, etc. will denote positive constants that may change with each appearance.

2. Known covariance matrix.

In this section, the covariance $\mathbf{\Sigma}$ is assumed to be known. The minimax detection rates are expressed with respect to the Mahalanobis distance

$$R_0(\theta) = \Delta\mu^\top \mathbf{\Sigma}^{-1} \Delta\mu.$$

2.1. *Minimax lower bound.*

Fix a mixing weight $\nu \in (0, 1)$ and a sparsity s, and consider

$$\Omega_0(\nu) = \{\theta = (\nu, \mu, \mu, \mathbf{\Sigma}), \mu \in \mathbb{R}^p, \mathbf{\Sigma} \text{ p.s.d.}\}, \tag{8}$$

and, for $r_n > 0$ and for a signal-to-noise ratio function R,

$$\begin{aligned} \Omega_1(\nu, R, r_n) := \{\theta = (\nu, \mu_0, \mu_1, \mathbf{\Sigma}) :& \\ \mu_0, \mu_1 \in \mathbb{R}^p \text{ satisfying (3)}, \mathbf{\Sigma} \text{ p.s.d.}, R(\theta) \geq r_n\}&, \end{aligned} \tag{9}$$

where we leave implicit the dependency of $\Omega_1(\nu, R_0, r_n)$ on s. As the tests considered in this section use the knowledge the covariance matrix $\mathbf{\Sigma}$, we also consider for any covariance $\mathbf{\Sigma}$,

$$\Omega_0(\nu, \mathbf{\Sigma}) = \{\theta = (\nu, \mu, \mu, \mathbf{\Sigma}), \mu \in \mathbb{R}^p\}, \tag{10}$$

and, fixing a mixing weight $\nu \in (0, 1)$, a sparsity s and $r_n > 0$, consider

$$\begin{aligned} \Omega_1(\nu, \mathbf{\Sigma}, R, r_n) := \{\theta = (\nu, \mu_0, \mu_1, \mathbf{\Sigma}) :& \\ \mu_0, \mu_1 \in \mathbb{R}^p \text{ satisfying (3)}, R(\theta) \geq r_n\}&. \end{aligned} \tag{11}$$

Then the minimax detection risk with known variance is defined by

$$\gamma^*_{\text{known}}(\Omega_0(\nu), \Omega_1(\nu, R, r_n)) = \sup_{\mathbf{\Sigma}} \inf_{\phi} \gamma(\phi; \Omega_0(\nu, \mathbf{\Sigma}), \Omega_1(\nu, \mathbf{\Sigma}, R, r_n)).$$

In order to emphasize the role of sparsity, we distinguish the sparse and nonsparse settings, corresponding to $s = p$ and $s = o(p)$, respectively.

PROPOSITION 1. *Consider testing* (10) *versus* (11). *For any fixed* $\nu \in (0, 1)$, *we have* $\liminf \gamma^*_{\text{known}}(\Omega_0(\nu), \Omega_1(\nu, R_0, r_n)) = 1$ *in the following two cases*:

- Nonsparse setting. *Assume* $s = p \to \infty$ *and*

$$r_n \ll \sqrt{p/n}.$$

- Sparse setting. *Assume* $p/s \to \infty$ *and*

$$r_n \ll \sqrt{p/n} \tag{12}$$

and

$$\limsup \frac{r_n}{\sqrt{\frac{s}{n}\log(\frac{ep}{s}) \vee \frac{s}{n}\log(1 + \sqrt{\frac{epn}{s^2}})}} < 1. \tag{13}$$

REMARK. As usual, for the derivation of a minimax lower bound it is sufficient to provide a lower bound on the risk for testing subclasses of $\Omega_0(\nu, \mathbf{\Sigma})$ and $\Omega_1(\nu, \mathbf{\Sigma}, R_0, r_n)$. In fact, we reduce the problem to testing $\theta \in \widetilde{\Omega}_0 := \{\theta = (\nu, 0, 0, \mathbf{I})\}$ against

$$\theta \in \widetilde{\Omega}_1(\nu, R_0, r_n) := \bigl\{(\nu, -(1-\nu)\mu, \nu\mu, \mathbf{I}), \mu \text{ is } s\text{-sparse}, R_0(\theta) \geq r_n\bigr\}.$$

2.2. *Methodology based on* (*sparse*) *principal component analysis.* We now turn to designing tests that are asymptotically powerful just above the lower bound given in Proposition 1. We note that the performance bounds for the tests based on (14) and (15) in Propositions 2 and 3 apply to a general (known) covariance matrix.

Our methodology is based on the expression for the covariance matrix of X displayed in (7). We standardize the observations to have identity covariance under the null, thus working with $X_\ddagger = \mathbf{\Sigma}^{-1/2}X$, which satisfies

$$\mathbf{\Sigma}_\ddagger := \operatorname{Cov}(X_\ddagger) = \mathbf{\Sigma}^{-1/2}\operatorname{Cov}(X)\mathbf{\Sigma}^{-1/2} = \nu(1-\nu)\Delta\mu_\ddagger\Delta\mu_\ddagger^\top + \mathbf{I},$$

where $\Delta\mu_\ddagger := \mathbf{\Sigma}^{-1/2}\Delta\mu$. Thus $\operatorname{Cov}(X_\ddagger)$ is a rank-one perturbation of the identity matrix under the alternative. Since $\operatorname{Cov}(X_\ddagger)$ is unknown, our inference is based on the sample equivalent, which is $\hat{\mathbf{\Sigma}}_\ddagger := \mathbf{\Sigma}^{-1/2}\hat{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1/2}$, where

$$\hat{\mathbf{\Sigma}} := \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})^\top, \qquad \bar{X} := \frac{1}{n}\sum_{i=1}^n X_i,$$

are the sample covariance matrix and sample mean, respectively.

- When $\Delta\mu$ is not sparse ($s = p$), this leads us to consider the top eigenvalue of $\hat{\mathbf{\Sigma}}_\ddagger$, namely

$$\hat{\lambda}_{\mathbf{\Sigma}}^{\max} := \max_{\|u\|=1} u^\top \hat{\mathbf{\Sigma}}_\ddagger u. \tag{14}$$

  We note that the maximizer of (14) is the first principal direction of the standardized observations, and that $\hat{\lambda}_{\mathbf{\Sigma}}^{\max}$ is the variance along that direction. As we shall see, this test is also competitive when $\Delta\mu$ is moderately sparse.
- When $\Delta\mu$ is s-sparse, we restrict the maximization over the set of vectors that are s-sparse in some appropriate basis. To guide our choice, we notice that $\Delta\mu_\ddagger$ is a top eigenvector for $\mathbf{\Sigma}_\ddagger$, and $\mathbf{\Sigma}^{1/2}\Delta\mu_\ddagger = \Delta\mu$ is s-sparse. This leads us to the following form of s-sparse (top) eigenvalue:

$$\hat{\lambda}_{s,\mathbf{\Sigma}}^{\max} := \max_{\|u\|=1, \|\mathbf{\Sigma}^{1/2}u\|_0 \leq s} u^\top \hat{\mathbf{\Sigma}}_\ddagger u. \tag{15}$$

  We note that the maximizer of (15) is the first s-sparse (after standardization) principal direction of the standardized observations, and that $\hat{\lambda}_{s,\mathbf{\Sigma}}^{\max}$ is the variance along that direction.

REMARK. With the notable exception of (14), all the statistics studied in Sections 2 and 3 are difficult to compute, which effectively makes them useless in practical settings, which are often high-dimensional. For this reason, we leave implicit the critical values of the corresponding tests. The interested reader may obtain their expression by inspecting the proofs of the corresponding propositions.

The following performance bound says, roughly, that the test based on (14) is reliable when (12) does not hold.

PROPOSITION 2. *Consider testing* (10) *versus* (11) *with* $\mathbf{\Sigma}$ *known*, $\nu \in (0, 1)$ *fixed*, $s \le p$ *and* $p \wedge n \to \infty$. *Let* T *denote the statistic* (14). *The test* $\phi = \{T \ge 1 + p/n + 12\sqrt{p/n}\}$ *is asymptotically powerful*, *meaning* $\gamma(\phi; \Omega_0(\nu, \mathbf{\Sigma}), \Omega_1(\nu, \mathbf{\Sigma}, R_0, r_n)) \to 0$, *if the minimum Mahalanobis distance* r_n *satisfies*

$$\liminf r_n \nu(1-\nu)\sqrt{\frac{n}{p}} > C, \tag{16}$$

where C *is a universal constant.*

In view of Proposition 1, the above test is adaptive to the mixing weight ν as long as it is fixed.

The following performance bound says, roughly, that the test based on (15) is reliable when (13) does not hold, and that consistent support estimation is possible with a slightly stronger signal-to-noise ratio. The procedure is also adaptive to ν.

PROPOSITION 3. *Assume* $\mathbf{\Sigma}$ *is known and that* $p \wedge n \to \infty$. *For any sequence* s *of sparsity*, *the following results hold.*

- Detection. *Consider testing* (10) *versus* (11) *with* $\nu \in (0, 1)$ *fixed. Let* T_s *denote the statistic* (15). *There is a sequence of critical values* t *such that the test* $\phi = \{T_s \ge t\}$ *is asymptotically powerful*, *meaning* $\gamma(\phi; \Omega_0(\nu, \mathbf{\Sigma}), \Omega_1(\nu, \mathbf{\Sigma}, R_0, r_n)) \to 0$, *if the minimum Mahalanobis distance* r_n *satisfies*

$$\liminf \frac{\nu(1-\nu) r_n}{\sqrt{\frac{s}{n}\log(\frac{ep}{s})} \vee \frac{s}{n}\log(\frac{ep}{s})} > C, \tag{17}$$

  where C *is a universal constant.*
- Variable selection. *Consider the model* (11). *Let* $\hat{u}_s$ *denote a maximizer of* (15) *and let* $\hat{v}_s = \mathbf{\Sigma}^{1/2}\hat{u}_s$. *Then under the slightly stronger condition*

$$\nu(1-\nu)\Delta\mu^\top \mathbf{\Sigma}^{-1}\Delta\mu \gg \sqrt{\frac{s}{n}\log\left(\frac{ep}{s}\right)} \vee \frac{s}{n}\log\left(\frac{ep}{s}\right), \tag{18}$$

  and the assumption that the effective dynamic range of $\Delta\mu$ *and the* $2s$*-sparse Riesz constant of* $\mathbf{\Sigma}$ *are both bounded*, *the support of* $\hat{v}_s$ *is consistent for the support of* $\Delta\mu$.

We note that without a bound on the dynamic range of $\Delta\mu$, its largest entries could overwhelm the smaller (nonzero) ones and make consistent support recovery difficult, or even impossible.

Special case: $\mathbf{\Sigma} = \mathbf{I}$. As a consequence of the remark below Proposition 1, the detection boundary is roughly at

$$\|\Delta\mu\|^2 \approx \left[\sqrt{\frac{s}{n}\log\left(\frac{ep}{s}\right)} \vee \frac{s}{n}\log\left(\frac{ep}{s}\right)\right] \wedge \sqrt{\frac{p}{n}},$$

except in the regime where $s \geq n$ and $s \approx \sqrt{np}$ where there is a logarithmic gap between the upper and lower bounds. The statistic of choice is (15), the top s-sparse eigenvalue of $\hat{\mathbf{\Sigma}}$, which is also known to be rate-optimal for the problem of testing for a top principal direction in a spiked Gaussian covariance model [Berthet and Rigollet (2013a)].

REMARK. In general, the statistic defined in (15) is not the top s-sparse eigenvalue of $\hat{\mathbf{\Sigma}}_{\ddagger}$, which is instead defined as

$$\lambda_s^{\max}(\hat{\mathbf{\Sigma}}_{\ddagger}) = \max_{\|u\|=1, \|u\|_0 \leq s} u^\top \hat{\mathbf{\Sigma}}_{\ddagger} u. \tag{19}$$

We are only able to show that the statistic (19) is asymptotically powerful in the following sense, that $\gamma(\phi; \Omega_0(\nu, \mathbf{\Sigma}), \Omega_1(\nu, \mathbf{\Sigma}, R_1, r_n)) \to 0$ for $R_1(\theta) := \nu(1 - \nu)\|\Delta\mu\|^4/\Delta\mu^\top \mathbf{\Sigma} \Delta\mu$ and r_n satisfying (17) for some constant $C > 0$. However, this bound is weaker than what we obtain in Proposition 3 for (15), simply because the function $R_1(\theta)$ is smaller than the Mahalanobis distance $R_0(\theta)$. Indeed, using the Cauchy–Schwarz inequality,

$$\begin{aligned}\|\Delta\mu\|^4 &= [(\mathbf{\Sigma}^{-1/2}\Delta\mu)^\top(\mathbf{\Sigma}^{1/2}\Delta\mu)]^2 \\ &\leq \|\mathbf{\Sigma}^{-1/2}\Delta\mu\|^2 \|\mathbf{\Sigma}^{1/2}\Delta\mu\|^2 \\ &= (\Delta\mu^\top \mathbf{\Sigma}^{-1}\Delta\mu)(\Delta\mu^\top \mathbf{\Sigma}\Delta\mu).\end{aligned}$$

3. Unknown covariance matrix.

We distinguish between the symmetric case ($\nu = 1/2$) and the asymmetric case ($\nu \neq 1/2$). In terms of methodology, skewness and kurtosis tests have played a major role in testing for multivariate normality, at least since the seminal work of Mardia (1970). Some of these tests are based on estimating the covariance matrix and, therefore, are not applicable in high-dimensional settings where $p > n$, at least not without additional assumptions on the covariance matrix. More malleable approaches are projection tests such as those proposed by Malkovich and Afifi (1973). We adapt such tests to the sparse setting considered here, and also design new variants to palliate some deficiencies.

3.1. *Symmetric setting.*

Consider the case where the covariance matrix is unknown and where the mixture distribution is symmetric, meaning that $\nu = 1/2$. The resulting mixture testing problem is more difficult than in the asymmetric setting treated in Section 3.2.

3.1.1. *Minimax lower bound.* We start with a minimax lower bound with respect to the signal-to-noise ratio:

$$R_1(\theta) = \frac{\|\Delta\mu\|^4}{\Delta\mu^\top \mathbf{\Sigma} \Delta\mu}. \tag{20}$$

We will see in Proposition 8 that the minimax detection rate with respect to the Mahalanobis distance R_0 is degenerate in a sparse high-dimensional setting.

PROPOSITION 4. *Consider testing* (8) *versus* (9) *with* $\nu = 1/2$. *Then* $\liminf \gamma^*(\Omega_0(\nu), \Omega_1(\nu, R_1, r_n)) = 1$ *in the following cases*:

- Nonsparse setting: $s = p \to \infty$ *and*

$$r_n \ll (p/n)^{1/4}; \tag{21}$$

  or $p \gg n$ *and*

$$\limsup r_n e^{-Cp/n} < 1, \tag{22}$$

  where $C > 0$ *is a universal constant.*
- Sparse setting: $p/s \to \infty$ *and*

$$\limsup r_n \left[\frac{s}{n} \log\left(\frac{ep}{s}\right)\right]^{-1/4} \le C_1, \tag{23}$$

  where $C_1 > 0$ *is a universal constant*; *or* $n \le \frac{1}{C_2} s \log(ep/s)$ *and*

$$\limsup r_n e^{-C_3 \frac{s}{n} \log(\frac{ep}{s})} \le 1, \tag{24}$$

  where $C_2, C_3 > 0$ *are universal constants.*

REMARK. From the above proposition, we deduce that the testing problem becomes extremely difficult when $\zeta := \frac{s}{n}\log(ep/s) \to \infty$, in the sense that the minimax detection rate is at least exponentially large with respect to ζ. A similar phenomenon occurs in other high-dimensional detection problems such as in sparse linear regression [Verzelen (2012)].

REMARK. Similar to what we do in the proof of Proposition 1, we reduce to testing subclasses of hypotheses. Specifically, we reduce to testing $\theta \in \widetilde{\Omega}_0 := \{\theta = (\frac{1}{2}, 0, 0, \mathbf{I})\}$ against

$$\theta \in \widetilde{\Omega}_1\left(\frac{1}{2}, R_1, r_n\right) := \left\{\left(\frac{1}{2}, -\mu, \mu, \mathbf{\Sigma}_\mu\right), \mu \text{ is } s\text{-sparse and } R_1(\theta) \ge r_n\right\},$$

where $\mathbf{\Sigma}_\mu := \mathbf{I} - \mu\mu^\top$. Note that, in this testing problem, the variables are centered and $\operatorname{Cov}(X) = \mathbf{I}$, both under the null and under the alternative.

3.1.2. *A classical approach based on the kurtosis.* Unlike in Section 2, here the covariance matrix $\mathrm{Cov}(X)$, by itself, does not contain in any sensible information to distinguish the null hypothesis from the alternative. It is therefore natural to consider higher order moments of X. In the symmetric setting, a traditional approach is the use of a kurtosis test. Malkovich and Afifi (1973) propose a projection test based on the kurtosis for the problem of testing for multivariate normality. This is easily adapted to the sparse setting. The resulting test is based on rejecting for *small* values of

$$\min_{\|u\|_0 \le s} \frac{\sum_i [u^\top (X_i - \bar{X})]^4}{(\sum_i [u^\top (X_i - \bar{X})]^2)^2}. \tag{25}$$

We note that Malkovich and Afifi (1973)—who are interested in testing for multivariate normality and do not make sparsity assumptions—reject for unusually large or small values of the above ratio along a general (meaning, not necessarily sparse) direction u.

REMARK. Although the null distribution of (25) depends on the unknown covariance matrix $\mathbf{\Sigma}$, it can calibrated by a simple Bonferroni correction, which is possible because (25) is the minimum over all subsets $S \subset [p]$ of size s of variables which have a null distribution that is independent of $\mathbf{\Sigma}$. The same applies to (27), (36) and (38).

PROPOSITION 5. *Consider testing* (8) *versus* (9) *with the assumption* $|\nu - 1/2| < \frac{\sqrt{3}}{6}$, *and assume that* $n \gg [s \log(ep/s)]^2$. *Let* T *denote the statistic* (25). *There is a sequence of critical values* t *such that such the test* $\phi = \{T \le t\}$ *is asymptotically powerful*, *meaning* $\gamma(\phi; \Omega_0(\nu), \Omega_1(\nu, R_1, r_n)) \to 0$, *if*

$$r_n \gg \left[\frac{s}{n}\log\left(\frac{ep}{s}\right)\right]^{1/4} \vee \left[\frac{s}{\sqrt{n}}\log\left(\frac{ep}{s}\right)\right]. \tag{26}$$

We see that there is a substantial discrepancy between the performance that we establish for the sparse kurtosis test (25) in Proposition 5 and the lower bound obtained in Proposition 4. The issue comes from the control of the numerator in (25), in that the estimator for the fourth moment has a heavy tail and does not concentrate enough when $s \log(ep/s)$ becomes large.

3.1.3. *A new approach based on the first absolute moment.* Instead of a kurtosis test, which is based on the fourth central moment, we propose a test based on the first central absolute moment in order to palliate the aforementioned issues. The test rejects for large values of

$$\max_{\|u\|_0 \le s} \frac{\sum_i |u^\top (X_i - \bar{X})|}{(\sum_i [u^\top (X_i - \bar{X})]^2)^{1/2}}. \tag{27}$$

PROPOSITION 6. *Consider testing* (8) *versus* (9) *with the assumption* $|\nu - 1/2| < \frac{1}{6}$, *and assume that* $n \gg s\log(ep/s)$. *Let* T *denote the statistic* (27). *There is a sequence of critical values* t *such that the test* $\phi = \{T \geq t\}$ *is asymptotically powerful*, *meaning* $\gamma(\phi; \Omega_0(\nu), \Omega_1(\nu, R_1, r_n)) \to 0$, *if*

$$r_n \gg \left[\frac{s}{n}\log(ep/s)\right]^{1/4}. \tag{28}$$

Consequently, the test based on (27) achieves the minimax detection boundaries (21) and (23). Note that the assumption $n \gg s\log(ep/s)$ is necessary in view of Proposition 4.

Variable selection. In regard to variable selection, we are unable to use the statistic (27) [or the original statistic (25)]. To see why, for concreteness, consider the situation where the variables have zero mean under the null and alternative, and assume the mixture is symmetric ($\nu = 1/2$). Using the arguments provided in the proof of Proposition 6, we can show that, if $n \to \infty$ fast enough, then the result of maximizing of the empirical ratio in (27) is consistent with

$$\max_{\|u\|_0 \leq s} \frac{\mathbb{E}|u^\top X|}{(\mathbb{E}[u^\top X]^2)^{1/2}}. \tag{29}$$

Elementary calculations yield

$$\frac{\mathbb{E}|u^\top X|}{(\mathbb{E}[u^\top X]^2)^{1/2}} = \frac{\mathbb{E}|h_u/2 + z|}{(1 + h_u^2/4)^{1/2}} = \sqrt{\frac{2}{\pi}}\left(1 + \frac{1}{192}h_u^4 + O(h_u^6)\right), \tag{30}$$

where $z \sim \mathcal{N}(0,1)$ and when $h_u := u^\top \Delta\mu/\sqrt{u^\top \mathbf{\Sigma} u} \to 0$, which is allowed in (28). The maximizer of (29) is therefore close to $\arg\max_{\|u\|_0 \leq s} |h_u|$, which does not necessarily have the same support as $\Delta\mu$.

In view of (30), we normalize (27) to cancel the denominator $(u^\top \mathbf{\Sigma} u)^2$ in h_u^4, so that the maximizer is approximately aligned with $\Delta\mu$. This motivates us to consider the support estimator $\hat{J} = \operatorname{supp}(\hat{u})$, where $\hat{u}$ maximizes

$$\max_{\|u\|_0 \leq s, \|u\|=1} \left[\frac{\sum_i |u^\top (X_i - \bar{X})|}{(\sum_i [u^\top (X_i - \bar{X})]^2)^{1/2}} - \sqrt{\frac{2}{\pi}}\right]\left(\sum_i [u^\top (X_i - \bar{X})]^2\right)^2. \tag{31}$$

PROPOSITION 7. *Consider the model* (11) *with the assumption* $|\nu - 1/2| < \frac{1}{6}$, *assume that* $n \gg s\log(ep/s)$ *and that* $\Delta\mu$ *is* s-*sparse*. *Then the estimator defined in* (31) *is consistent for the support of* $\Delta\mu$ *if*

$$1 \gg \frac{\|\Delta\mu\|^4}{\Delta\mu^\top \mathbf{\Sigma} \Delta\mu} \gg \left[\frac{s}{n}\log\left(\frac{ep}{s}\right)\right]^{1/4}, \tag{32}$$

and the effective dynamic range of $\Delta\mu$ *and the* $2s$-*sparse Riesz constant of* $\mathbf{\Sigma}$ *are both bounded.*

Consequently, the estimator (31) is consistent when the signal strength is just above the detection threshold. When the signal is strong, the procedure above seems to fail. However, we mention that the simpler support estimator based on

$$\hat{u} \in \underset{\|u\|_0 \le s, \|u\|=1}{\arg\max} \sum_i \left| u^\top (X_i - \bar{X}) \right|,$$

is consistent when $\|\Delta\mu\|^4 / \Delta\mu^\top \mathbf{\Sigma} \Delta\mu \to \infty$ under the same conditions otherwise. Details are omitted as the arguments are similar, but simpler, than those underlying Proposition 7. Compare also with the coordinatewise support estimator introduced in Section 4.1.2.

3.1.4. *The Mahalanobis metric.* The lower bounds obtained in Proposition 4 are in terms of R_1, while those we obtained for the case where the covariance matrix is known in Proposition 1 are in terms of the Mahalanobis metric R_0. While these two metrics are equivalent if the s-sparse Riesz constant of $\mathbf{\Sigma}$ is bounded, this is not so for any arbitrary $\mathbf{\Sigma}$. We state below an information bound in terms of the Mahalanobis distance that is exponential in p/n, even when $\Delta\mu$ is 1-sparse. This suggests that R_1 is more relevant than R_0 in the present context.

PROPOSITION 8. *If* $p \gg n$, *then* $\liminf \gamma^*(\Omega_0(\nu), \Omega_1(\nu, R_0, r_n)) = 1$ *when*

$$\Delta\mu^\top \mathbf{\Sigma}^{-1} \Delta\mu \ll \frac{e^{p/(2n)}}{np}. \tag{33}$$

Again, the lower bound is proved by a reduction to the following simpler testing problem. Fix a 1-sparse vector $\Delta\mu$, and consider $\theta \in \Omega_0^\ddagger := \{(\frac{1}{2}, 0, 0, \mathbf{I})\}$ against $\theta \in \Omega_1^\ddagger(\frac{1}{2}, R_0, r_n)$, where

$$\Omega_1^\ddagger\left(\frac{1}{2}, R_0, r_n\right) := \left\{\theta = \left(\frac{1}{2}, -\frac{1}{2}\Delta\mu, \frac{1}{2}\Delta\mu, \mathbf{\Sigma}\right) : \right.$$
$$\left. \mathbf{\Sigma} - \mathbf{I} \text{ has rank 1 and } R_0(\theta) \ge r_n \right\}.$$

In contrast to the collection $\widetilde{\Omega}_1(\frac{1}{2}, R_0, r_n)$ used in the proof of Proposition 4, $\Omega_1^\ddagger(\frac{1}{2}, R_0, r_n)$ contains the collections of all rank 1 perturbation of the identity covariance matrix.

3.2. *Asymmetric setting.* Consider the case where the covariance matrix is unknown and where the mixture distribution is asymmetric, meaning that $\nu \neq 1/2$. As we shall see, detection in the asymmetric setting is quantifiably easier than in the symmetric setting, due to the ability to test for asymmetry (in a particular manner).

3.2.1. *Minimax lower bound.* We start with a minimax lower bound. As in the symmetric setting covered in Section 3.1, we use the signal-to-noise function R_1 defined in (20).

PROPOSITION 9. *Consider testing* (8) *versus* (9) *with* $\nu \neq 1/2$ *fixed. Then* $\liminf \gamma^*(\Omega_0(\nu), \Omega_1(\nu, R_1, r_n)) = 1$ *in the following cases*:

- Nonsparse setting. *Assume* $s = p \to \infty$ *and* $p = o(n)$ *and*

$$r_n \ll (p/n)^{1/3}. \tag{34}$$

- Sparse setting. *Assume* $p/s \to \infty$ *and* $n \gg s \log(ep/s)$ *and*

$$\limsup r_n \left[\frac{s}{n} \log\left(\frac{ep}{s}\right)\right]^{-1/3} \le C_\nu, \tag{35}$$

  where $C_\nu > 0$ *is a constant.*

3.2.2. *A classical approach based on the skewness.* The standard approach in this asymmetric setting is a skewness test. We adapt the projection skewness test of Malkovich and Afifi (1973) to our sparse setting. This leads us to rejecting for large values of the following statistic:

$$\max_{\|u\|_0 \le s} \frac{\sum_i [u^\top (X_i - \bar{X})]^3}{(\sum_i [u^\top (X_i - \bar{X})]^2)^{3/2}}. \tag{36}$$

PROPOSITION 10. *Consider testing* (8) *versus* (9) *with the assumption* $\nu \neq 1/2$ *fixed, and* $n \gg s \log(ep/s)$. *Let* T *denote the statistic* (36). *There is a sequence of critical values* t *such that the test* $\phi = \{T \ge t\}$ *is asymptotically powerful, meaning* $\gamma(\phi; \Omega_0(\nu), \Omega_1(\nu, R_1, r_n)) \to 0$, *if*

$$\big(\nu(1-\nu)|1-2\nu|\big)^{2/3} r_n \gg \left[\frac{s}{n} \log\left(\frac{ep}{s}\right)\right]^{1/3} \vee \left[n^{1/3} \frac{s}{n} \log(ep/s)\right]. \tag{37}$$

We notice a substantial discrepancy between this rate and the lower bound obtained in Proposition 9. As with the kurtosis statistic, the main issue is our difficulty with proving that the third moment concentrates enough under the null.

3.2.3. *A new approach based on the signed second moment.* We replace the third moment with the second signed moment, leading to

$$\max_{\|u\|_0 \le s} \frac{\sum_i [u^\top (X_i - \bar{X})]^2 \operatorname{sign}(u^\top (X_i - \bar{X}))}{\sum_i [u^\top (X_i - \bar{X})]^2}. \tag{38}$$

PROPOSITION 11. *Consider testing* (8) *versus* (9) *with the assumption* $\nu \neq 1/2$ *fixed, and* $n \gg s \log(ep/s)$. *Let* T *denote the statistic* (38). *There is a sequence* *of critical values t such that the test $\phi = \{T \geq t\}$ is asymptotically powerful, meaning $\gamma(\phi; \Omega_0(\nu), \Omega_1(\nu, R_1, r_n)) \to 0$, if*

$$\liminf \bigl(\nu(1-\nu)|1-2\nu|\bigr)^{2/3} r_n \left[\frac{s}{n}\log\left(\frac{ep}{s}\right)\right]^{-1/3} \geq C, \tag{39}$$

where C is a universal constant.

We see that this test achieves the minimax rate established in (35). Note that the minimax detection rate is substantially faster in the asymmetric case compared with the symmetric case.

Variable selection. Here, too, we are unable to use the statistic (38) to perform variable selection. In analogy with the symmetric case, we consider the estimator $\hat{J} = \operatorname{supp}(\hat{u})$, where $\hat{u}$ maximizes

$$\max_{\|u\|=1, \|u\|_0 \leq s} \left[\sum_i [u^\top (X_i - \bar{X})]^2 \operatorname{sign}\bigl(u^\top (X_i - \bar{X})\bigr)\right]\left[\sum_i [u^\top (X_i - \bar{X})]^2\right]^{1/2}. \tag{40}$$

Despite the strong parallel with the statistic (31), we were not able to obtain a satisfactory performance for (40). We mention, as we did before, that other estimators may be needed when the signal is strong. And we also refer the reader to Section 4.1.2, where a coordinatewise support estimator is introduced.

3.3. *Diagonal model.* A popular approach in situations where the covariance is unknown is to assume it is diagonal. In (supervised) classification, this leads to diagonal linear discriminant analysis, which corresponds to the naive Bayes classifier in the Gaussian mixture model [Bickel and Levina (2004)]. Define

$$\kappa = \|\Delta\mu_\ddagger\|_\infty / \|\Delta\mu_\ddagger\|, \qquad \Delta\mu_\ddagger := \mathbf{\Sigma}^{-1/2}\Delta\mu. \tag{41}$$

Given $\nu \in (0,1)$, $a \in (0,1)$, and $s \leq p$, we consider the mixture testing problem with unknown diagonal covariance matrix, which we define as testing

$$\breve{\Omega}_0 = \{\theta = (\nu, \mu, \mu, \mathbf{\Sigma}), \mu \in \mathbb{R}^p, \mathbf{\Sigma} \text{ diagonal p.s.d.}\} \tag{42}$$

versus

$$\breve{\Omega}_1(\nu, R, r_n) := \breve{\Omega}_1(\nu) \cap \{\theta : R(\theta) \geq r_n\}, \tag{43}$$

where

$$\begin{aligned}\breve{\Omega}_1(\nu) := \{\theta = (\nu, \mu_0, \mu_1, \mathbf{\Sigma}) :& \\ &\mu_0, \mu_1 \in \mathbb{R}^p \text{ satisfying (3)}, \mathbf{\Sigma} \text{ diagonal p.s.d.}, \kappa \leq a\}.\end{aligned}$$

In this situation, it is natural to estimate the covariance matrix by the diagonal of the sample covariance matrix. We can then use this estimator in place of $\mathbf{\Sigma}$ in (15), yielding the following statistic:

$$\max_{\|u\|_0 \leq s} \frac{u^\top \hat{\mathbf{\Sigma}} u}{u^\top \operatorname{diag}(\hat{\mathbf{\Sigma}}) u}, \tag{44}$$

with the convention that $0/0 = 0$, where for a square matrix $A = (a_{ij})$, $\operatorname{diag}(A)$ denotes the diagonal matrix with diagonal elements (a_{ii}). The null distribution of the test statistic (44) does not depend on $\mathbf{\Sigma}$ as long as it is diagonal.

PROPOSITION 12. *Consider testing* (42) *versus* (43) *with* $\nu \in (0, 1)$ *fixed, and* $1 \ll \log p \ll n$. *Assume that* κ *in* (41) *is bounded away from* 1.

- Detection. *Let* T *denote the statistic* (44). *There is a sequence of critical values* t *such that the test* $\phi = \{T \geq t\}$ *is asymptotically powerful, meaning* $\gamma(\phi; \breve{\Omega}_0, \breve{\Omega}_1(\nu, R, r_n)) \to 0$, *if*

$$\nu(1-\nu)r_n \geq \frac{C}{1-a^2}\left[\sqrt{\frac{s}{n}\log(ep/s)} \vee \frac{s}{n}\log(ep/s)\right], \tag{45}$$

  where $C > 0$ *is a universal constant.*
- Variable selection. *Let* $\hat{u}$ *denote a maximizer of* (44). *Then under the slightly stronger condition* (18), *and assuming that* $\|\Delta\mu\|_0 > 1$ *and that the effective dynamic range of* $\Delta\mu_\ddagger$ *is bounded, the support of* $\hat{u}$ *is consistent for the support of* $\Delta\mu$.

Proposition 12 presents an interesting phenomenon. When the covariance matrix is supposed to be diagonal but is unknown, there is a qualitative difference between the case $\|\Delta\mu\|_0 = 1$ and $\|\Delta\mu\|_0 > 1$. The conditions of Proposition 12 imply that $\|\Delta\mu\|_0 > 1$. When $\|\Delta\mu\|_0 = 1$, the statistic (44) is useless at either detection or variable selection, since in that situation $\operatorname{Cov}(X)$ is also diagonal under the alternative. In that case, the optimal detection rate is the same as that for general unknown covariances, that is, $(\log(p)/n)^{1/4}$ when $\nu = 1/2$ and $(\log(p)/n)^{1/3}$ when $\nu \neq 1/2$. Indeed, when $s = 1$, the proofs of Propositions 4 and 9 are based on diagonal covariance matrices. When $\|\Delta\mu\|_0 > 1$, (45) is the same as (17), meaning we can do as well as if $\mathbf{\Sigma}$ were known, as long as κ remains bounded away from 1, meaning that $\Delta\mu_\ddagger$ is not approximately 1-sparse.

4. Computationally tractable methods and numerical experiments.

A test statistic of the form $\max\{G(u; X_1, \ldots, X_n) : \|u\|_0 \leq s\}$, where G is a real-valued function, results in a combinatorial maximization over the subsets of $[p]$ of size at most s, and this is very quickly intractable when $s \to \infty$ as $n \to \infty$, because there are $\binom{p}{s} \geq (p/s)^s$ such subsets.

More specifically, here we say that a method is computationally tractable if it can be computed in time polynomial in (n, p, s). Although such a method may still be practically intractable for large problems, on a theoretical level, it provides a qualitative definition in line with a central concern in theoretical computer science. Among the statistics considered in Sections 2 and 3, only the largest eigenvalue $\hat{\lambda}_{\mathbf{\Sigma}}^{\max}$ defined in (14) is known to be computable in polynomial time. All the other methods are tailored to the sparse setting and are combinatorial in nature. This motivates the development of computationally tractable methods for this setting.

4.1. *Coordinatewise methods.* The simplest computationally tractable methods are arguably those based on testing each coordinate at a time. Such a method is of the form

$$M\big(T_{\ddagger}(X^1), \dots, T_{\ddagger}(X^p)\big), \tag{46}$$

where $X^j = (X_{i,j} : i \in [n])$ is the jth variable, $T_{\ddagger}$ is a test statistic for mixture testing in dimension one, and M implements a multiple testing procedure. In what follows, we opt for the simple Bonferroni correction, which corresponds to $M(t_1, \dots, t_p) = \max_j t_j$. Coordinatewise testing and/or variable selection of this type is considered in Azizyan, Singh and Wasserman (2013), Chan and Hall (2010) and also in Amini and Wainwright (2009), Berthet and Rigollet (2013a), Johnstone and Lu (2009) in the context of sparse PCA. Such approaches are also considered in recent work[3] by Jin and Wang (2014) and Jin, Ke and Wang (2015), who obtain very precise minimax results when the covariance matrix has relatively small condition number. Except for Chan and Hall (2010), where a nonparametric setting is considered, these papers assume that the covariance matrix is known.

4.1.1. *Known covariance.* Denote $\mathbf{\Sigma} = (\sigma_{jk})$ and $\hat{\mathbf{\Sigma}} = (\hat{\sigma}_{jk})$. Inspired by the statistic (19), we arrive at the maximum canonical variance statistic

$$\max_{j \in [p]} \frac{\hat{\sigma}_{jj}}{\sigma_{jj}}, \tag{47}$$

and at the corresponding support estimator

$$\hat{J} = \{j \in [p] : \hat{\sigma}_{jj}/\sigma_{jj} > t\}, \qquad t := 1 + 5\left(\sqrt{\frac{\log(p)}{n}} \vee \frac{\log(p)}{n}\right), \tag{48}$$

for a given threshold $\omega \to \infty$. Note that (47) corresponds to working with the test statistic $T_{\ddagger}(x_1, \dots, x_n) = \frac{1}{n}\sum_i (x_i - \bar{x})^2$ in (46).

PROPOSITION 13. *Consider testing* (10) *versus* (11) *with* $\nu \in (0, 1)$ *fixed and* $p \to \infty$. *Denoting* T *the statistic* (47), *we consider the test* $\phi = \{T \geq t\}$ *with* t *defined in* (48). *The test* ϕ *has asymptotic level* 0. *Moreover, it has asymptotic power one if*

$$\nu(1 - \nu) \max_{j \in [p]} \frac{\Delta\mu_j^2}{\sigma_{jj}} > C_1\left(\sqrt{\frac{\log(p)}{n}} \vee \frac{\log(p)}{n}\right), \tag{49}$$

where $C_1 > 0$ *is a universal constant. Moreover, the estimator* (48) *is consistent for the support of* $\Delta\mu$ *if*

$$\nu(1 - \nu) \min_{j \in J} \frac{\Delta\mu_j^2}{\sigma_{jj}} > C_2\left(\sqrt{\frac{\log(p)}{n}} \vee \frac{\log(p)}{n}\right),$$

where $C_2 > 0$ *is a universal constant.*

[3] This work was made publicly available after ours.

The proof is a straightforward adaptation of that of Proposition 3, and is omitted.

REMARK. A stronger result can be obtained by using (15) instead of (19), leading to $\hat{\lambda}^{\max}_{1,\boldsymbol{\Sigma}}$ instead of (47), but the approach is somewhat less natural and the resulting performance bound somewhat less intuitive.

Special case $\boldsymbol{\Sigma} = \mathbf{I}$. In Section 2, we proved that the test based on (15) is asymptotically powerful under (17), that is,

$$\nu(1-\nu)\|\Delta\mu\|^2 \geq C\left[\sqrt{\frac{s}{n}\log(ep/s)} \vee \frac{s}{n}\log(ep/s)\right].$$

The coordinatewise test was shown here to be asymptotically powerful under (49), that is,

$$\nu(1-\nu)\|\Delta\mu\|_\infty^2 \geq C\left[\sqrt{\frac{\log(p)}{n}} \vee \frac{\log(p)}{n}\right].$$

(C is a sufficiently large constant.) When the energy of $\Delta\mu$ is spread over its support, we have $\|\Delta\mu\|_\infty \approx \|\Delta\mu\|/\sqrt{s}$, in which case the latter condition becomes

$$\nu(1-\nu)\|\Delta\mu\|^2 \geq C\left[\sqrt{\frac{s^2}{n}\log(ep/s)} \vee \left(\frac{s}{n}\log(ep/s)\right)\right].$$

Hence, the coordinatewise method is shown to achieve a detection rate within a multiplicative factor $\sqrt{s}$ of the optimal rate. In the special situation where $n = O(\log p)$, the coordinatewise method even achieves the optimal rate. In general, however, there is this multiplicative factor of $\sqrt{s}$ between the detection bounds. We speculate that this factor of $\sqrt{s}$ is unavoidable and incurred by any polynomial time method. Our speculation is based on an analogy with the sparse PCA detection problem and the recent work of Berthet and Rigollet (2013b). These authors prove that a multiplicative factor of $\sqrt{s}$ applies to any polynomial time algorithm, if some classical problem in computational complexity, known as the planted clique problem, is not solvable in polynomial time; see Berthet and Rigollet (2013b) for definitions and pointers to the literature. (Although we focused on the case $\boldsymbol{\Sigma} = \mathbf{I}$, this discussion is in fact valid for general covariance matrices $\boldsymbol{\Sigma}$ as long as the s-sparse Riesz constants are bounded.)

4.1.2. *Unknown covariance.* We adapt the statistics (27) and (38) to coordinatewise methods by considering $s = 1$, thus working with

$$T_1 = \max_{j\in[p]} T_{1,j}, \qquad T_{1,j} := \sum_{i=1}^{n} \frac{|X_{i,j} - \bar{X}_j|}{\sqrt{\hat{\sigma}_{jj}}} \tag{50}$$

and

$$T_2 = \max_{j\in[p]} T_{2,j}, \qquad T_{2,j} := \left|\sum_{i=1}^{n} \frac{(X_{i,j} - \bar{X}_j)^2}{\hat{\sigma}_{jj}} \operatorname{sign}(X_{i,j} - \bar{X}_j)\right|. \tag{51}$$

Although the null distribution of (50) depends on the unknown covariance matrix $\mathbf{\Sigma}$, it can be calibrated by a simple Bonferroni correction, which is possible because the terms in the maximum have a null distribution that is independent of $\mathbf{\Sigma}$. The same applies to (51).

For any $u \in (0,1)$, denote by $q_1^{-1}(u)$ the $(1-u)$-quantile of the distribution of $T_{1,1}$ under the null hypothesis. Given some level $\alpha \in (0,1)$, denote by $\hat{J}_1$ the set of indices such that $T_{1,j}$ is significant at level α/p, namely,

$$\hat{J}_1 := \{j : T_{1,j} > q_1^{-1}(\alpha/p)\}.$$

The estimator $\hat{J}_2$ is defined analogously based on (51). In practice, the quantile functions q_1^{-1} and q_2^{-1} can be easily estimated by Monte Carlo simulations.

PROPOSITION 14. *Consider testing* (8) *versus* (9) *with* $n \gg \log(p) \gg 1$. *Consider a sequence of levels* α *satisfying* $\alpha = o(1)$ *and* $\alpha \geq p^{-a}$ *for some fixed* $a > 0$ *in the definition of* $\hat{J}_1$ *and* $\hat{J}_2$.

- Detection. *The test* $\phi_1 := \{\hat{J}_1 \neq \varnothing\}$ *has a level smaller than* α. *Moreover, it has asymptotic power* 1 *if if* $|\nu - 1/2| < \frac{1}{6}$ *and*

$$\max_{j\in[p]} \frac{(\Delta\mu_j)^2}{\sigma_{jj}} \gg \left[\frac{\log(p)}{n}\right]^{1/4}. \tag{52}$$

  The test $\phi_2 := \{\hat{J}_2 \neq \varnothing\}$ *has a level smaller than* α. *Moreover, it has asymptotic power* 1 *if*

$$\liminf\bigl(\nu(1-\nu)|1-2\nu|\bigr)^{2/3} \max_{j\in[p]} \frac{(\Delta\mu_j)^2}{\sigma_{jj}} \left[\frac{\log(p)}{n}\right]^{-1/3} \geq C. \tag{53}$$

- Variable selection. *Assume that the effective dynamic range of* $\Delta\mu$ *and the* $2s$*-sparse Riesz constant of* $\mathbf{\Sigma}$ *are both bounded.*
  - *If* $|\nu - 1/2| < \frac{1}{6}$, *then under the stronger condition that*

$$\frac{\|\Delta\mu\|^4}{\Delta\mu^\top \mathbf{\Sigma} \Delta\mu} \gg s\left[\frac{\log(p)}{n}\right]^{1/4}, \tag{54}$$

    $\hat{J}_1$ *is consistent for the support of* $\Delta\mu$.
  - *If* $\nu \neq 1/2$ *is fixed, then under the stronger condition that*

$$\frac{\|\Delta\mu\|^4}{\Delta\mu^\top \mathbf{\Sigma} \Delta\mu} \gg s\left[\frac{\log(p)}{n}\right]^{1/3}, \tag{55}$$

    $\hat{J}_2$ *is consistent for the support of* $\Delta\mu$.

REMARK. Assuming the energy of $\Delta\mu$ is spread over its support, and that the s-sparse Riesz constant of $\mathbf{\Sigma}$ is bounded, (52) and (53) reduce to

$$\frac{\|\Delta\mu\|^4}{\Delta\mu^\top \mathbf{\Sigma}\Delta\mu} \gg s\left[\frac{\log(p)}{n}\right]^{1/4} = s^{3/4}\left[s\frac{\log(p)}{n}\right]^{1/4},$$

$$\bigl(\nu(1-\nu)|1-2\nu|\bigr)^{2/3}\frac{\|\Delta\mu\|^4}{\Delta\mu^\top \mathbf{\Sigma}\Delta\mu} \geq Cs\left[\frac{\log(p)}{n}\right]^{1/3} = Cs^{2/3}\left[s\frac{\log(p)}{n}\right]^{1/3}.$$

Compared to (28) and (39), the performances of the coordinatewise methods are within $s^{3/4}$ and $s^{2/3}$ multiplicative factors, respectively, of the optimal rates. We do not know to what extent this is intrinsic to the problem, namely, whether there are polynomial time methods with performance bounds that come closer to the optimal bounds.

4.2. *Other computationally tractable methods.* Beyond methods based on examining k-tuples of coordinates, instead of just $k = 1$ coordinate at a time, and other heuristics based on principal component analysis [Srivastava (1984)], more sophisticated methods may be needed. We present two methods based on relaxations of the sparse eigenvalue problem, which we learned from Berthet and Rigollet (2013a), who applied it to the problem of detecting a top principal component in a spiked covariance model. See also Amini and Wainwright (2009). Assume for simplicity that $\mathbf{\Sigma} = \mathbf{I}$ (or, equivalently, that it is diagonal and known), so that the sparse eigenvalues defined in (15) and (19) coincide, and in both cases, the maximization is over s-sparse unit vectors.

- The first relaxation is the semidefinite program (SDP) of d'Aspremont et al. (2007):

$$\text{(56)} \qquad \begin{gathered} \mathrm{SDP}_s(\mathbf{A}) = \max_{\mathbf{Z}} \operatorname{trace}(\mathbf{AZ}), \\ \text{subject to } \mathbf{Z} \succeq 0, \operatorname{trace}(\mathbf{Z}) = 1, |\mathbf{Z}|_1 \leq s, \end{gathered}$$

  where the maximum is over positive semidefinite matrices $\mathbf{Z} = (Z_{st})$ and $|\mathbf{Z}|_1 := \sum_{s,t}|Z_{st}|$. We would then use $\mathrm{SDP}_s(\hat{\mathbf{\Sigma}})$.
- The second relaxation leads to using the minimum dual perturbation

$$\text{(57)} \qquad \mathrm{MDP}_s(\mathbf{A}) := \min_{z\geq 0}\bigl[\lambda^{\max}\bigl(\tau_z(\mathbf{A})\bigr) + sz\bigr],$$

  where τ_z is entrywise soft-thresholding at z, meaning that for a matrix $\mathbf{A} = (a_{jk})$, $\tau_z(\mathbf{A}) = (b_{jk})$, where $b_{jk} = \operatorname{sign}(a_{jk})\max(|a_{jk}| - z, 0)$. We would then use $\mathrm{MDP}_s(\hat{\mathbf{\Sigma}})$.

Both relaxations operate in polynomial time. That said, the semidefinite program does not scale well, while the second relaxation is computationally more friendly as it boils down to a one-dimensional grid search over $z \in \mathbb{R}$ requiring the computation of the top eigenvalue of symmetric matrix at every grid point.

PROPOSITION 15. *Consider testing* (10) *versus* (11) *with* $\nu \in (0, 1)$ *fixed, and* $n \wedge p \to \infty$. *Let* T *denote either of the statistics* $\mathrm{SDP}_s(\hat{\mathbf{\Sigma}})$ *or* $\mathrm{MDP}_s(\hat{\mathbf{\Sigma}})$. *For some universal constant* $C_0 > 0$, *consider the test*

$$\phi = \left\{T \geq 1 + C_0\left[\sqrt{\frac{s^2}{n}\log(ep/s)} \vee \frac{s}{n}\log(ep/s)\right]\right\}.$$

The test ϕ *has asymptotic level* 0. *Moreover, it has asymptotic power* 1 *if*

$$\nu(1-\nu)\Delta\mu^\top \mathbf{\Sigma}^{-1}\Delta\mu \geq C_1\left[\sqrt{\frac{s^2}{n}\log\left(\frac{ep}{s}\right)} \vee \frac{s}{n}\log\left(\frac{ep}{s}\right)\right],$$

where $C_1 > 0$ *is a universal constant.*

The proof of Proposition 15 is a straightforward adaptation of the work of Berthet and Rigollet (2013a). The critical ingredient is the following inequality:

$$\lambda_s^{\max}(\mathbf{A}) \leq \mathrm{SDP}_s(\mathbf{A}) \leq \mathrm{MDP}_s(\mathbf{A}), \tag{58}$$

valid for any p.s.d. matrix $\mathbf{A}$ and any sparsity level s. Then, on the one hand, we find in Berthet and Rigollet (2013a), Proposition 6.2, that

$$\mathrm{MDP}_s(\hat{\mathbf{\Sigma}}) \leq 1 + C_1\left[\sqrt{\frac{s^2}{n}\log\left(\frac{ep}{s}\right)} \vee \frac{s}{n}\log\left(\frac{ep}{s}\right)\right]$$

with probability tending to one under the null (where the sample is i.i.d. standard normal); while, on the other hand, following what we did in the proof of Proposition 3, we find that $\lambda_s^{\max}(\hat{\mathbf{\Sigma}}) \geq 1 - \frac{1}{n} + C_2\nu(1-\nu)\Delta\mu^\top\mathbf{\Sigma}^{-1}\Delta\mu$ with probability tending to one under the alternative. From these two bounds and (58), we conclude.

REMARK. We notice the same $\sqrt{s}$ multiplicative factor and one wonders whether the added sophistication of these relaxations (SDP or MDP) is worth it; clearly, not from a theoretical standpoint, but it shows in our numerical experiments presented in Section 4.3. This is analogous to what Berthet and Rigollet (2013a) observed in the context of detecting a first principal component.

REMARK. We do not know of any analogous relaxations for the statistics presented in Section 3 for the case where the covariance matrix is unknown.

4.3. *Numerical experiments.*

We present here the result of some small-scale computer simulations meant to compare some of the computationally tractable tests introduced above. In all the experiments, we chose $p = 500$, $n = 200$, and the underlying covariance matrix $\mathbf{\Sigma}$ (whether assumed known or unknown) was taken to be the identity. The variables were generated with zero mean under both the null and the alternative. The difference in means, $\Delta\mu$, was chosen to be equally spread (in terms of energy) over all its nonzero coordinates. Specifically, we chose

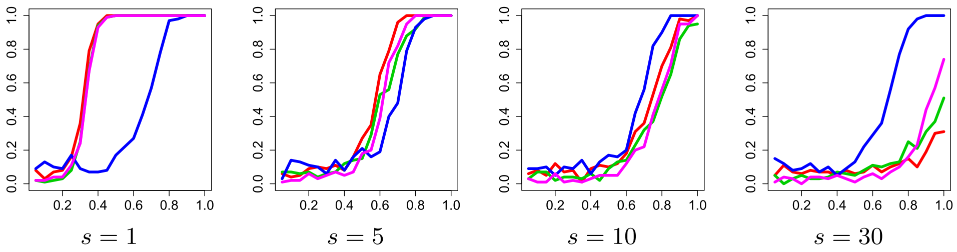

FIG. 1. *Power curves for the largest canonical variance* (*red*), *sth largest canonical variance* (*green*), *top sample eigenvalue* (*blue*) *and MDP* (*magenta*) *for various sparsity levels s as displayed. The level was set at* 0.05 *by simulation. On the horizontal axis is $A = \|\Delta\mu\|$, while on the vertical axis is the proportion of rejections* (*out of* 100 *repeats*).

$\Delta\mu_j = A\mathbb{1}_{\{j \le s\}}/\sqrt{s}$, where the sparsity s ranged over $\{1, 5, 10, 30\}$, while the amplitude $A = \|\Delta\mu\|$ varied the difficulty of the detection problem. We focused entirely on the symmetric model where $\nu = 1/2$. Each setting was repeated 100 times.

4.3.1. *Known covariance.* In this set of experiments, we assume that $\mathbf{\Sigma}$ is known to be the identity, and compared the maximum canonical variance (47), the s-largest canonical variance $\hat{\sigma}_{jj}/\sigma_{jj}$, the top sample eigenvalue (14), and the MDP statistics defined in (57). Note that the s-largest canonical variance and the MDP$_s$ both require knowledge of s. The results from these experiments are shown as power curves in Figure 1. Among other things, they confirm that the maximum canonical eigenvalue performs best when $\Delta\mu$ is really sparse whereas top sample eigenvalues performs best for less sparse signals; see Table 3. At least in the particular setting of these simulations, the combination of the maximum canonical variance and the top sample eigenvalue is competitive. An alternative—which we did not implement and is most relevant when $\mathbf{\Sigma}$ is diagonal—would be a higher-criticism approach applied to the canonical variances $\hat{\sigma}_{jj}/\sigma_{jj}$, which under the null are i.i.d. $\frac{1}{n}\chi^2_{n-1}$.

4.3.2. *Unknown covariance* (*kurtosis versus first moment*). In this set of experiments, we assume that $\mathbf{\Sigma}$ is unknown (even though it remains the identity), and compared the coordinatewise kurtosis and first absolute moment. We used the maximum canonical variance (whose calibration is only possible when $\mathbf{\Sigma}$ is known) as an oracle benchmark. Although we have a tighter control of the first moment under the null compared to the kurtosis, in these experiments the two behave very similarly (Figure 2).

5. Discussion. This paper leaves a number of interesting open problems regarding the theory of clustering under sparsity. We list a few of them below.

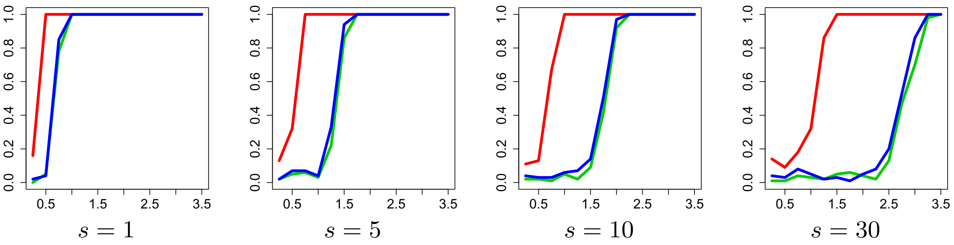

FIG. 2. *Power curves for the largest canonical variance* (*red*), *the largest canonical kurtosis* (*green*), *and the largest canonical absolute moment* (*blue*), *for various sparsity levels* s *as displayed. The level was set at* 0.05 *by simulation. On the horizontal axis is* $A = \|\Delta\mu\|$, *while on the vertical axis is the proportion of rejections* (*out of* 100 *repeats*).

The generalized likelihood ratio test (*GLRT*). The GLRT performs well in very many testing problems. In this paper, we simply focused on obtaining tests that achieved the various optimal detection rates, and we are curious to know whether the GLRT is one of them, at least in some of the settings. If anything, the GLRT is computationally very intensive in high dimensions, even more so than the moment-based methods analyzed here and, therefore, not practical, while heuristic implementations à la EM are very hard to analyze.

Theoretical adaptation to unknown sparsity. Throughout the paper, except in Section 4.1, we work under the assumption that the sparsity level s is known. This is in fact a mild assumption. Indeed, on the one hand, the problem is harder when s is unknown (since the set of alternatives is larger), so that the minimax lower bounds developed in the paper apply to the case where s is unknown. On the other hand, one can easily check that there is enough lee-way in the concentration bounds developed (under the null) for the various procedures that rely on s to accommodate a scan over $s \in [p]$.

Adaptation to unknown sparsity for computationally tractable procedures. We also note that the coordinatewise methods studied in Section 4.1 do not require the knowledge of the sparsity. When the population covariance matrix $\mathbf{\Sigma}$ is known, one can rely on the maximal canonical variance statistic (47) and the top eigenvalue statistic (47) together with a Bonferroni correction to simultaneously achieve the rates of Table 3 for all s.

Unknown mixing probability. We have assumed that ν is unknown. However, when the covariance matrix is unknown, it matters whether $\nu = 1/2$ or $\nu \neq 1/2$, for the proposed methods are different—based on the first absolute moment and the second signed moment, respectively. Let us focus on the coordinatewise methods introduced and studied in Section 4.1.2. For the detection problem, an easy way to adapt to situations where it is unknown whether $\nu = 1/2$ or $\nu \neq 1/2$ is to combine the tests based on T_1 and T_2 with a Bonferroni correction. For the variable selection problem, one can simply consider the union $\hat{J}_1 \cup \hat{J}_2$ of the variables selected by the two methods.

Mixture models with different covariance matrices. We have assumed everywhere in the paper that the two populations had the same covariance matrix. When this is not the case, assuming the two population covariance matrices are known (both under the null and under the alternative) does not seem as meaningful, and the case where they are unknown is more complex, and we speculate that more sophisticated methods that attempt to cluster the data into two groups (as the GLRT does) may be required. We note, however, that the procedure presented in Section 3.3 applies in exactly the same way to the special case where the population covariance matrices are diagonal—although the performance bound established in Proposition 12 is not valid.

Mixture models with more than two components. Suppose the mixture, under the alternative, has $K \geq 2$ components, with the kth component having mean μ_k and proportion ν_k, and consider for simplicity the case where the population covariance matrix is known to be the identity, both under the null and the alternative. Then, under the alternative,

$$\operatorname{Cov}(X) = \sum_{k=1}^{K} \nu_k(1-\nu_k)\mu_k\mu_k^\top + \sum_{1\leq k<\ell\leq K} \nu_k\nu_\ell\big(\mu_k\mu_\ell^\top + \mu_\ell\mu_k^\top\big) + \mathbf{I}.$$

In the general situation where the group means are affine independent, $\operatorname{Cov}(X)$ is a rank $K-1$ perturbation of the identity matrix. It is therefore natural to consider a test based on the top $K-1$ s-sparse eigenvalues of the sample covariance matrix. We note, though, that when K is fixed, the top s-sparse eigenvalue is still able to achieve the optimal detection rate. See Hsu and Kakade (2013) for related results in a nonsparse setting. Recently, Jin and Wang (2014) have also studied the case where the population covariance matrices are diagonal but unknown.

Computational issues. Computational considerations have lead a number of researchers to propose coordinatewise methods as we did in Section 4.1. In Section 4.2 we studied an SDP relaxation in the context of mixture detection when the covariance matrix is known to be the identity. It seems possible to extend this to the case of a general known covariance matrix. If anything, the suboptimal test based on (19) can be relaxed in the same exact way since it is based on computing a top sparse eigenvalue. And the same is true of the diagonal model. However, we do not know how to relax any of the tests considered in the case where the covariance matrix is unknown.

Acknowledgements. We thank two anonymous referees for suggestions that greatly improved the presentation.

SUPPLEMENTARY MATERIAL

Supplement to "Detection and feature selection in sparse mixture models" (DOI: 10.1214/16-AOS1513SUPP; .pdf). This supplement contains the proofs of the results.

REFERENCES

AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control* **AC-19** 716–723. MR0423716

AMINI, A. A. and WAINWRIGHT, M. J. (2009). High-dimensional analysis of semidefinite relaxations for sparse principal components. *Ann. Statist.* **37** 2877–2921. MR2541450

AZIZYAN, M., SINGH, A. and WASSERMAN, L. (2013). Minimax theory for high-dimensional Gaussian mixtures with sparse mean separation. *Neural Information Processing Systems* (*NIPS*).

AZIZYAN, M., SINGH, A. and WASSERMAN, L. (2015). Efficient sparse clustering of high-dimensional non-spherical Gaussian mixtures. In *AISTATS*.

BELKIN, M. and SINHA, K. (2010). Polynomial learning of distribution families. In 2010 *IEEE* 51*st Annual Symposium on Foundations of Computer Science FOCS* 2010 103–112. IEEE Computer Soc., Los Alamitos, CA. MR3024780

BERTHET, Q. and RIGOLLET, P. (2013a). Optimal detection of sparse principal components in high dimension. *Ann. Statist.* **41** 1780–1815. MR3127849

BERTHET, Q. and RIGOLLET , P. (2013b). Complexity theoretic lower bounds for sparse principal component detection. In *Conference on Learning Theory (COLT)* 1046–1066.

BICKEL, P. J. and LEVINA, E. (2004). Some theory of Fisher's linear discriminant function, "naive Bayes", and some alternatives when there are many more variables than observations. *Bernoulli* **10** 989–1010. MR2108040

BIRNBAUM, A., JOHNSTONE, I. M., NADLER, B. and PAUL, D. (2013). Minimax bounds for sparse PCA with noisy high-dimensional data. *Ann. Statist.* **41** 1055–1084. MR3113803

BOUCHERON, S., BOUSQUET, O., LUGOSI, G. and MASSART, P. (2005). Moment inequalities for functions of independent random variables. *Ann. Probab.* **33** 514–560. MR2123200

BRUBAKER, S. C. and VEMPALA, S. S. (2008). Isotropic PCA and affine-invariant clustering. In *Building Bridges*. *Bolyai Soc. Math. Stud.* **19** 241–281. Springer, Berlin. MR2484643

CAI, T. T., MA, Z. and WU, Y. (2013). Sparse PCA: Optimal rates and adaptive estimation. *Ann. Statist.* **41** 3074–3110. MR3161458

CAI, T., MA, Z. and WU, Y. (2015). Optimal estimation and rank detection for sparse spiked covariance matrices. *Probab. Theory Related Fields* **161** 781–815. MR3334281

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

CHAN, Y. and HALL, P. (2010). Using evidence of mixed populations to select variables for clustering very high-dimensional data. *J. Amer. Statist. Assoc.* **105** 798–809. MR2724862

CHANG, W.-C. (1983). On using principal components before separating a mixture of two multivariate normal distributions. *J. Roy. Statist. Soc. Ser. C* **32** 267–275. MR0770316

CHAUDHURI, K., DASGUPTA, S. and VATTANI, A. (1999). Learning mixtures of Gaussians using the k-means algorithm. Preprint. Available at arXiv:0912.0086.

CHEN, S. S., DONOHO, D. L. and SAUNDERS, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.* **20** 33–61. MR1639094

D'ASPREMONT, A., EL GHAOUI, L., JORDAN, M. I. and LANCKRIET, G. R. G. (2007). A direct formulation for sparse pca using semidefinite programming. *SIAM Rev.* **49** 434–448.

DONOHO, D. and JIN, J. (2009). Feature selection by higher criticism thresholding achieves the optimal phase diagram. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **367** 4449–4470. MR2546396

FAN, J. and PENG, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32** 928–961. MR2065194

FRIEDMAN, J. H. and MEULMAN, J. J. (2004). Clustering objects on subsets of attributes. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **66** 815–849. MR2102467

HARDT, M. and PRICE, E. (2015). Tight bounds for learning a mixture of two Gaussians [extended abstract]. In *STOC'*15—*Proceedings of the* 2015 *ACM Symposium on Theory of Computing* 753–760. ACM, New York. MR3388255

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, New York. MR2722294

HSU, D. and KAKADE, S. M. (2013). Learning mixtures of spherical Gaussians: Moment methods and spectral decompositions. In *ITCS'*13—*Proceedings of the* 2013 *ACM Conference on Innovations in Theoretical Computer Science* 11–19. ACM, New York. MR3385380

INGSTER, Y. I., POUET, C. and TSYBAKOV, A. B. (2009). Classification of sparse high-dimensional vectors. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **367** 4427–4448. MR2546395

JI, P. and JIN, J. (2012). UPS delivers optimal phase diagram in high-dimensional variable selection. *Ann. Statist.* **40** 73–103. MR3013180

JIN, J. (2009). Impossibility of successful classification when useful features are rare and weak. *Proc. Natl. Acad. Sci. USA* **106** 8859–8864. MR2520682

JIN, J., KE, Z. T. and WANG, W. (2015). Phase transitions for high dimensional clustering and related problems. Preprint. Available at arXiv:1502.06952.

JIN, J. and WANG, W. (2014). Important feature pca for high dimensional clustering. Preprint. Available at arXiv:1407.5241.

JOHNSTONE, I. M. and LU, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.* **104** 682–693. MR2751448

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

MALKOVICH, J. F. and AFIFI, A. (1973). On tests for multivariate normality. *J. Amer. Statist. Assoc.* **68** 176–179.

MALLAT, S. and ZHANG, Z. (1993). Matching pursuit with time-frequency dictionaries. *IEEE Trans. Image Process.* **41** 3397–3415.

MALLOWS, C. (1973). Some comments on cp. *Technometrics* **15** 661–675.

MARDIA, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika* **57** 519–530. MR0397994

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

MAUGIS, C. and MICHEL, B. (2011). A non asymptotic penalized criterion for Gaussian mixture model selection. *ESAIM Probab. Stat.* **15** 41–68. MR2870505

PAN, W. and SHEN, X. (2007). Penalized model-based clustering with application to variable selection. *J. Mach. Learn. Res.* **8** 1145–1164.

RAFTERY, A. E. and DEAN, N. (2006). Variable selection for model-based clustering. *J. Amer. Statist. Assoc.* **101** 168–178. MR2268036

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014

SRIVASTAVA, M. S. (1984). A measure of skewness and kurtosis and a graphical method for assessing multivariate normality. *Statist. Probab. Lett.* **2** 263–267. MR0777837

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

TROPP, J. A. (2004). Greed is good: Algorithmic results for sparse approximation. *IEEE Trans. Inform. Theory* **50** 2231–2242. MR2097044

TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer, New York. MR2724359

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes with Applications to Statistics*. Springer, New York. MR1385671

VERZELEN, N. (2012). Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electron*. *J*. *Stat*. **6** 38–90. MR2879672

VERZELEN, N. and ARIAS-CASTRO, E. (2016). Supplement to "Detection and feature selection in sparse mixture models." DOI:10.1214/16-AOS1513SUPP.

VU, V. Q. and LEI, J. (2012). Minimax rates of estimation for sparse pca in high dimensions. In *International Conference on Artificial Intelligence and Statistics* 1278–1286.

VU, V. Q. and LEI, J. (2013). Minimax sparse principal subspace estimation in high dimensions. *Ann*. *Statist*. **41** 2905–2947. MR3161452

WANG, S. and ZHU, J. (2008). Variable selection for model-based high-dimensional clustering and its application to microarray data. *Biometrics* **64** 440–448, 666. MR2432414

WITTEN, D. M. and TIBSHIRANI, R. (2010). A framework for feature selection in clustering. *J*. *Amer*. *Statist*. *Assoc*. **105** 713–726. MR2724855

XIE, B., PAN, W. and SHEN, X. (2008). Penalized model-based clustering with cluster-specific diagonal covariance matrices and grouped variables. *Electron*. *J*. *Stat*. **2** 168–212. MR2386092

ZHU, J. and HASTIE, T. (2004). Classification of gene microarrays by penalized logistic regression. *Biostatistics* **5** 427–443.

INRA, UMR 0729 MISTEA
2 PLACE VIALA, BÂT.29
F-34060 MONTPELLIER
FRANCE
E-MAIL: nicolas.verzelen@inra.fr

DEPARTEMENT OF MATHEMATICS
UNIVERSITY OF CALIFORNIA, SAN DIEGO
LA JOLLA, CALIFORNIA 92093-0112
USA
E-MAIL: eariasca@ucsd.edu