Optimal block-sparse PCA for high dimensional correlated samples[☆]

Romain Couillet[a], Matthew McKay[b]

[a] *Telecommunication department, Supélec, Gif sur Yvette, France*
[b] *ECE Department, Hong Kong University of Science and Technology*

Abstract

A new principal component analysis (PCA) method is proposed which is performed on a subset of blocks of consecutive entries of the population data vectors. This block-based dimensionality reduction introduces a trade-off by which the accuracy of the dominant eigenvector of the dimension-reduced sample covariance matrix is enhanced while some population entries are discarded. This scheme is particularly suited (but not restricted) to population eigenvectors with localized energy and rather sparse structures. Unlike many sparse PCA algorithms, the originality of our scheme lies in its providing an online selection of the subset of blocks which, in the large dimensional regime where both population and sample sizes grow large, provably ensures optimal alignment between population and sample eigenvectors. Moreover, our method inherently handles (a priori unknown) linear correlation between sample data.

Keywords: random matrix theory, PCA, sparsity, correlated samples.

1. Introduction

The advent of data mining over large dimensional datasets has called into question many classical statistical methods designed originally under the assumption of small dimensional population sizes (N) and large sample sizes (n). A prime example of current interest is that of principal component analysis (PCA), a ubiquitous dimensionality reduction and analysis tool in diverse applications spanning statistical finance (Ledoit and Wolf, 2003; Laloux et al., 2000), molecular biology and virology (Dahirel et al., 2011; Quadeer et al., 2014), signal processing and wireless communications (Moore et al., 2014), among numerous others.

Considering the simple "spiked" data model $Y = (I_N + \omega uu^*)^{\frac{1}{2}} X$ with $X \in \mathbb{R}^{N\times n}$ having standard Gaussian entries, $\omega > 0$ and $u \in \mathbb{R}^N$ of unit

[☆] Couillet's work is supported by the ANR RMT4GRAPH ANR-14-CE28-0006. McKay's work is supported by the Hong Kong Research Grants Council under 16206914

Email addresses: romain.couillet@supelec.fr (Romain Couillet), m.mckay@ust.hk (Matthew McKay)

norm, traditional PCA aims to estimate u by the vector $\hat{u}$ corresponding to the dominant eigenvector of the sample covariance matrix $n^{-1}YY^*$. It is well known that $|\hat{u}^*u| \to 1$ as $n \to \infty$ while ω and N are fixed, thus resulting (up to a rotation) in a consistent estimation of u. However, this consistency property no longer holds if both N and n grow simultaneously large or if ω decreases with n. In particular, if $N, n \to \infty$ with $N/n \to c \in (0, \infty)$ while ω is fixed, it is known from (Baik and Silverstein, 2006) that $|\hat{u}^*u|^2 \to (1 - c/\omega^2)(1 + c/\omega)^{-1}1_{\omega > \sqrt{c}}$. Thus, under such conditions, $\hat{u}$ is an inconsistent estimator for u to the extent that, if $\omega < \sqrt{c}$, $\hat{u}$ tends to be completely orthogonal to u.

The performance of PCA may be potentially improved by exploiting certain structural characteristics in the model. In particular, often it is the case, perhaps after a suitable basis transformation, that u is intrinsically sparse in the sense that its entries have few dominant values. If these values are found for indices in a set $\mathcal{I} \subset \{1, \ldots, N\}$, it is then advantageous to perform PCA on $Y_{\mathcal{I}} \in \mathbb{R}^{|\mathcal{I}| \times n}$, the observation matrix restricted to the rows indexed by $\mathcal{I}$. By reducing the ratio N/n to $|\mathcal{I}|/n$, the resulting dominant eigenvector $\hat{u}_{\mathcal{I}} \in \mathbb{R}^{|\mathcal{I}|}$ of $n^{-1}Y_{\mathcal{I}}Y_{\mathcal{I}}^*$ provides a better approximation for the vector u restricted to its dominant values. Obviously, smaller $|\mathcal{I}|$ implies more accurate $\hat{u}_{\mathcal{I}}$, but it also may lead to more energy in u being discarded. Hence, there is a natural trade-off between the need for accuracy in the estimation of dominant values and the need to preserve most of the population vector energy. Statistically, one may also see this as a variance-bias trade-off, where more aggressive index selection leads to a reduction in estimator variance but an increased bias. These observations have recently triggered an important wave of research aimed at (i) defining the necessary and sufficient conditions on the sparsity level of u and the relative growth of ω, N, n to ensure $|\hat{u}^*u| \to 1$ for some sparse estimate $\hat{u}$, (ii) determining the optimal convergence rate, and (iii) deriving algorithms that meet such a convergence rate. A seminal early work is (Johnstone and Lu, 2009), which considered a single-spiked model as introduced above, and this was further generalised to multiple spikes in (Paul and Johnstone, 2012). Remarkably, in major contrast to traditional PCA, it was shown that sparse methods can yield consistent estimates for N even as large as $\exp(n)$ for $\omega = O(1)$, provided that u is sufficiently sparse; specifically, $\limsup_N \sum_{i=1}^N |u_i|^q < \infty$ for some $q \in (0, 2)$. Several extensions mostly proposing alternative schemes with improved convergence rate properties were then provided in e.g., (Ma, 2013; Birnbaum et al., 2013; Yuan and Zhang, 2013; Shen et al., 2013; Cai et al., 2013; Wang et al., 2014). Other approaches have also been proposed (Jolliffe et al., 2003; Zou et al., 2006; Shen and Huang, 2008; Witten et al., 2009; d'Aspremont et al., 2007, 2008; Ulfarsson and Solo, 2008; Journée et al., 2010; Mairal et al., 2010; Candès et al., 2011) that turn PCA into an optimization problem and induce sparsity through the introduction of penalty functions or constraints. In the present work, we shall explicitly compare our proposed method against the legacy (although now recognized as suboptimal in practice) sparse PCA (SPCA) algorithm from (Johnstone and Lu, 2009) and the powerful truncated power (TPower) method from (Yuan and Zhang, 2013). Note that we compare with the TPower method as

a reference, as this was shown to yield competitive performance against a wide range of sparse PCA alternatives in (Yuan and Zhang, 2013), including (Zou et al., 2006; d'Aspremont et al., 2008; Shen and Huang, 2008; Witten et al., 2009; Journée et al., 2010; Mairal et al., 2010).

The above results, while providing optimal rates of convergence and schemes that achieve these rates, are based on various assumptions which may impede their performance for practical systems and for finite N, n. In particular, the schemes of (Johnstone and Lu, 2009; Paul and Johnstone, 2012) are based on selecting sets $\mathcal{I}$ of rows of Y that satisfy e.g., $i \in \mathcal{I} \Leftrightarrow [YY^*]_{ii} > \alpha_N$ for α_N given (rather imprecisely) in terms of orders of magnitude of N. An exact definition of an optimal choice for α_N (or equivalently of the cardinality of $|\mathcal{I}|$) is thus missing and it turns out in practice that a slight modification in such choices can entail important performance losses for finite dimensions. A similar comment applies to the rate-achieving scheme (Ma, 2013), and its single eigenvector predecessor (Yuan and Zhang, 2013). For example, in (Yuan and Zhang, 2013), successive truncations of the largest k absolute components of an iterated vector are used at the core of the TPower algorithm, but no precise estimate of the optimal k is provided. Another issue is linked to imposing sparsity constraints of the ℓ_q type that may not be met in practice, despite the fact that for many applications one may naturally expect the eigenvector energy to be localized in few places of the eigenvector.

With these remarks in mind, in this article, we take a different approach from the aforementioned works, adopting a so-called "G-estimation" strategy from random matrix theory which has been successfully applied in various contexts including portfolio optimization in finance (Laloux et al., 2000; Rubio et al., 2012; Yang et al., 2014), filter design in array processing (Mestre and Lagunas, 2008; Couillet et al., 2011; Rubio and Mestre, 2009), low-rank matrix denoising (Nadakuditi, 2014), and wireless communications (Couillet and Debbah, 2011), though it has yet to be applied to the application of sparse PCA.

We shall work under a setting for which $N, n \to \infty$ in such a way that $N/n \to c \in (0, \infty)$ and with $\omega = O(1)$. Under this setting, in general terms, the G-estimation strategy works in three major steps: (i) find a deterministic equivalent for the (random) objective function of interest, (ii) propose a sample-based consistent estimator for this objective function, and (iii) select on-line the relevant parameters to optimize the sample-based objective function. In the current context, the objective function of interest is the quantity $|\hat{u}_{\mathcal{I}}^{e*} u|$, where $\hat{u}_{\mathcal{I}}^{e*}$ is calculated based on a row selection approach similar in spirit to (Johnstone and Lu, 2009; Paul and Johnstone, 2012; Yuan and Zhang, 2013; Birnbaum et al., 2013), though here the rows are selected in a blockwise manner for which the possible sets $\mathcal{I} \subset \{1, \ldots, N\}$ under consideration are finitely many and each of cardinality $O(N)$ (the superscript 'e' stands for the zero-padded extension of $\hat{u}_{\mathcal{I}}$). Our approach, which is most appropriate for eigenvectors with localized energy, brings several important advantages. Most notably, it allows us to consistently evaluate on-line the optimal such set $\mathcal{I}$ for all large N, n, which maximizes the objective $|\hat{u}_{\mathcal{I}}^{e*} u|$. This approach is fully automated and requires no heuristic selection of thresholds nor other parameters. One limitation of the base method

is that it entails rather high computational complexity for large N; however, we show that this can be greatly reduced with a selection technique under certain assumptions. Practical comparisons versus the aforementioned schemes are also provided that show sometimes large performance improvements.

An additional contrasting feature of our approach relative to previous work is that we account for a possibly unknown linear correlation between the samples. Such unknown inter-sample correlation may be present in practice; for example, when dealing with time series of temporally correlated asset returns, or when processing biological data extracted from patients with certain temporal, geographical, or sequencing biases, which are difficult to remove through pre-processing. To account for unknown correlation effects among samples, we shall let the random Gaussian matrix X become $XT^{\frac{1}{2}}$ for some unknown deterministic positive definite matrix T that plays the role of a data correlation matrix.

Using the notation mentioned above, our main results can be more precisely summarized as follows. Assume the data model $Y = (I_N + \omega uu^*)^{\frac{1}{2}} XT^{\frac{1}{2}}$ with all parameters unknown, and let $\mathcal{K}$ denote a finite and deterministically-selected set of subsets $\mathcal{I} \subset \{1, \ldots, N\}$, such that for each $\mathcal{I}$, $|\mathcal{I}|/n$ converges in $(0, c]$. We propose a method that asymptotically determines the set $\mathcal{I} \in \mathcal{K}$ minimizing the angle between the vector $u \in \mathbb{R}^N$ and the vector $\hat{u}^e_{\mathcal{I}} \in \mathbb{R}^N$ composed of $\hat{u}_{\mathcal{I}}$ in the indices defined by $\mathcal{I}$ and zeros elsewhere. For practical purposes, $\mathcal{K}$ is then particularized to the ensemble of the 2^K unions of the regular division of $\{1, \ldots, N\}$ into K sets of size N/K each (assuming $N/K \in \mathbb{N}$), and then to the further union of such unions across M values $K_1, \ldots, K_M$ of K. To avoid the computationally expensive evaluation of the method over these $\sum_{m=1}^{M} 2^{K_m}$ sets, an algorithm (referred to as BlockPCA) is produced, which in favorable scenarios reduces this evaluation to only $\sum_{m=1}^{M} K_m$ sets and which is proved nonetheless to asymptotically achieve the minimum angle between u and $\hat{u}^e_{\mathcal{I}}$ across $\mathcal{K}$. Simulations are performed for two examples, the first with less sparsity than the second. The second scenario is based on a synthetic wavelet-sparse example introduced in (Johnstone and Lu, 2009). BlockPCA will be shown to largely outperform conventional sparse PCA alternatives in the former setting and to perform close to optimal in the latter, with the strong advantage of proposing an estimated optimal (blockwise) subset selection. As for (Johnstone and Lu, 2009; Yuan and Zhang, 2013), our focus is on the simplest case of a single spike model, leaving non-trivial extensions to multi-spike models for subsequent work.

Notation: In the remainder of the article, for a set $\mathcal{I}$, the notation $|\mathcal{I}|$ will stand for the cardinality of $\mathcal{I}$. Vectors are understood as column vectors with u^* the transposed version of u (taken as Hermitian transpose if complex variables are considered). The notation $\boldsymbol{\delta}_x$ is the Dirac mass at x. The support of a measure μ shall be denoted $\mathrm{supp}(\mu)$ and its Stieltjes transform $\int (t-z)^{-1} d\mu(t)$ valid for $z \in \mathbb{C} \setminus \mathrm{supp}(\mu)$ as $m_\mu(z)$. For Hermitian matrices $X \in \mathbb{R}^{N \times N}$, $\lambda_1(X) \geq \ldots \geq \lambda_N(X)$ denotes the eigenvalues of X in non-increasing order. The convergence sign "$\xrightarrow{\text{a.s.}}$" stands for almost sure (a.s.) convergence.

2. Model and assumptions

We consider the spiked data model,

$$Y = (I_N + \omega uu^*)^{\frac{1}{2}} XT^{\frac{1}{2}} \tag{1}$$

where we assume $u \in \mathbb{R}^N$ of unit norm, $\omega > 0$, $X \in \mathbb{R}^{N\times n}$ with independent real Gaussian entries with zero mean and unit variance,[1] and $T \in \mathbb{R}^{n\times n}$ deterministic nonnegative definite Hermitian such that $\frac{1}{n}\sum_{i=1}^n \boldsymbol{\delta}_{\lambda_i(T)} \to \nu$ weakly, with $\nu \neq \boldsymbol{\delta}_0$ a probability measure of compact support with $\limsup_n \operatorname{dist}(\lambda_i(T), \operatorname{supp}(\nu)) \to 0$, dist denoting Euclidean distance. We shall consider the regime where $N, n \to \infty$ with $N/n \to c \in (0, \infty)$.

For $\mathcal{I} \subset \{1, \ldots, N\}$, we denote $Y_{\mathcal{I}} \in \mathbb{R}^{|\mathcal{I}|\times n}$ the matrix of the rows of Y indexed by $\mathcal{I}$. The vector $u_{\mathcal{I}} \in \mathbb{R}^{|\mathcal{I}|}$ denotes equivalently the restriction to $\mathcal{I}$ of the vector u. The extended vector $u^e_{\mathcal{I}} \in \mathbb{R}^N$ is the vector with zero entries but for the indices in $\mathcal{I}$ that contain $u_{\mathcal{I}}$. We will also denote $c_r = r/n$, so in particular $c_{|\mathcal{I}|} = |\mathcal{I}|/n$. The ordered eigenvalue and eigenvector pairs of $Y_{\mathcal{I}}Y^*_{\mathcal{I}}/n$ are denoted $(\hat\lambda_{i,\mathcal{I}}, \hat u_{i,\mathcal{I}})$ with $\hat\lambda_{1,\mathcal{I}} \geq \ldots \geq \hat\lambda_{|\mathcal{I}|,\mathcal{I}}$. When $\mathcal{I} = \{1, \ldots, N\}$, the eigenvalue and eigenvector pairs will be simply denoted $(\hat\lambda_i, \hat u_i)$. Since we are mostly concerned with the pair $(\hat\lambda_{1,\mathcal{I}}, \hat u_{1,\mathcal{I}})$, we shall also use the shortcut notation $(\hat\lambda_{\mathcal{I}}, \hat u_{\mathcal{I}}) = (\hat\lambda_{1,\mathcal{I}}, \hat u_{1,\mathcal{I}})$.

Following the ideas from (Johnstone and Lu, 2009) and subsequent works on sparse PCA, our objective is to estimate the vector u by $\hat u^e_{\mathcal{I}}$ for some conveniently selected set $\mathcal{I}$. The metric $F_{\mathcal{I}}$ to maximize across $\mathcal{I}$ is given by the random variable

$$F_{\mathcal{I}} \triangleq \omega\, |u^* \hat u^e_{\mathcal{I}}|^2 \tag{2}$$

with minimum at zero and maximum at ω.

The performance measure $F_{\mathcal{I}}$ is maximized for some $\mathcal{I} \subset \{1, \ldots, N\}$, hence for one out of 2^N such sets (with probability one). To avoid the prohibitive procedure of testing all sets, sparse PCA approaches often propose to only retain those indices i for which $|u_i|$ is maximum. This, however, comes with several practical problems. First, to determine which entries of u have the largest $|u_i|$, the natural procedure is to select those indices i for which the diagonal terms $[YY^*]_{ii}$ are maximal; this is the original idea of (Johnstone and Lu, 2009). But, by the central limit theorem, $n^{-1}[YY^*]_{ii} = 1 + \omega|u_i|^2 + O_p(n^{-\frac{1}{2}})$ so only those indices i for which $\limsup_n n^{\frac{1}{2}}|u_i|^2 > 0$ can be expected to be effectively recovered. Since $\|u\|^2 = 1$, these indices have cardinality at most $O(n^{\frac{1}{2}})$, thus forming a vanishingly small proportion of $\{1, \ldots, N\}$. Practical settings in which the energy of u is concentrated in a non-vanishing proportion of its entries are thus ruled out. Second, even if for some L, the L largest $|u_i|$ could be perfectly recovered, the choice of L is still left to the practitioner, with only approximate

[1] Complex variables may also be considered, with all results applying verbatim.

rules of thumb provided in the works discussed previously. In this paper, we aim to address these issues.

Our starting point is to observe that, by definition of $Y_{\mathcal{I}}$ for a set $\mathcal{I} \subset \{1, \ldots, N\}$ and by the Gaussianity of X,

$$Y_{\mathcal{I}} \sim \left(I_{|\mathcal{I}|} + \omega u_{\mathcal{I}} u_{\mathcal{I}}^*\right)^{\frac{1}{2}} X_{\mathcal{I}} T^{\frac{1}{2}} \tag{3}$$

$$= \left(I_{|\mathcal{I}|} + \Omega_{\mathcal{I}} \frac{u_{\mathcal{I}}}{\|u_{\mathcal{I}}\|} \frac{u_{\mathcal{I}}^*}{\|u_{\mathcal{I}}\|}\right)^{\frac{1}{2}} X_{\mathcal{I}} T^{\frac{1}{2}} \tag{4}$$

where we have defined

$$\Omega_{\mathcal{I}} \triangleq \omega \|u_{\mathcal{I}}\|^2 \tag{5}$$

and where "$\sim$" indicates equality in distribution. This unfolds from noticing that every column of the matrix $(I_N + \omega u u^*)^{\frac{1}{2}} X$ restricted to the rows indexed by $\mathcal{I}$ is a zero mean Gaussian vector with covariance $I_{|\mathcal{I}|} + \omega u_{\mathcal{I}} u_{\mathcal{I}}^*$.

Of particular importance is that, upon selecting the indices $\mathcal{I}$, the effective model (4) is still a spiked model having the same form as (1), but with the spiked eigenvector u now replaced with $\frac{u_{\mathcal{I}}}{\|u_{\mathcal{I}}\|}$ and the spiked eigenvalue ω replaced with $\Omega_{\mathcal{I}}$. An important property, which follows from (5), is that for any two *disjoint* sets $\mathcal{I}_1, \mathcal{I}_2 \subset \{1, \ldots, N\}$, we have

$$\Omega_{\mathcal{I}_1 \cup \mathcal{I}_2} = \Omega_{\mathcal{I}_1} + \Omega_{\mathcal{I}_2}. \tag{6}$$

Hence, if one can estimate the effective spike eigenvalue of individual disjoint block selections, $\Omega_{\mathcal{I}_i}$, $i = 1, 2, \ldots$, then these can be combined to estimate the effective spike eigenvalues of larger selections. This observation will be important for developing an efficient computational algorithm in the following.

3. Results

We provide a new block-based selection method for sparse PCA. The idea is as follows. We first show that for each non random set $\mathcal{I}$ of cardinality $O(N)$, it is possible to asymptotically evaluate $F_{\mathcal{I}}$ with arbitrary accuracy, without any knowledge of ω, u, and T. The estimate will be denoted $\hat{F}_{\mathcal{I}}$. Moreover, for a partition of the N indices into K blocks of size $N/K \in \mathbb{N}$, with K fixed and independent of N, this result holds uniformly over the set $\mathcal{K}$ of 2^K possible unions of these blocks. This ensures that, by comparing $\hat{F}_{\mathcal{I}}$ for all $\mathcal{I} \in \mathcal{K}$, we may estimate $\mathcal{I}_{\text{opt}} = \text{argmax}_{\mathcal{I} \in \mathcal{K}} F_{\mathcal{I}}$ to arbitrary accuracy. Denoting the corresponding estimate as $\hat{\mathcal{I}}_{\text{opt}}$, for K not too small, this gives in turn an accurate estimate $\hat{u}^e_{\hat{\mathcal{I}}_{\text{opt}}}$ for u. To reduce the computational complexity of the proposed algorithm for large K, we propose an efficient greedy approach which reduces the search from 2^K sets to only K. Our proposed method is summarized in Algorithm 1. An accompanying important theoretical result is provided in Theorem 1, which essentially states that, under mild conditions and

with $\hat{\mathcal{I}}_{\text{opt}}$ now the output of Algorithm 1, $F_{\hat{\mathcal{I}}_{\text{opt}}} - F_{\mathcal{I}_{\text{opt}}} \xrightarrow{\text{a.s.}} 0$ as $N, n \to \infty$ with $N/n \to c > 0$.

Throughout the paper, we will let $\mathcal{I} \subset \{1, \ldots, N\}$ be a *deterministically selected* set (that is, not depending on X) with cardinality $|\mathcal{I}|$ such that $c_{|\mathcal{I}|} \to \bar{c} \in (0, c]$. For such $\mathcal{I}$, we have

$$F_{\mathcal{I}} = \omega \left| u^* \hat{u}^e_{\mathcal{I}} \right|^2 \;=\; \Omega_{\mathcal{I}} \left| \frac{u^*_{\mathcal{I}}}{\|u_{\mathcal{I}}\|} \hat{u}_{\mathcal{I}} \right|^2 .$$

We will present a consistent estimator for this quantity, first for the case $T = \sigma^2 I$ and subsequently for more general T.

3.1. Consistent estimators for $F_{\mathcal{I}}$

For $T = \sigma^2 I$, we require the following result, which follows from (4) and as a consequence of (Baik and Silverstein, 2006, Theorems 1.1 and 1.2) and (Paul, 2007, Theorems 1, 2, and 4):

Proposition 1. *Let $T = \sigma^2 I$ with $\sigma > 0$. For deterministically-selected $\mathcal{I}$ as above, if*

$$\liminf_n \Omega_{\mathcal{I}} > \sqrt{\bar{c}} \tag{7}$$

we have, as $n \to \infty$,

$$\hat{\lambda}_{\mathcal{I}} - \rho_{\mathcal{I}} \xrightarrow{\text{a.s.}} 0, \qquad \hat{\lambda}_{2,\mathcal{I}} \xrightarrow{\text{a.s.}} \sigma^2 \left(1 + \sqrt{\bar{c}}\right)^2 \tag{8}$$

$$F_{\mathcal{I}} - \bar{F}_{\mathcal{I}} \xrightarrow{\text{a.s.}} 0, \tag{9}$$

where

$$\rho_{\mathcal{I}} = \sigma^2 \left(1 + \Omega_{\mathcal{I}} + \bar{c} \frac{1 + \Omega_{\mathcal{I}}}{\Omega_{\mathcal{I}}}\right)$$
$$\bar{F}_{\mathcal{I}} = \frac{\Omega^2_{\mathcal{I}} - \bar{c}}{\Omega_{\mathcal{I}} + \bar{c}} .$$

If (7) is not met, then there exists at least one subsequence over which

$$\hat{\lambda}_{\mathcal{I}}, \hat{\lambda}_{2,\mathcal{I}} \xrightarrow{\text{a.s.}} \sigma^2 \left(1 + \sqrt{\bar{c}}\right)^2$$
$$F_{\mathcal{I}} \xrightarrow{\text{a.s.}} 0.$$

Proposition 1 provides a deterministic equivalent for $\hat{\lambda}_{\mathcal{I}}$ and $F_{\mathcal{I}}$ in the large n limit. The right-hand quantity in (7) defines a well-known "phase-transition" threshold: if the effective spiked eigenvalue $\Omega_{\mathcal{I}}$ is above this, then the top sample eigenvalue $\hat{\lambda}_{\mathcal{I}}$ (for the $\mathcal{I}$-truncated model) converges to a value $\rho_{\mathcal{I}} > \sigma^2(1+\sqrt{\bar{c}})^2$, and the corresponding top sample eigenvector $\hat{u}_{\mathcal{I}}$ traces an asymptotic angle with $u_{\mathcal{I}}$ as determined by $\bar{F}_{\mathcal{I}}$. Note that $\hat{\lambda}_{\mathcal{I}}$ is asymptotically distinguishable

from the second largest sample eigenvalue $\hat{\lambda}_{2,\mathcal{I}}$ which converges to $\sigma^2(1+\sqrt{\bar{c}})^2$, representing the right-hand edge of the limiting empirical eigenvalue distribution of $Y_{\mathcal{I}}Y_{\mathcal{I}}^*/n$ (the Marchenko-Pastur law in this case). On the other hand, if $\Omega_{\mathcal{I}}$ falls below this threshold, as will be the case when $\|u_{\mathcal{I}}\|^2$ is small for the selected subset $\mathcal{I}$, neither the top sample eigenvalue nor the top sample eigenvector become informative, and in this case $F_{\mathcal{I}}$ is asymptotically zero.

Defining

$$\hat{\sigma}^2 \triangleq \frac{1}{N-1}\sum_{i=2}^{N}\hat{\lambda}_i = \frac{1}{N-1}\left(\mathrm{tr}(YY^*)/n - \hat{\lambda}_1\right)$$

(recall that $\hat{\lambda}_i = \hat{\lambda}_{i,\{1,\ldots,N\}}$), we also have:

$$\hat{\sigma}^2 \xrightarrow{\text{a.s.}} \sigma^2 \ ,$$

providing a consistent estimate for the parameter σ^2, if unknown.

In order to decide whether (7) is met, one may verify that $\hat{\lambda}_{\mathcal{I}} > \hat{\sigma}^2(1+\sqrt{c_{|\mathcal{I}|}})^2(1+\varepsilon)$ for some $\varepsilon > 0$ (Zeng and Liang, 2009; Bianchi et al., 2011), or alternatively that $\hat{\lambda}_{\mathcal{I}} - \hat{\lambda}_{2,\mathcal{I}} > \varepsilon$. Both methods are equivalent and we shall use the former in the following.

According to these remarks and the results of Proposition 1, defining[2]

$$\hat{\Omega}_{\mathcal{I}} \triangleq \frac{1}{2}\left(\hat{\sigma}^{-2}\hat{\lambda}_{\mathcal{I}} - 1 - c_{|\mathcal{I}|} + \sqrt{(c_{|\mathcal{I}|} - 1 - \hat{\sigma}^{-2}\hat{\lambda}_{\mathcal{I}})^2 - 4c_{|\mathcal{I}|}}\right) 1_{\{\hat{\lambda}_{\mathcal{I}} > \hat{\sigma}^2(1+\sqrt{c_{|\mathcal{I}|}})^2(1+\varepsilon)\}}$$

$$\hat{F}_{\mathcal{I}} \triangleq \frac{(\hat{\Omega}_{\mathcal{I}})^2 - c_{|\mathcal{I}|}}{\hat{\Omega}_{\mathcal{I}} + c_{|\mathcal{I}|}} 1_{\{\hat{\lambda}_{\mathcal{I}} > \hat{\sigma}^2(1+\sqrt{c_{|\mathcal{I}|}})^2(1+\varepsilon)\}}$$

we have that, if $\liminf_n F_{\mathcal{I}} > 0$, then there exists $\varepsilon > 0$, sufficiently small, such that

$$\hat{\Omega}_{\mathcal{I}} - \Omega_{\mathcal{I}} \xrightarrow{\text{a.s.}} 0, \qquad \hat{F}_{\mathcal{I}} - F_{\mathcal{I}} \xrightarrow{\text{a.s.}} 0. \tag{10}$$

Hence, $\hat{F}_{\mathcal{I}}$ provides a consistent estimator for $F_{\mathcal{I}}$ for the case $T = \sigma^2 I$.

Now, consider the case of more general T. The result in Proposition 1 has a (albeit more complicated) generalization for arbitrary T. This follows from (4) and as a consequence of (Benaych-Georges and Nadakuditi, 2011, Theorem 2.8) and (Couillet and Hachem, 2013, Theorems 1 and 2):

Proposition 2. *Let $\mathcal{I}$ be as in Proposition 1. Then, as $n \to \infty$, $\frac{1}{|\mathcal{I}|}\sum_{i=1}^{|\mathcal{I}|}\boldsymbol{\delta}_{\hat{\lambda}_{i,\mathcal{I}}} \xrightarrow{\text{a.s.}}$ $\mu_{\bar{c}}$ weakly, where $\mu_{\bar{c}}$ is a probability measure with compact support* $\mathrm{supp}(\mu_{\bar{c}})$

[2]The formula for $\hat{\Omega}_{\mathcal{I}}$ results from inverting the relation $\hat{\lambda}_{\mathcal{I}} = \hat{\sigma}^2(1+\hat{\Omega}_{\mathcal{I}}+c_{|\mathcal{I}|}\frac{1+\hat{\Omega}_{\mathcal{I}}}{\hat{\Omega}_{\mathcal{I}}})+o(1)$ and retaining the unique positive solution.

which only depends on $\bar{c}$ and ν. (Recall that ν is the weak limit of $\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\delta}_{\lambda_i(T)}$.) Moreover, if

$$\liminf_n \Omega_{\mathcal{I}} > -\left(\lim_{x\downarrow \mathrm{sup}(\mathrm{supp}(\mu_{\bar{c}}))} x m_{\mu_{\bar{c}}}(x) + 1\right)^{-1} \tag{11}$$

where $m_{\mu_{\bar{c}}}(x) \triangleq \int (t-x)^{-1}\mu_{\bar{c}}(dt)$ is the restriction to the real axis of the Stieltjes transform of $\mu_{\bar{c}}$, then

$$\hat{\lambda}_{\mathcal{I}} - \rho_{\mathcal{I}} \xrightarrow{\text{a.s.}} 0 \tag{12}$$

$$\hat{\lambda}_{2,\mathcal{I}} \xrightarrow{\text{a.s.}} \mathrm{sup}(\mathrm{supp}(\mu_{\bar{c}})) \tag{13}$$

$$F_{\mathcal{I}} - \bar{F}_{\mathcal{I}} \xrightarrow{\text{a.s.}} 0 \tag{14}$$

where $\rho_{\mathcal{I}} > \mathrm{sup}(\mathrm{supp}(\mu_{\bar{c}}))$ is the unique solution (for all n large) to the equation in $x \in (\mathrm{sup}(\mathrm{supp}(\mu_{\bar{c}})), \infty)$

$$\Omega_{\mathcal{I}} = -\left(x m_{\mu_{\bar{c}}}(x) + 1\right)^{-1}$$

and where

$$\bar{F}_{\mathcal{I}} = -\frac{m_{\mu_{\bar{c}}}(\rho_{\mathcal{I}})}{m_{\mu_{\bar{c}}}(\rho_{\mathcal{I}}) + \rho_{\mathcal{I}} m'_{\mu_{\bar{c}}}(\rho_{\mathcal{I}})}.$$

If (11) is not met, then there exists at least one subsequence over which

$$\hat{\lambda}_{\mathcal{I}}, \hat{\lambda}_{2,\mathcal{I}} \xrightarrow{\text{a.s.}} \mathrm{sup}(\mathrm{supp}(\mu_{\bar{c}}))$$
$$F_{\mathcal{I}} \xrightarrow{\text{a.s.}} 0.$$

For the special case, $T = \sigma^2 I$, $\mu_{\bar{c}}$ takes the form of the Marchenko-Pastur law, and $\rho_{\mathcal{I}}$ and $\bar{F}_{\mathcal{I}}$ admit the explicit forms given in Proposition 1. Note also that the model assumption $\mathrm{dist}(\lambda_i(T), \nu) \to 0$, introduced in Section 2, ensures that $\hat{\lambda}_{2,\mathcal{I}} \to \mathrm{sup}(\mathrm{supp}(\mu_{\bar{c}}))$ (i.e., no other eigenvalue than $\hat{\lambda}_{\mathcal{I}}$ may escape the limiting support of $\mu_{\bar{c}}$); if relaxed, several eigenvalues may escape the limiting support, making it impossible to identify the eigenvector $\hat{u}_{i,\mathcal{I}}$ having a non-trivial alignment to $u_{\mathcal{I}}$.

While directly estimating the right-hand side of (11) is difficult in general, we instead observe that the condition (11) is equivalent, as for the previous case, to there existing $\varepsilon > 0$ such that $\hat{\lambda}_{\mathcal{I}} - \hat{\lambda}_{2,\mathcal{I}} > \varepsilon$ for all large n a.s. For the estimation of $F_{\mathcal{I}}$, note that from Proposition 2 (by weak convergence and the asymptotic absence of eigenvalues beyond $\mathrm{sup}(\mathrm{supp}(\mu_{\bar{c}}))$), we have that, for every $x > \mathrm{sup}(\mathrm{supp}(\mu_{\bar{c}}))$, and every deterministically-selected set $\mathcal{I}$ such that $|\mathcal{I}|/n \to \bar{c}$,

$$\hat{m}_{\mathcal{I}}(x) \triangleq \frac{1}{|\mathcal{I}| - 1}\sum_{i=2}^{|\mathcal{I}|} \frac{1}{\hat{\lambda}_{i,\mathcal{I}} - x} \xrightarrow{\text{a.s.}} m_{\mu_{\bar{c}}}(x) \tag{15}$$

$$\hat{m}'_{\mathcal{I}}(x) = \frac{1}{|\mathcal{I}| - 1}\sum_{i=2}^{|\mathcal{I}|} \frac{1}{(\hat{\lambda}_{i,\mathcal{I}} - x)^2} \xrightarrow{\text{a.s.}} m'_{\mu_{\bar{c}}}(x). \tag{16}$$

Thus, denoting

$$\hat{\Omega}_{\mathcal{I}} \triangleq -\left(\hat{\lambda}_{\mathcal{I}}\hat{m}_{\mathcal{I}}(\hat{\lambda}_{\mathcal{I}})+1\right)^{-1} 1_{\{\hat{\lambda}_{\mathcal{I}}-\hat{\lambda}_{2,\mathcal{I}}>\varepsilon\}} \tag{17}$$

$$\hat{F}_{\mathcal{I}} \triangleq -\frac{\hat{m}_{\mathcal{I}}(\hat{\lambda}_{\mathcal{I}})}{\hat{m}_{\mathcal{I}}(\hat{\lambda}_{\mathcal{I}})+\hat{\lambda}_{\mathcal{I}}\hat{m}'_{\mathcal{I}}(\hat{\lambda}_{\mathcal{I}})} 1_{\{\hat{\lambda}_{\mathcal{I}}-\hat{\lambda}_{2,\mathcal{I}}>\varepsilon\}} \tag{18}$$

we conclude that, if $\liminf_n F_{\mathcal{I}} > 0$, then there exists $\varepsilon > 0$ such that

$$\hat{\Omega}_{\mathcal{I}} - \Omega_{\mathcal{I}} \xrightarrow{\text{a.s.}} 0, \qquad \hat{F}_{\mathcal{I}} - F_{\mathcal{I}} \xrightarrow{\text{a.s.}} 0. \tag{19}$$

Hence, we have a consistent estimator of $F_{\mathcal{I}}$ for general T.

3.2. Uniform consistency

The results (19) and (10) hold uniformly on a finite family of sets $\mathcal{K} = \{\mathcal{I}_1, \ldots, \mathcal{I}_{|\mathcal{K}|}\}$ such that, for each $1 \leq j \leq |\mathcal{K}|$, $|\mathcal{I}_j|/n$ converges in $(0, c]$. So in particular, if $\liminf_n \max_j F_{\mathcal{I}_j} > 0$, then there exists $\varepsilon > 0$ such that

$$\max_{\mathcal{I}\in\mathcal{K}} \hat{F}_{\mathcal{I}} - \max_{\mathcal{I}\in\mathcal{K}} F_{\mathcal{I}} \xrightarrow{\text{a.s.}} 0. \tag{20}$$

This is important as it implies that one may find the optimal set $\mathcal{I}_{\text{opt}} \subset \{1, \ldots, N\}$ among a finite family $\mathcal{K}$ which maximizes the "true" objective $F_{\mathcal{I}}$ by finding the set which maximizes the sample-based objective $\hat{F}_{\mathcal{I}}$. However, while the optimization of $F_{\mathcal{I}}$ in (20) is completely data driven, it entails high complexity when $|\mathcal{K}|$ is not too small as it requires to perform $|\mathcal{K}|$ eigen-decompositions. We will demonstrate in the following that with a pragmatic construction of $\mathcal{K}$, this complexity can be greatly reduced.

In the subsequent development, we will focus attention on general T and exploit results given in Proposition 2, recalling that it absorbs Proposition 1 as a particular case. It is important to point out, however, that the case $T = \sigma^2 I_n$ has the practical advantage in that the computation of $\hat{F}_{\mathcal{I}}$ does not require the evaluation of $\hat{m}_{\mathcal{I}}(\hat{\lambda}_{\mathcal{I}})$ which is a function of *all* eigenvalues of $Y_{\mathcal{I}}Y_{\mathcal{I}}^*/n$. Here, only $\hat{\lambda}_{\mathcal{I}}$ is needed, which can be estimated efficiently using classical iterative power methods, without the need to compute the entire eigen-decomposition.

3.3. Reducing the search space

For simplicity of exposition, let us assume that N is taken to be a (large) multiple of some fixed integers $K_1, \ldots, K_M$. For each $K \in \{K_1, \ldots, K_M\}$, define $\mathcal{K}^K \triangleq \{\mathcal{I}_1^K, \ldots, \mathcal{I}_{2^K}^K\}$ where $\mathcal{I}_j^K \triangleq \cup_{i\in D_j} \mathcal{J}_i^K$, where $\mathcal{J}_i^K \triangleq \{1 + (i-1)N/K, \ldots, iN/K\}$ and $D_j \subset \{1, \ldots, K\}$ is the index set of components equal to one in the base-2 expression of j. A convenient choice for $\mathcal{K}$ is then to take $\mathcal{K} = \cup_{j=1}^M \mathcal{K}^{K_j}$, which contains all possible combinations of regular splittings of $\{1, \ldots, N\}$ in K sets of size N/K each, with K being any one of $K_1, \ldots, K_M$.

For $K \in \{K_1, \ldots, K_M\}$, let $s_1, \ldots, s_K$ denote the permutation of $\{1, \ldots, K\}$ satisfying

$$\mathcal{J}_{s_1}^K \succeq \ldots \succeq \mathcal{J}_{s_K}^K \tag{21}$$

in which the ordering relation "$\succeq$" is defined for two sets $\mathcal{I}, \mathcal{I}'$ by

$$\mathcal{I} \succeq \mathcal{I}' \Leftrightarrow \Omega_{\mathcal{I}} \geq \Omega_{\mathcal{I}'}. \tag{22}$$

With this, we may further define the K sets $\mathcal{I}^K_{1,\succeq}, \ldots, \mathcal{I}^K_{K,\succeq}$ of $\mathcal{K}^K$ by $\mathcal{I}^K_{i,\succeq} \triangleq \cup_{j=1}^{i} \mathcal{J}^K_{s_j}$, such that $\mathcal{I}^K_{1,\succeq} = \mathcal{J}^K_{s_1}$, $\mathcal{I}^K_{2,\succeq} = \mathcal{J}^K_{s_1} \cup \mathcal{J}^K_{s_2}$, and so on.

With the definitions above, upon recalling (6), we see that for a given K and i, the maximum of $\Omega_{\mathcal{I}}$, for $\mathcal{I} \in \mathcal{K}^K$ and $|\mathcal{I}| = iN/K$, is met for $\mathcal{I} = \mathcal{I}^K_{i,\succeq}$. We will show now that this choice also leads to the asymptotic optimum of $F_{\mathcal{I}}$. That is, for a given K, the optimal set within $\mathcal{K}^K$ is one of $\mathcal{I}^K_{1,\succeq}, \ldots, \mathcal{I}^K_{K,\succeq}$. It then follows immediately that the required task of estimating $F_{\mathcal{I}}$ over all $\mathcal{I} \in \mathcal{K}$ is reduced to simply estimating $F_{\mathcal{I}}$ for the sets $\mathcal{I} \in \cup_{j=1}^{M} \{\mathcal{I}^{K_j}_{1,\succeq}, \ldots, \mathcal{I}^{K_j}_{K_j,\succeq}\}$. Computationally, this is important, since only $\sum_{j=1}^{M} K_j$ sets must be considered, rather than $\sum_{j=1}^{M} 2^{K_j}$. Key to the result is a technical lemma, Lemma 1, established in Section 6, which implies that for each deterministically-selected $\mathcal{I}$ satisfying (11), the application

$$\Omega_{\mathcal{I}} \mapsto \bar{F}_{\mathcal{I}} = -\frac{m_{\mu_{\bar{c}}}(\rho_{\mathcal{I}})}{m_{\mu_{\bar{c}}}(\rho_{\mathcal{I}}) + \rho_{\mathcal{I}} m'_{\mu_{\bar{c}}}(\rho_{\mathcal{I}})} \tag{23}$$

is continuous and monotonically increasing (recall that $\rho_{\mathcal{I}} = \rho_{\mathcal{I}}(\Omega_{\mathcal{I}})$, as defined in Proposition 2). Also note that, for all $\mathcal{I} \in \mathcal{K}^K$ such that $|\mathcal{I}| = iN/K$, we have $|\mathcal{I}|/n \to \bar{c} = ci/K$ and thus the right-hand side of (23) becomes

$$\bar{F}_{\mathcal{I}} = -\frac{m_{\mu_{ci/K}}(\rho_{\mathcal{I}})}{m_{\mu_{ci/K}}(\rho_{\mathcal{I}}) + \rho_{\mathcal{I}} m'_{\mu_{ci/K}}(\rho_{\mathcal{I}})}$$

which only differs for various $\mathcal{I}$'s by the value of the quantity $\rho_{\mathcal{I}}$ (equivalently, by the value of $\Omega_{\mathcal{I}}$). As a consequence, from Proposition 2, the mapping

$$\begin{aligned} \left\{\mathcal{I} \in \mathcal{K}^K,\ |\mathcal{I}| = iN/K\right\} &\to \mathbb{R}^+ \\ \mathcal{I} &\mapsto F_{\mathcal{I}} \end{aligned}$$

is asymptotically maximal for $\mathcal{I}$ such that $\Omega_{\mathcal{I}}$ is arbitrarily close to $\Omega_{\mathcal{I}^K_{i,\succeq}}$.[3] By continuity of the mapping (23), it is then asymptotically sufficient to consider the set $\mathcal{I} = \mathcal{I}^K_{i,\succeq}$ among all sets of $\left\{\mathcal{I} \in \mathcal{K}^K,\ |\mathcal{I}| = iN/K\right\}$.

3.4. Estimating the index sets $\mathcal{I}^K_{i,\succeq}, i = 1, \ldots, K$

Having established the asymptotic sufficiency of restricting attention to the reduced collection of sets $\mathcal{I}^K_{i,\succeq}, i = 1, \ldots, K$, the remaining problem is to retrieve

[3] Note that the mapping may (infinitely often) not be maximal for $\mathcal{I} = \mathcal{I}^K_{i,\succeq}$ as $\Omega_{\mathcal{I}^K_{i,\succeq}}$ might be infinitely often within $o(1)$ of some other $\Omega_{\mathcal{I}}$. Nonetheless, $\mathcal{I}^K_{i,\succeq}$ is always an asymptotically appropriate candidate for the maximization of $|u^*_{\mathcal{I}} \hat{u}_{\mathcal{I}}|^2$.

these sets from $\mathcal{K}^K$. Given that $\mathcal{I}^K_{i,\succeq} = \cup_{k=1}^{i} \mathcal{J}^K_{s_k}$ by definition, one requires the successive indices $s_1, \ldots, s_K$, which are determined from $\Omega_{\mathcal{J}_i^K}$ for $i \in \{1, \ldots, K\}$. In terms of estimating these quantities, according to Proposition 2, for large N, if all $\Omega_{\mathcal{J}_i^K}$ were sufficiently large (i.e., they met the condition (11)), then it would be sufficient to evaluate $\hat{\Omega}_{\mathcal{J}_1^K}, \ldots, \hat{\Omega}_{\mathcal{J}_K^K}$, as these would form consistent estimates of $\Omega_{\mathcal{J}_1^K}, \ldots, \Omega_{\mathcal{J}_K^K}$. This, in turn, would lead to consistent estimates of the indices $s_1, \ldots, s_K$, and thus of the desired index sets $\mathcal{I}^K_{i,\succeq}$, $i = 1, \ldots, K$. This is an ideal situation however, and in practice the problem is more complicated. If for some block indices i, $\Omega_{\mathcal{J}_i^K}$ is positive but falls below the phase transition threshold in (11), then $\hat{\Omega}_{\mathcal{J}_i^K}$ will be asymptotically zero, thus forming an *inconsistent* estimate of $\Omega_{\mathcal{J}_i^K}$. This, in turn, will lead to classification errors in the estimation of $s_1, \ldots, s_K$. In the following, we outline a method for dealing with this important issue.

We first describe the basic idea. Consider a scenario for which for some subset of block indices $\mathcal{M}_1^K \subset \{1, \ldots, K\}$, $\hat{\Omega}_{\mathcal{J}_i^K} > 0$ for all $i \in \mathcal{M}_1^K$ only. By ordering these values, one can consistently estimate $s_1, \ldots, s_{|\mathcal{M}_1^K|}$ and consequently $\mathcal{I}^K_{1,\succeq}, \ldots, \mathcal{I}^K_{|\mathcal{M}_1^K|,\succeq}$. Importantly, while $\mathcal{I}^K_{1,\succeq}$ represents the best single block, $\mathcal{I}^K_{2,\succeq}$ the best pair of blocks, through to $\mathcal{I}^K_{|\mathcal{M}_1^K|,\succeq}$ the best $|\mathcal{M}_1^K|$-fold block, these index sets are estimated based on the single-block scores $\hat{\Omega}_{\mathcal{J}_i^K}$ *only*. Thus, it is highly efficient. Nonetheless, we still must estimate $\mathcal{I}^K_{|\mathcal{M}_1^K|+1,\succeq}, \ldots, \mathcal{I}^K_{K,\succeq}$. This can be done by successively appending to $\mathcal{M}_1^K$ the block indices taken from $\{1, \ldots, K\} \setminus \mathcal{M}_1^K$, and computing scores for these appended sets. That is, we start by computing $\hat{\Omega}_{\left(\cup_{i \in \mathcal{M}_1^K} \mathcal{J}_i^K\right) \cup \mathcal{J}_j^K}$ for all $j \in \{1, \ldots, K\} \setminus \mathcal{M}_1^K$. Two scenarios may result:

1. If these are non-zero for $j \in \mathcal{A}_2^K$ for some $\mathcal{A}_2^K \subset \{1, \ldots, K\} \setminus \mathcal{M}_1^K$ only, then these values can be ordered and, in the same way as before, used directly to obtain consistent estimates of $\mathcal{I}^K_{|\mathcal{M}_1^K|+1,\succeq}, \ldots, \mathcal{I}^K_{|\mathcal{M}_1^K|+|\mathcal{A}_2^K|,\succeq}$.

2. If they are all zero, then all pairwise extensions to $\mathcal{M}_1^K$ must be considered, and so on, until one or more non-zero value is found. Assuming that it takes $k > 1$ iterations, denote $\mathcal{A}_2^K$ as the set of all k-sets of $\{1, \ldots, K\} \setminus \mathcal{M}_1^K$ which, when extended to $\mathcal{M}_1^K$, yields $\hat{\Omega} > 0$. This can be used to directly produce a consistent estimate of $\mathcal{I}^K_{|\mathcal{M}_1^K|+k,\succeq}$. Importantly, similar to the case above, if $\mathcal{A}_2^K$ contains multiple k-set entries, then by computing the $\hat{\Omega}$ scores for $\mathcal{M}_1^K$ extended by all *unions* of these, one can also directly produce consistent estimators for each $\mathcal{I}^K_{|\mathcal{M}_1^K|+\ell,\succeq}$, for $\ell \in \{k+1, \ldots, |\mathcal{A}_2^K|\}$, provided that there exists a cardinality-ℓ union for which $\hat{\Omega} > 0$.

One can now iterate the process, with $\mathcal{M}_1^K$ replaced with $\mathcal{M}_2^K = \mathcal{M}_1^K \cup \mathcal{A}_2^K$.

The approach described above can be formalized. Letting $\mathcal{M}_0^K \triangleq \emptyset$, we

define the following sets recursively: for $K \in \{K_1, \ldots, K_M\}$ and $k = 1, 2, \ldots$,

$$\mathcal{M}_k^K \;=\; \mathcal{M}_{k-1}^K \cup \mathcal{A}_k^K \;=\; \mathcal{M}_{k-1}^K \cup \left(\cup_j \mathcal{A}_{kj}^K\right)$$

with $\mathcal{A}_{kj}^K$, $j = 1, 2, \ldots$, the subsets of $\{1, \ldots, K\} \setminus \mathcal{M}_{k-1}^K$ such that

$$\hat{\Omega}_{\cup_{a \in \mathcal{M}_{k-1}^K \cup \mathcal{A}_{k1}^K} \mathcal{J}_a^K} \geq \hat{\Omega}_{\cup_{a \in \mathcal{M}_{k-1}^K \cup \mathcal{A}_{k2}^K} \mathcal{J}_a^K} \geq \cdots > 0$$

and such that $|\mathcal{A}_{k1}| = |\mathcal{A}_{k2}| = \ldots$ is *minimal* in the sense that there exists no subset $\mathcal{A} \subset \{1, \ldots, K\} \setminus \mathcal{M}_{k-1}^K$ of smaller positive cardinality for which $\hat{\Omega}_{\cup_{a \in \mathcal{M}_{k-1}^K \cup \mathcal{A}} \mathcal{J}_a^K} > 0$. The recursion is performed iteratively for $k = 1, \ldots, \tau^K$, where τ^K is the smallest value of k for which $\mathcal{M}_k = \{1, \ldots, K\}$.

Now, for any given $k < \tau^K$, consider the sets

$$\mathcal{M} \in \{\mathcal{M}_{k-1}^K \cup A_{k1}^K,\; \mathcal{M}_{k-1}^K \cup A_{k1}^K \cup A_{k2}^K,\; \ldots\} \,.$$

These $\mathcal{M}$ are the (estimated) best index set candidates—in the sense of partial ordering "$\succeq$"—among all cardinality-$|\mathcal{M}|$ subsets of $\{1, \ldots, K\}$. We may denote for each k, l with $\hat{\Omega}_{\cup_{a \in \mathcal{M}_{k-1}^K \cup (\cup_{j=1}^l \mathcal{A}_{kj}^K)} \mathcal{J}_a^K} > 0$,

$$\hat{\mathcal{I}}^K_{|\mathcal{M}_{k-1}^K \cup (\cup_{j=1}^l \mathcal{A}_{kj}^K)|, \succeq} \triangleq \bigcup_{a \in \mathcal{M}_{k-1}^K \cup (\cup_{j=1}^l \mathcal{A}_{kj}^K)} \mathcal{J}_a^K \subset \{1, \ldots, N\}. \tag{24}$$

It comes immediately that, if (11) holds for $\mathcal{I}^K_{i,\succeq}$ for some i, then there exists k, l and sets $\mathcal{M}_{k-1}^K$ and $\mathcal{A}_{k1}, \ldots, \mathcal{A}_{kl}$ with $|\mathcal{M}_{k-1}^K \cup (\cup_{j=1}^l \mathcal{A}_{kj}^K)| = i$, and thus such that $F_{\hat{\mathcal{I}}^K_{i,\succeq}} - F_{\mathcal{I}^K_{i,\succeq}} \xrightarrow{\text{a.s.}} 0$.

3.5. Proposed Block PCA algorithm

The results given above can be synthesized in the form of an algorithm which seeks to estimate the optimum set $\mathcal{I}$ for $K \in \{K_1, \ldots, K_M\}$. This is described in Algorithm 1, which pertains to our proposed Block PCA method. Therein, the notation $\text{Comb}(\mathcal{A}, k)$ stands for the set of all possible sets of k elements of $\mathcal{A}$, and $\hat{\Omega}_{\mathcal{I}}$ and $\hat{F}_{\mathcal{I}}$ are defined in (17) and (18), respectively. In association with this proposed algorithm, our main theoretical result can then be summarized as follows.

Theorem 1 (Asymptotic Optimality of Algorithm 1). *For $K \in \{K_1, \ldots, K_M\}$ and $1 \leq k \leq K$, let $\mathcal{J}_k^K = \{(k-1)N/K + 1, \ldots, kN/K\}$. Assume that, for all large N, there exists $K_0 \in \{K_1, \ldots, K_M\}$ and $\{k_1, \ldots, k_m\} \subset \{1, \ldots, K_0\}$ such that (11) holds for $\mathcal{I} = \cup_j \mathcal{J}_{k_j}^{K_0}$. Then, denoting $\mathcal{K} = \{\cup_{a \in \mathcal{A}} \mathcal{J}_a^K, K \in \{K_1, \ldots, K_M\}, \mathcal{A} \subset \{1, \ldots, K\}\}$, there exists $\varepsilon > 0$ such that*[4]

$$\hat{F}_{\hat{\mathcal{I}}_{\text{opt}}} - F_{\mathcal{I}_{\text{opt}}} \xrightarrow{\text{a.s.}} 0 \tag{25}$$

[4]Recall that ε enters in the definition of $\hat{\Omega}_{\mathcal{I}}$ and $\hat{F}_{\mathcal{I}}$ for each $\mathcal{I} \in \{1, \ldots, N\}$.

with $\mathcal{I}_{\text{opt}} \triangleq \text{argmax}_{\mathcal{I}\in\mathcal{K}} \{F_{\mathcal{I}}\}$ *and* $\hat{\mathcal{I}}_{\text{opt}}$ *the output of Algorithm 1.*

Moreover, if there exists $\eta > 0$ *such that* $F_{\mathcal{I}_{\text{opt}}} > \max_{\mathcal{I}\in\mathcal{K}\setminus\{\mathcal{I}_{\text{opt}}\}} F_{\mathcal{I}} + \eta$ *for all large* n*,*

$$\hat{\mathcal{I}}_{\text{opt}} = \mathcal{I}_{\text{opt}} \tag{26}$$

for all large n *a.s.*

This specifies that, asymptotically, Algorithm 1 achieves the best performance among any blockwise subset selection. Equation (26) indicates that the subset selection returned by Algorithm 1 is optimal, provided the performance achieved by $\mathcal{I}_{\text{opt}}$ remains sufficiently away from that achieved by all other sets.

3.6. Practical considerations

Here we provide several remarks concerning the practical implementation of the proposed BlockPCA algorithm, Algorithm 1. The first remark concerns computational cost, while the next remark deals with the accuracy of the estimates.

Remark 1 (Early stoppage and sparse setting). *As we will demonstrate, BlockPCA shows substantial performance gains over standard PCA whenever the vector* u *exhibits sparsity, in the sense that its support size is essentially a small fraction of* N*. Under this setting, the scores* $\hat{F}_{\hat{\mathcal{I}}^K_{i,\succeq}}$ *computed by the BlockPCA algorithm will typically tend to significantly decrease with growing* i *(for instance, when* $\hat{F}_{\cup_{k=1}^{i}\mathcal{I}^{(k)}}$ *becomes much smaller than* $\hat{F}_{\max}$ *in the statement of Algorithm 1). Indeed, while not true in the general (non-sparse) setting, for sparse vectors* u*,* $\hat{F}_{\hat{\mathcal{I}}^K_{i,\succeq}}$*, when seen as a function of* i*, generally grows until a maximum and then monotonically decreases (as most of the energy in* u *is rapidly exhausted). Early stoppage is thus possible when this stage is reached, which considerably reduces the computational effort.*

Remark 2 (Estimation of $m_{\mu_{\bar{c}}}$**,** $m'_{\mu_{\bar{c}}}$**).** *In practice, for rather small sets* $\mathcal{I}$*, the estimates* (15) *and* (16) *may not be sufficiently good approximations of* $m_{\mu_{\bar{c}}}$ *and* $m'_{\mu_{\bar{c}}}$*, respectively. Instead it might be preferable to consider more robust estimates, such as*

$$\hat{m}_{\bar{c}}(x) \triangleq \frac{1}{|\mathcal{A}|}\sum_{\mathcal{I}\in\mathcal{A}} \hat{m}_{\mathcal{I}}(x), \qquad \hat{m}'_{\bar{c}}(x) \triangleq \frac{1}{|\mathcal{A}|}\sum_{\mathcal{I}\in\mathcal{A}} \hat{m}'_{\mathcal{I}}(x)$$

with $\mathcal{A}$ *a set of deterministic subsets* $\mathcal{I} \subset \{1,\ldots,N\}$ *all satisfying* $|\mathcal{I}|/n \to \bar{c}$*. For larger* $|\mathcal{I}|$*, in order to estimate* $\Omega_{\mathcal{I}}$ *for a given set* $\mathcal{I}$*,* $\hat{\Omega}_{\mathcal{I}}$ *in* (17) *may nonetheless be a preferable data-adapted estimator than would*

$$\left(\hat{\lambda}_{\mathcal{I}}\hat{m}_{\bar{c}}(\hat{\lambda}_{\mathcal{I}}) + 1\right)^{-1} 1_{\{\hat{\lambda}_{\mathcal{I}} - \hat{\lambda}_{2,\mathcal{I}} > \varepsilon\}}.$$

Algorithm 1 Block PCA algorithm for $K \in \{K_1, \ldots, K_M\}$.

function BLOCKPCA($Y, \{K_1, \ldots, K_M\}$)
  $\hat{F}_{\max} \leftarrow 0$
  $\hat{\mathcal{I}}_{\text{opt}} \leftarrow \emptyset$
  for $K \in \{K_1, \ldots, K_M\}$ **do**
    $\mathcal{M} \leftarrow \emptyset$
    FoundPositiveOmega $\leftarrow 1$
    while FoundPositiveOmega $= 1$ **do**
      $A \leftarrow 0$
      FoundPositiveOmega $\leftarrow 0$
      $\mathcal{A}_{\text{set}} \leftarrow \emptyset$
      while FoundPositiveOmega $= 0$ **and** $|\mathcal{M}| + A < K$ **do**
        $A \leftarrow A + 1$
        $\mathcal{C} \leftarrow \emptyset$
        for $\mathcal{A} \in \text{Comb}(\{1, \ldots, K\} \setminus \mathcal{M}, A)$ **do**
          $\mathcal{I} \leftarrow \left(\cup_{a \in \mathcal{M}} \mathcal{J}_a^K\right) \cup \left(\cup_{a \in \mathcal{A}} \mathcal{J}_a^K\right)$
          if $\hat{\Omega}_{\mathcal{I}} > 0$ **then**
            $\mathcal{C} \leftarrow \mathcal{C} \cup \{\mathcal{I}\}$
            $\mathcal{A}_{\text{set}} \leftarrow \mathcal{A}_{\text{set}} \cup \mathcal{A}$
            FoundPositiveOmega $\leftarrow 1$
          end if
        end for
        if $\mathcal{C} \neq \emptyset$ **then**
          Write $\mathcal{C} = \{\mathcal{I}^{(1)}, \ldots, \mathcal{I}^{(|\mathcal{C}|)}\}$ such that $\hat{\Omega}_{\mathcal{I}^{(1)}} \geq \ldots \geq \hat{\Omega}_{\mathcal{I}^{(|\mathcal{C}|)}}$.
          for $i \in \{1, \ldots, |\mathcal{C}|\}$ **do**
            if $\hat{F}_{\cup_{k=1}^{i} \mathcal{I}^{(k)}} > \hat{F}_{\max}$ **then**
              $\hat{F}_{\max} \leftarrow \hat{F}_{\cup_{k=1}^{i} \mathcal{I}^{(k)}}$
              $\hat{\mathcal{I}}_{\text{opt}} \leftarrow \cup_{k=1}^{i} \mathcal{I}^{(k)}$
            end if
          end for
        end if
      end while
      $\mathcal{M} \leftarrow \mathcal{M} \cup \mathcal{A}_{\text{set}}$
    end while
  end for
  return $\hat{\mathcal{I}}_{\text{opt}}$
end function

These considerations are left to the appreciation of the practitioner. Note that under this new approximation of $m_{\bar{c}}$, sorting the values of $\hat{\Omega}_{\mathcal{I}}$ for various sets $\mathcal{I}$ of same cardinality is strictly equivalent to sorting the values according to the ordering of the values of $\hat{\lambda}_{\mathcal{I}}$, which also comes along with a substantial computational cost reduction.

3.7. Connection to existing sparse PCA methods

As mentioned in Section 1, the BlockPCA algorithm is closely related to various sparse PCA algorithms found in e.g., (Johnstone and Lu, 2009) (SPCA) or (Paul and Johnstone, 2012; Ma, 2013). However, both the target and the model assumptions of BlockPCA versus sparse PCA methods are different. For one, under the present assumption that $N/n \to c \in (0, \infty)$ and fixed ω, the sparse PCA techniques fundamentally assume that there exists $\mathcal{I} \subset \{1, \ldots, N\}$ such that $|\mathcal{I}|/N \to 0$ and $\sum_{i \in \mathcal{I}} |u_i|^2 \to 1$, i.e., the energy in u concentrates in a vanishingly small (relative to N) support. This implies in particular that the leading $|u_i|^2$ are of order $1/|\mathcal{I}| \gg 1/N$. In contrast, our approach is not designed under any such sparsity assumption, but is beneficial whenever u is "rather sparse" in the sense that most energy of u concentrates in one or more isolated places, each of size $O(N)$. Being block-based, our algorithm is not designed to perform well for vectors u with very scattered non-zero entries. For instance, if $u_{\lfloor kN^{1-\alpha} \rfloor} = N^{-\alpha}$ for some $0 < \alpha < 1$ and for all $1 \leq k \leq N^{\alpha}$, and $u_i = 0$ otherwise, BlockPCA will likely select all N entries of u as the optimal subset, whereas SPCA will perform optimally. As such, BlockPCA favors applications for which the index ordering in u matters, and for which sparsity is naturally blockwise. This is often the case when dealing with transformed data (e.g., via Fourier or wavelet) in signal or image processing.

A key feature of both SPCA and BlockPCA is that they select entries or blocks of entries of $\{1, \ldots, N\}$ by successive order of magnitudes. For SPCA, the selection is based on sorting the values of $[YY^*]_{ii}$ from largest to smallest. By central limit theorem arguments forgetting for a while the possible dependence between successive estimates, we have $n^{-1}[YY^*]_{ii} = 1 + \omega|u_i|^2 + O_p(n^{-\frac{1}{2}})$, therefore being only relevant as long as $|u_i|^2$ dominates $n^{-\frac{1}{2}}$. Since N and n scale together, this can only occur for a maximum of $N^{\frac{1}{2}}$ entries of u, thus a vanishing amount. This estimation procedure is therefore relevant for sufficiently sparse vectors u, but is not suitable for vectors which are "not-too-sparse" (e.g., if most of the $|u_i|^2$ are of order N^{-1}). In contrast, our proposed BlockPCA approach leverages the blockwise approach to achieve estimates of the type $\hat{\Omega}_{\mathcal{I}} = \Omega_{\mathcal{I}} + o(1)$ as long as $|\mathcal{I}| = O(N)$. In fact, from (Couillet and Hachem, 2013), this estimate obeys $\hat{\Omega}_{\mathcal{I}} = \Omega_{\mathcal{I}} + O_p(N^{-\frac{1}{2}})$, i.e., $\|u_{\mathcal{I}}\|^2$ is estimated with accuracy $N^{-\frac{1}{2}}$. This allows for the consistent estimation of every block with energy dominating $N^{-\frac{1}{2}}$. We shall see in Section 4 that, for a synthetically-generated example from (Johnstone and Lu, 2009) that can be considered sparse, the SPCA algorithm does not precisely select the relevant indices of u beyond the first few ones, which compromises the performance. The BlockPCA algorithm,

on the other hand, often picks the correct most powerful index blocks and leads to improved performance, even for such sparsity-induced examples.

But the major feature of BlockPCA lies in its ability to consistently estimate the optimal block selection. This capability draws mainly on the fact that the number 2^K of possible block selections remains finite while N, n grow, and on the fact that the elementary subsets $\mathcal{J}_k^K$ are deterministic. The former aspect enables uniform consistency of the 2^K estimates in the large N, n limit, while the latter aspect ensures that random matrix techniques are valid as $X_{\mathcal{I}}$ is independent of the complementary $X_{\mathcal{I}^c}$. The individual entry selection made by SPCA induces too much correlation between the selected sets $\mathcal{I}$ and the random variable X. Beyond the first few selected indices corresponding to maximal $[YY^*]_{ii}$ (which for all large N, n are surely the *deterministic* indices of maximal $|u_i|^2$ values), many selected entries depend on X. This makes it challenging to determine the theoretical limit of $F_{\mathcal{I}}$ when $\mathcal{I}$ contains such indices, let alone deciding on the optimal threshold τ for which only the i's satisfying $[YY^*]_{ii} > \tau$ are selected.

Whereas SPCA is limited by the choice of τ, BlockPCA is in turn limited by the choice of the set $\{K_1, \ldots, K_M\}$. Indeed, while large K values theoretically guarantee refined optimal subsets $\mathcal{I}_{\text{opt}}$, in practice for not-too-large values of N, n, large K's will likely lead to $\Omega_{\mathcal{J}_k^K}$'s of smaller magnitude and to a loss of uniformity in the convergence results. This goes to the extent where, for $K = N$, BlockPCA boils down to SPCA. As a rule of thumb, given the central limit discussion above, it is advised to keep in practice $K < \sqrt{N}$. But, as shall be seen in simulations, it is possible to detect limitations in the choice of a particular K by observing significant mismatches between the performance achieved by $\hat{\mathcal{I}}_{i,\succeq}^K$ and by $\hat{\mathcal{I}}_{i/2,\succeq}^{K/2}$ for several i's. This may suggest errors in the selection of these sets, especially if the performance achieved by the former is worse than that of the latter.

Another selection argument has to do with the application at hand. Depending on the data under consideration, one may have some a-priori knowledge on the typical support size of the localized energy in u. The choice of K should then be adapted to this support. Fourier and wavelet transforms of practical data tend in particular not to provide clean single-valued frequencies but rather spreads of frequencies around given cut-frequencies. Engineering considerations can thus guide the selection of $\{K_1, \ldots, K_M\}$.

4. Simulation results

We provide simulation results which compare BlockPCA against alternative sparse PCA approaches; particularly, the SPCA (Johnstone and Lu, 2009) and TPower (Yuan and Zhang, 2013) methods. Recall that the SPCA algorithm estimates u via $\hat{u}_{\mathcal{I}}$ with $\mathcal{I} = \hat{\mathcal{S}}_k$ for some k, where

$$\hat{\mathcal{S}}_k \triangleq \operatorname{argmax}_{\substack{\mathcal{I} \subset \{1,\ldots,N\} \\ |\mathcal{I}|=k}} \frac{1}{N} \operatorname{tr}\left(\frac{1}{n} Y_{\mathcal{I}} Y_{\mathcal{I}}^*\right).$$

A definite choice of k is in general not available. We shall also denote

$$\mathcal{S}_k \triangleq \operatorname{argmax}_{\substack{\mathcal{I}\subset\{1,\ldots,N\}\\|\mathcal{I}|=k}} \Omega_{\mathcal{I}}$$

the set which the SPCA algorithm seeks and estimates by $\hat{\mathcal{S}}_k$. The TPower method estimates u by $\hat{u}^{\infty}_{(k)}$, the limiting vector obtained by the iterations $\hat{u}^{i+1}_{(k)} = \mathrm{TrunNorm}_k(\frac{1}{n}YY^*\hat{u}^i_{(k)})$ with $\mathrm{TrunNorm}_k(x)$ the vector x with only k largest absolute indices maintained, and normalized to one. Again here k is a parameter chosen by the experimenter. We denote $\hat{\mathcal{T}}_k$ as the support of $\hat{u}^{\infty}_{(k)}$. For these sets, the notation $\mathcal{S}_{\mathrm{opt}}$, $\hat{\mathcal{S}}_{\mathrm{opt}}$, and $\hat{\mathcal{T}}_{\mathrm{opt}}$ are understood as the respective optima across the set indexes.

We shall consider two successive examples. The first one concerns the setting where u has a short but not sparse support, depicted in the first quadrant of Figure 2, and the columns of Y form an autoregressive process of order one. We shall show that in this setting, BlockPCA outperforms sparse PCA alternatives even if the latter are given optimal parametrization of the support size. The second example concerns a genuinely sparse setting borrowed from (Johnstone and Lu, 2009, Figure 1) in which u was manually designed to be sparse in a wavelet domain, ω is rather large, and $T = I_n$. This example serves as a fair comparison of our method against sparse PCA alternatives. It shall be shown on this example that, while BlockPCA largely outperforms SPCA, it performs slightly worse than TPower if the latter were given the optimal support size. Practical considerations on an optimal usage of BlockPCA shall also be discussed. Both instantaneous and 1000-averaged Monte Carlo realizations of the various algorithms will be provided.

We start with the non-sparse scenario. To evaluate the performance of BlockPCA for not-too-large N, n, we take here $N = 512$, $n = 256$. The vector u is defined, up to a rescaling to unit norm, as $u_i = u(i/N)$, where the function $u(x)$ of support $[0, 1]$ is given by

$$u(x) = \mathcal{N}(x; 5/8, 1/80) + 4\mathcal{N}(x; 3/4, 1/40) + 9\mathcal{N}(x; 7/8, 3/80)$$

with $\mathcal{N}(x; a, b)$ the real Gaussian probability density function at x with mean a and standard deviation b. Although u could be clearly made sparse in some well chosen wavelet basis, we assume here that the experimenter is not aware of such a basis so that u is treated as a non-sparse vector. The parameter ω is set to $\omega = 20$ and $[T]_{ij} = 0.7^{|i-j|}$. We finally set $\{K_1, \ldots, K_M\}$ to $\{2, 4, 8, 16, 32\}$. Figure 1 depicts in + signs the averaged values of $\hat{F}_{\hat{\mathcal{I}}_{k,\succeq}}$ obtained by BlockPCA for $K = 32$ (larger black '+' signs) and $K = 8$ (smaller gray '+' signs). These are compared against the deterministic equivalent $\bar{F}_{\mathcal{I}_{k,\succeq}}$ in dots, which the algorithm aims at estimating, and against the actual values of $F_{\hat{\mathcal{I}}_{k,\succeq}}$ in circles, which provide the actual achieved performance (smaller gray dots/circles for $K = 8$). It is observed that the approximation is quite accurate on the first half of the range of $|\mathcal{I}|$ but decays on the second half, with the decay most significant for larger K.

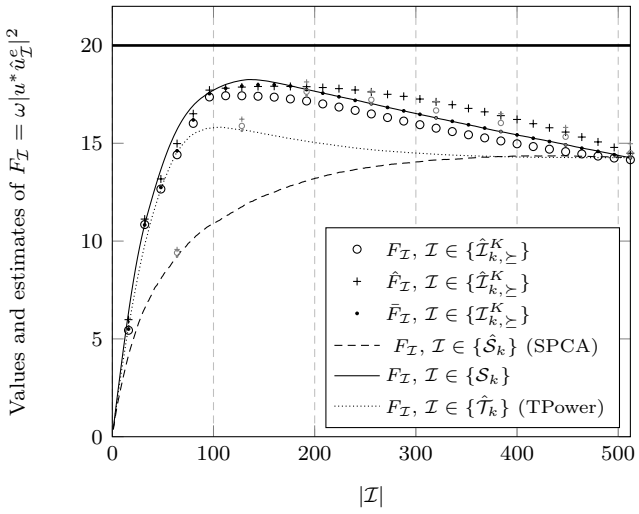

Figure 1: Non-sparse example. Averaged values of $F_{\mathcal{I}}$ ($\circ$) and its random approximation by Algorithm 1 ($+$), for $\mathcal{I} \in \{\hat{\mathcal{I}}_{k,\succeq}^K\}$ found by Algorithm 1, versus deterministic approximation $\bar{F}_{\mathcal{I}}$ for the optimal sets $\mathcal{I} \in \{\mathcal{I}_{k,\succeq}^K\}$ ($\cdot$). Comparison against SPCA performance on SPCA-discovered sets $\hat{\mathcal{S}}_k$ versus optimal sets $\mathcal{S}_k$ and against TPower. Correlation $[T]_{ij} = 0.7^{|i-j|}$ unknown.

This is explained by the fact that, for $N = 512$, $K = 32$ cannot be considered sufficiently small for the joint approximations of $F_{\hat{\mathcal{I}}_{k,\succeq}}$ by $\hat{F}_{\hat{\mathcal{I}}_{k,\succeq}}$ across all k to hold.[5] Nonetheless, it appears that in the region of interest, where $\hat{F}_{\hat{\mathcal{I}}_{k,\succeq}}$ is maximal, BlockPCA on average retrieves the correct optimum set $\mathcal{I}_{\text{opt}}$. The comparison made against SPCA and TPower (in dashed and dotted lines) shows that SPCA is inappropriate for estimating the given u, and reveals optimal performance for a given subset size only on the far right of the plot when $\hat{\mathcal{S}}_k = \{1, \ldots, N\}$, which thus falls back to the standard PCA scenario. The TPower method, on the other hand, retrieves a smaller optimal set $\hat{\mathcal{T}}_k$, but has much degraded performance compared with BlockPCA. We also note that the ideal sought-for $F_{\mathcal{S}_k}$ achieves its maximum very close to BlockPCA's optimum, and thus BlockPCA is close to optimal in this setting. This example therefore clearly shows the ability of our proposed approach to estimate "not-so-sparse" vectors, compared with the sparse PCA approaches, SPCA and TPower, as well as

[5] It is instructive to note that the upper arc of the larger '$+$' signs, which parallels the lower arc of larger '$\circ$' signs, is a strong indicator of the impact of the matrix X in the estimation of $\mathcal{I}_{k,\succeq}^K$. Indeed, the upper arc appears to indicate a successive selection of "energetic" $Y_{\mathcal{I}} Y_{\mathcal{I}}^*$ irrespective of u, while the lower arc reflects that there is increasing mis-selection. This effect cannot be avoided if all the energy in u was exhausted in the first sets $\mathcal{I}_{k,\succeq}^K$. However, with growing N, n, or alternatively for smaller K, both arcs need to tend to one another, which is what is observed here for $K = 8$. This effect is even more visible in Figure 3.

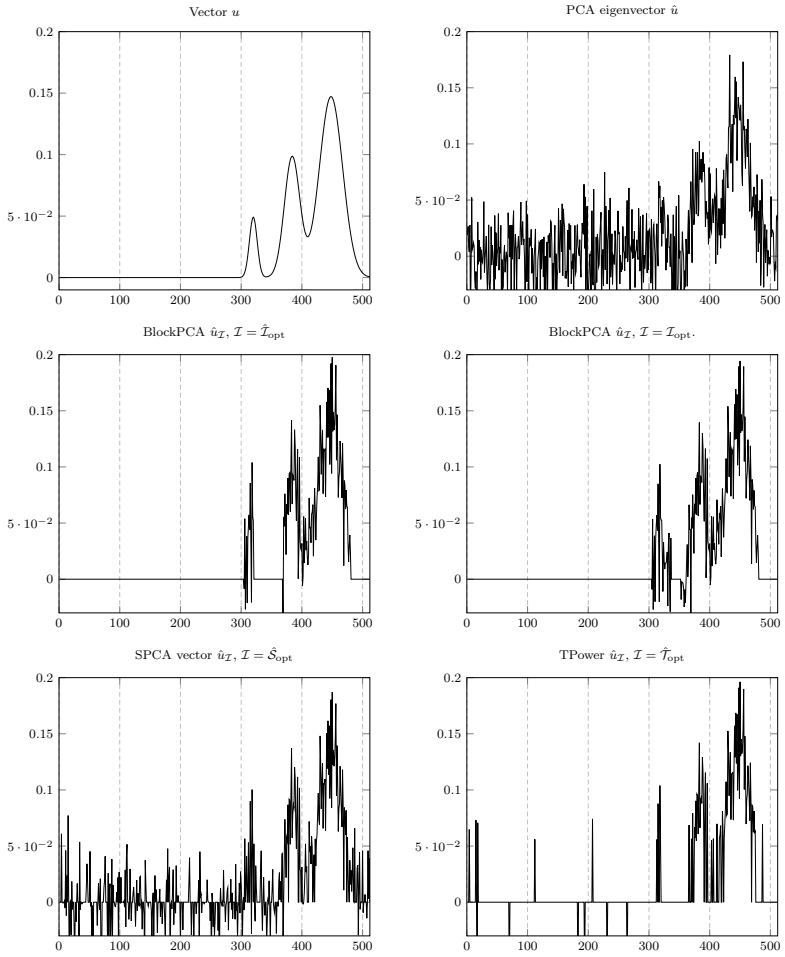

Figure 2: Non-sparse example. Eigenvectors obtained by each algorithm. From left to right and top to bottom: original eigenvector; standard PCA; leading eigenvector found by Algorithm 1 assuming T unknown; optimum theoretical BlockPCA eigenvector; SPCA with oracle decision on optimal set $\hat{\mathcal{S}}_k$; TPower with oracle knowledge of best $\hat{\mathcal{T}}_k$.

classical PCA.

A single (but typical) realization of the eigenvector estimates retrieved by each algorithm is provided in Figure 2. The substantial improvement of Block PCA over standard PCA is clearly evident, with the standard approach giving noisy results and having difficulties in picking up the smaller peaks in u in particular (which BlockPCA retrieves accurately). We caution, however, that this is not a systematic scenario, as BlockPCA is not perfect, and sometimes misses the weakest peak or creates some extra peaks. Note also that $\hat{\mathcal{I}}_{\text{opt}}$ differs from $\mathcal{I}_{\text{opt}}$ in this example. The SPCA and TPower figures are obtained at their respective optimal point (which we recall—in contrast BlockPCA—is *not* made

available by the algorithm), hence providing here a best case scenario. It is seen that SPCA, while not being exactly the standard PCA output, shows an important number of spurious values outside the support of u. As for TPower, it displays a smaller support but misses the connection between the two largest peaks and presents extra small peaks outside the genuine support of u.

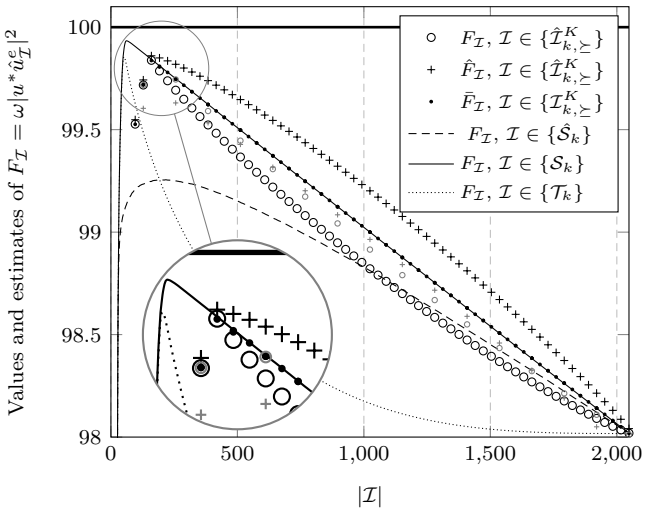

Figure 3: Setting of (Johnstone and Lu, 2009, Figure 1). Averaged values of $F_{\mathcal{I}}$ ($\circ$) and its random approximation by Algorithm 1 ($+$), for $\mathcal{I} \in \{\hat{\mathcal{I}}_{k,\succeq}^K\}$ found by Algorithm 1, versus deterministic approximation $\bar{F}_{\mathcal{I}}$ for the optimal sets $\mathcal{I} \in \{\mathcal{I}_{k,\succeq}^K\}$ ($\cdot$). Comparison against SPCA performance on SPCA-discovered sets $\hat{\mathcal{S}}_k$ versus optimal sets $\mathcal{S}_k$. Correlation $T = I_n$ known.

One may argue, of course, that both the SPCA and TPower methods were mainly designed for the estimation of sparse vectors, as opposed to not-so-sparse vectors, which are the basis for the BlockPCA algorithm. Thus, for a fair comparison, it is instructive to compare the performance of the different algorithms under a suitably sparse setting, particularly to see if the BlockPCA technique degrades substantially and is overwhelmed in performance by SPCA and TPower. To investigate this, we now consider the setting of (Johnstone and Lu, 2009, Figure 1) where, when seen in a wavelet domain, u is sparse. Following (Johnstone and Lu, 2009) exactly, we take $N = 2048$, $n = 1024$, $\omega = 100$, $T = I_n$ (which we assume known to BlockPCA). The original-domain vector u is displayed in the upper-left quadrant of Figure 4. Its value is given by $u_i = u(i/N)$, where the function $u(x)$ is defined on $[0, 1]$ as

$$u(x) = C\left(0.7b(1500, 3000, x) + 0.5b(1200, 900, x) + 0.5b(600, 160, x)\right)$$

with C a normalization constant and $b(a, b, x)$ the Beta-density given by $b(a, b, x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^a (1-x)^{b-1}$. This vector u is sparse when wavelet transformed using

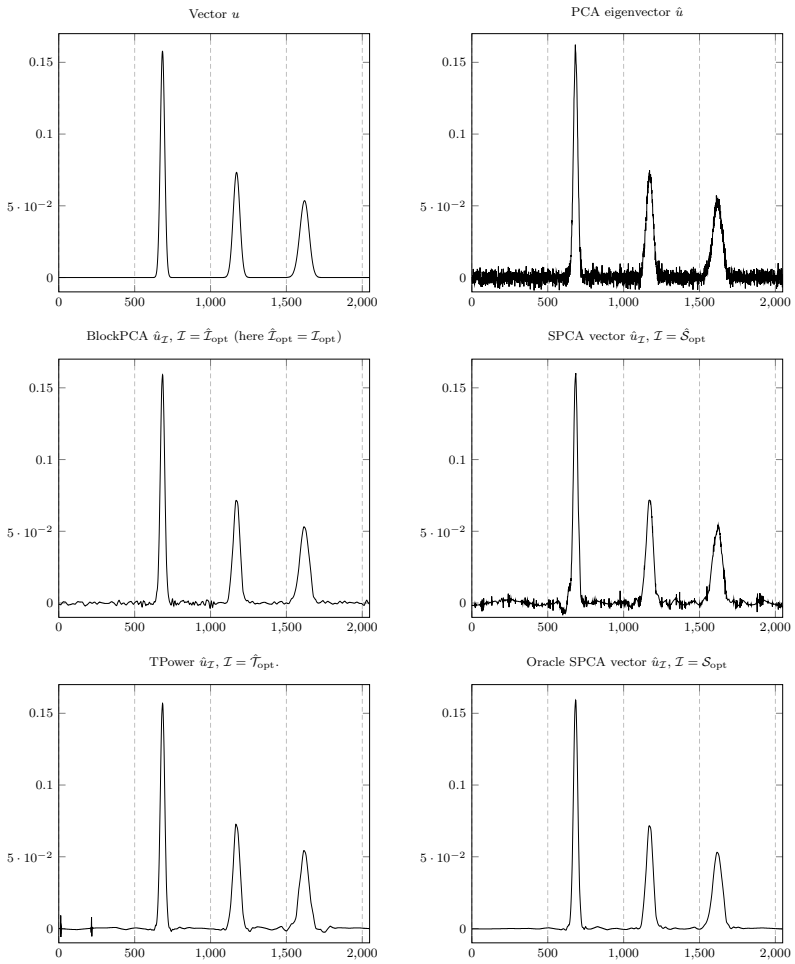

Figure 4: Setting of (Johnstone and Lu, 2009, Figure 1). Resulting inverse wavelet-transformed eigenvectors. From left to right and top to bottom: original eigenvector; standard PCA; leading eigenvector found by Algorithm 1 assuming $T = I_n$ is known (corresponds on this realization to the leading eigenvector for the theoretically optimal subset selection $\mathcal{I}_{\rm opt}$); leading eigenvector found by Algorithm 1 assuming $T = I_n$ is unknown; SPCA with oracle decision on optimal thresholding of $[YY^*]_{ii}$; SPCA with oracle knowledge of both ordering of $|u_i|^2$ and optimal thresholding of $[YY^*]_{ii}$.

a Symmlet basis. This transformation, not documented in (Johnstone and Lu, 2009, Figure 1), was performed here via the WaveLab Matlab© library using the functions[6]

[6]The authors in (Johnstone and Lu, 2009) may have used different parameter values than 1 and 4 in their own definition of `WT` and `IWT`.

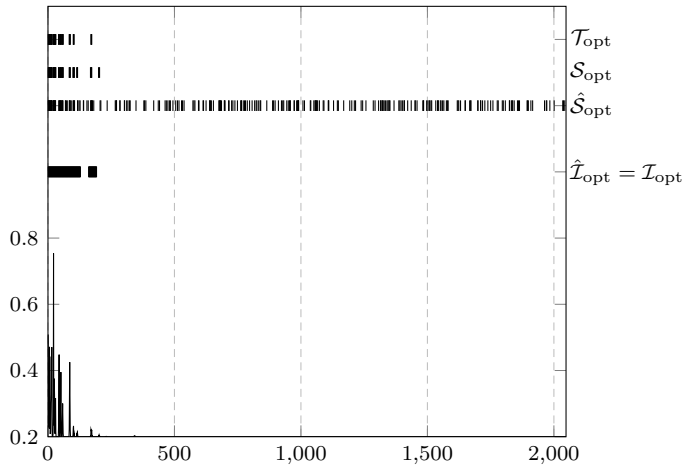

Figure 5: Setting of (Johnstone and Lu, 2009, Figure 1). Bottom: absolute eigenvector $|u|$ (seen in wavelet domain). Top: optimal selection sets, theoretical and empirical.

```
WT  = @(x) FWT_PO(x,1,MakeONFilter('Symmlet',4))
IWT = @(x) IWT_PO(x,1,MakeONFilter('Symmlet',4))
```

The absolute value of the resulting wavelet transformed vector is displayed in the bottom part of Figure 5. It is observed that the eigenvector support spans about 5% to 10% of the vector size and is rather connected, therefore making it quite sparse but also accessible to BlockPCA. Despite BlockPCA not being at a particular advantage in this scenario, we shall observe that it still outperforms SPCA and, remarkably, achieves virtually the same performance as TPower when the latter is assumed to know the optimal support size selection.

Since $T = I_n$ is assumed known, we consider here for BlockPCA the estimators in (10), with $K \in \{2, 4, 8, 16, 32, 64\}$. For the sake of clarity, only the results for $K \in \{16, 64\}$ will be depicted. Similar to Figure 1, Figure 3 displays in continuous lines the averaged performance of SPCA and TPower, along with the oracle SPCA with knowledge of $\mathcal{S}_k$. Similar to previously, aside from the very first strongest values of $|u_i|$ (here about 30 of them), quickly the SPCA algorithm shows difficulties in retrieving the correct indices of energetic $|u_i|$. TPower demonstrates considerable performance improvement, though it also shows a particularly sensitive turning point beyond which the performance drops significantly and $F_{\hat{\mathcal{T}}_k}$ becomes a weak approximation of the optimal $F_{\mathcal{S}_k}$. By comparison, BlockPCA is a close approximation to $F_{\mathcal{S}_k}$ on most of its range. Since N is four times larger than in the example of Figure 1, observe that the approximation $\hat{F}_{\hat{\mathcal{I}}^K_{k,\geq}}$ for $K = 16$ is quite accurate along the whole range of values of $F_{\mathcal{I}^K_{k,\geq}}$. For $K = 64$, the approximations tend to be worse but are still extremely accurate on the range of most interest (around the peak). In particular, observe that $\hat{F}_{\hat{\mathcal{I}}_{\text{opt}}}$ is a particularly close approximation to $F_{\mathcal{I}_{\text{opt}}}$ which strongly suggests that $\hat{\mathcal{I}}_{\text{opt}} = \mathcal{I}_{\text{opt}}$ for most realizations. Finally, it is worth not-

ing that all three approaches in this case can perform better than standard PCA, the performance of which is given by the value corresponding to the right-most point (i.e., $\mathcal{I} = \{1, \ldots, 2048\}$).

We next provide in Figure 4 a representative single realization of the eigenvectors retrieved by each algorithm. These eigenvectors are all obtained after performing the inverse wavelet transform (via the function `IWT` above) of the eigenvectors obtained as the output of the algorithms. Since $\hat{\mathcal{I}}_{\text{opt}} = \mathcal{I}_{\text{opt}}$ for most realizations, we do not depict both cases here. In terms of performance, it is evident that BlockPCA largely outperforms the SPCA eigenvector obtained for the optimal selection $\hat{\mathcal{S}}_{\text{opt}}$, while BlockPCA and TPower are comparable. As a further comparison, we provide in the lower right quadrant the sought-for outcome of SPCA, corresponding to optimal subset selection $\mathcal{S}_{\text{opt}}$ (which can thus be considered as an overall optimum).

In Figure 5, taking the same realization as for Figure 4, in addition to plotting the magnitude of the wavelet coefficients of vector u, at the top of the figure we depict the sets $\mathcal{I}_{\text{opt}}$, $\hat{\mathcal{I}}_{\text{opt}}$, along with $\mathcal{S}_{\text{opt}}$, $\hat{\mathcal{T}}_{\text{opt}}$, and $\hat{\mathcal{S}}_{\text{opt}}$. As already recalled $\hat{\mathcal{I}}_{\text{opt}} = \mathcal{I}_{\text{opt}}$ on this realization, therefore leading to an optimal BlockPCA set selection. This set turns out to be $\mathcal{I}^{64}_{5,\succeq} = \mathcal{J}^{64}_1 \cup \mathcal{J}^{64}_2 \cup \mathcal{J}^{64}_3 \cup \mathcal{J}^{64}_4 \cup \mathcal{J}^{64}_6$. Note that the set $\mathcal{S}_{\text{opt}}$ resembles $\mathcal{I}_{\text{opt}}$ but for a more discrete (higher resolution) support, which explains the superiority of the oracle SPCA result. The optimum SPCA set selection $\hat{\mathcal{S}}_{\text{opt}}$ has an extremely extended support despite its sharing most of the support of $\mathcal{S}_{\text{opt}}$ (the reader must be aware that, despite the visual impression, the support of $\mathcal{S}_{\text{opt}}$ does in general not exceed 10% of the support of $\{1, \ldots, N\}$ as hinted at in Figure 3). This can be explained by the fact that, while the first indices selected by SPCA correspond to genuinely optimal indices in u, those selected subsequently are competing against the fluctuations of X which induce large values in the diagonal elements of YY^* independently of u: since N is much larger than the support of u, these errors are likely numerous and it takes quite a few erroneous index selections to exhaust most of the support of u. Finally, as opposed to $\hat{\mathcal{S}}_{\text{opt}}$ which tends to be larger than $\mathcal{S}_{\text{opt}}$, TPower seems to produce a subset of $\mathcal{S}_{\text{opt}}$. This is consistent with Figure 3, suggesting that $\hat{\mathcal{T}}_k = \mathcal{S}_k$ for the first few values of k, until a point where every additional index strongly taints the performance of TPower.

5. Discussion and future perspectives

We have presented a novel blockwise sparse PCA approach which leverages recent results regarding the extreme eigenvalues and eigenvectors of spike random matrix models. Our results present a key distinction from most existing work in the sparse PCA literature, in that they do not impose a strong sparse structure for which the number of contributing variables is a rapidly vanishing fraction of the total number of variables, and we do not seek algorithms which provide consistent (and rate-optimal) estimates. Instead, we consider "not so sparse" models in which the fraction of contributing variables is non-vanishing, and for which consistent sparse PCA algorithms are generally not

attainable under the regime where the number of samples and number of variables grow at the same rate. Our proposed approach presents an efficient and asymptotically-optimal blockwise variable selection which yields the right balance between minimizing errors due to finite sampling (achieved with more aggressive variable selection), and maximizing the retained eigenvector energy (achieved with less aggressive variable selection). Unlike most existing sparse PCA algorithms, the proposed method is completely data driven with no free parameters requiring ad-hoc tuning, while also allowing for possibly unknown linear correlations between the samples. Our numerical studies demonstrate the performance merits of the approach with respect to competing sparse PCA algorithms.

We point out that the purpose of this paper was to pose, for the first time, the idea of using the limiting eigenvalue and eigenvector results from random matrix theory to provide consistent estimators of the errors achieved with different variable selections, and to apply such estimators to propose an objective data-driven variable selection algorithm. These results were demonstrated under various assumptions, some more restrictive than others. First, our proposed method assumed that the underlying model eigenvector u has its energy mainly contained in a possibly sparse group of blocks of consecutive entries. We believe that this assumption is practically reasonable, particularly when working in a transformed domain such as in a Fourier (or wavelet) basis, where the signal energy is often clustered around certain frequency bands (typically low frequencies). A more restrictive assumption of our model is that we consider a single spike only; effectively representing the case of a rank-1 signal in noise. The relevant random matrix properties concerning the asymptotic eigenvalue limits and eigenvector projections are indeed known for spike models beyond the rank-1 case (see e.g., (Baik and Silverstein (2006); Couillet and Hachem (2013))), however a refined variable selection strategy will be necessary for such cases. Such a refinement appears non-trivial. For example, even for the simplest 2-spike departure, a key challenge arises by the fact that for any proposed blockwise variable selection $\mathcal{I}$, the extracted components from the top two spike eigenvectors will generally no longer be orthogonal. Moreover, even if they were, the subset selection may lead to a re-ordering of the corresponding sample eigenvector-eigenvalue pairs which appears difficult to resolve. These issues may be partially alleviated if the supports of the spike eigenvectors were assumed to be non-overlapping. While seemingly a strong assumption, this type of structured correlation model is in fact representative of various applications for which distinct groups of mutually correlated variables exist, but with the groups being uncorrelated with each other. For example, this has been shown to be relevant in biological applications and used to define "biological sectors" (Dahirel et al. (2011); Quadeer et al. (2014)) identifying co-evolving protein sites in HIV and Hepatitis C, and is a reasonable approximation for classifying "financial sectors" of economic activity in financial time series (Plerou et al. (2002)). While beyond the scope of the current paper, further development of our ideas along these lines is an interesting avenue for further work. Alternative correlation scenarios may also be considered; for example, by replacing the iden-

tity matrix in (1) with a more general correlation construction. In that case, one could potentially leverage recent results which have been derived for so-called "generalized spike" models (see e.g., (Bai and Yao (2012); Ding (2015))). Finally, as for many of the sparse PCA algorithms in the literature, we have assumed a Gaussian data model. This was required in order to establish the reduced spike model equivalence in (4), though other data distributions may be possibly considered; for example, heavy-tailed elliptical models, which may utilize very recent results that have been established for the corresponding random matrices (Couillet (2015)). This is yet another interesting extension for future work.

6. A technical lemma

Lemma 1 (A monotonicity result). *Let μ be a probability measure with compact support* $\mathrm{supp}(\mu) \subset \mathbb{R}^+$ *and Stieltjes transform m_μ. Then the functions*

$$\begin{aligned}(\sup(\mathrm{supp}(\mu)), \infty) &\to \mathbb{R}^+ \\ x &\mapsto -\frac{1}{1 + x m_\mu(x)}\end{aligned}$$

and

$$\begin{aligned}(\sup(\mathrm{supp}(\mu)), \infty) &\to \mathbb{R}^+ \\ x &\mapsto -\frac{m_\mu(x)}{m_\mu(x) + x m'_\mu(x)}\end{aligned}$$

are both positive and increasing.

Proof. For the first function,

$$-\frac{1}{1 + x m_\mu(x)} = \left(\int \frac{t}{x - t} \mu(dt) \right)^{-1}$$

which is positive on $(\mathrm{supp}(\mu), \infty)$ and increasing. For the second function, positivity is ensured by observing that the numerator is negative while the denominator satisfies $m_\mu(x) + x m'_\mu(x) = \int t(x-t)^2 \mu(dt) > 0$. To verify monotonicity, differentiating along x gives a derivative of the same sign as

$$\begin{aligned}&- m'_\mu(x) \left(m_\mu(x) + x m'_\mu(x) \right) + m_\mu(x) \left(2m'_\mu(x) + m''_\mu(x) \right) \\ &= - \int \frac{1}{(x-t)^2} \mu(dt) \int \frac{t}{(x-t)^2} \mu(dt) + 2 \int \frac{1}{x-t} \mu(dt) \int \frac{t}{(x-t)^3} \mu(dt).\end{aligned}$$

Writing $-\int t(x-t)^p \mu(dt) = \int (x-t)^{p-1} \mu(dt) - x \int (x-t)^p$ for both terms involving t in the numerators, this becomes

$$- \int \frac{\mu(dt)}{(x-t)^2} \int \frac{\mu(dt)}{x-t} - x \left(\int \frac{\mu(dt)}{(x-t)^2} \right)^2 + 2x \int \frac{\mu(dt)}{x-t} \int \frac{\mu(dt)}{(x-t)^3}.$$

By the Cauchy–Schwarz inequality, $\left(\int \frac{\mu(dt)}{(x-t)^2}\right)^2 \leq \int \frac{\mu(dt)}{x-t} \int \frac{\mu(dt)}{(x-t)^3}$, so that

$$\begin{aligned}
&- m'_\mu(x)\left(m_\mu(x) + x m'_\mu(x)\right) + m_\mu(x)\left(2m'_\mu(x) + m''_\mu(x)\right) \\
&\geq \int \frac{\mu(dt)}{x-t}\left(x\int \frac{\mu(dt)}{(x-t)^3} - \int \frac{\mu(dt)}{(x-t)^2}\right) \\
&= \int \frac{\mu(dt)}{x-t} \int \frac{t}{(x-t)^3}\mu(dt) \\
&> 0\,.
\end{aligned}$$

References

Bai, Z., Yao, Y., 2012. On sample eigenvalues in a generalized spike population model. Journal of Multivariate Analysis 106, 167–177.

Baik, J., Silverstein, J. W., 2006. Eigenvalues of large sample covariance matrices of spiked population models. Journal of Multivariate Analysis 97 (6), 1382–1408.

Benaych-Georges, F., Nadakuditi, R. R., 2012. The singular values and vectors of low rank perturbations of large rectangular random matrices. Journal of Multivariate Analysis 111, 120–135.

Benaych-Georges, F., Nadakuditi, R. R., 2011. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. Advances in Mathematics 227 (1), 494–521.

Bianchi, P., Najim, J., Maida, M., Debbah, M., 2011. Performance of some eigen-based hypothesis tests for collaborative sensing. IEEE Transactions on Information Theory 57 (4), 2400–2419.

Bickel, P. J., Levina, E., 2008. Regularized estimation of large covariance matrices. The Annals of Statistics 36 (1), 199–227.

Birnbaum, A., Johnstone, I. M., Nadler, B., Paul, D., 2013. Minimax bounds for sparse PCA with noisy high-dimensional data. The Annals of Statistics 41 (3), 1055–1084.

Cai, T., Ma, Z., Wu, Y., 2013. Sparse PCA: Optimal rates and adaptive estimation. The Annals of Statistics 41 (6), 3074–3110.

Candès, E., Li, X., Ma, Y., Wright, J., 2011. Robust principal component analysis? Journal of the ACM (JACM) 58 (3), 11.

Couillet, R., Debbah, M., 2011. Random Matrix Methods for Wireless Communications, 1st Edition. Cambridge University Press, New York, NY, USA.

Couillet, R., Hachem, W., 2013. Fluctuations of spiked random matrix models and failure diagnosis in sensor networks. IEEE Transactions on Information Theory 59 (1), 509–525.

Couillet, R., Silverstein, J. W., Bai, Z. D., Debbah, M., 2011. Eigen-inference for energy estimation of multiple sources. IEEE Transactions on Information Theory 57 (4), 2420–2439.

Couillet, R., 2015. Robust spiked random matrices and a robust G-MUSIC estimator. Journal of Multivariate Analysis 140, 139–161.

Dahirel, V., Shekhar, K., Pereyra, F., Miura, T., Artyomov, M., Talsania, S., Allen, T. M., Altfeld, M., Carrington, M., Irvine, D. J., et al., 2011. Coordinate linkage of HIV evolution reveals regions of immunological vulnerability. Proceedings of the National Academy of Sciences 108 (28), 11530–11535.

d'Aspremont, A., Bach, F., El Ghaoui, L., 2008. Optimal solutions for sparse principal component analysis. The Journal of Machine Learning Research 9, 1269–1294.

d'Aspremont, A., El Ghaoui, L., Jordan, M. I., Lanckriet, G., 2007. A direct formulation for sparse PCA using semidefinite programming. SIAM Review 49 (3), 434–448.

Ding, X., 2015. Convergence of sample eigenvectors of spiked population model. Communications in Statistics–Theory and Methods 44, 3825–3840.

Johnstone, I. M., Lu, A. Y., 2009. On consistency and sparsity for principal components analysis in high dimensions. Journal of the American Statistical Association 104 (486).

Jolliffe, I. T., Trendafilov, N. T., Uddin, M., 2003. A modified principal component technique based on the LASSO. Journal of Computational and Graphical Statistics 12 (3), 531–547.

Journée, M., Nesterov, Y., Richtárik, P., Sepulchre, R., 2010. Generalized power method for sparse principal component analysis. The Journal of Machine Learning Research 11, 517–553.

Laloux, L., Cizeau, P., Potters, M., Bouchaud, J. P., Jul. 2000. Random matrix theory and financial correlations. International Journal of Theoretical and Applied Finance 3 (3), 391–397.

Ledoit, O., Wolf, M., 2003. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. Journal of Empirical Finance 10 (5), 603–621.

Ma, Z., 2013. Sparse principal component analysis and iterative thresholding. The Annals of Statistics 41 (2), 772–801.

Mairal, J., Bach, F., Ponce, J., Sapiro, G., 2010. Online learning for matrix factorization and sparse coding. The Journal of Machine Learning Research 11, 19–60.

Mestre, X., Lagunas, M., Feb. 2008. Modified subspace algorithms for DoA estimation with large arrays. IEEE Transactions on Signal Processing 56 (2), 598–614.

Moore, B. E., Nadakuditi, R. R., Fessler, J. A., 2014. Improved robust PCA using low-rank denoising with optimal singular value shrinkage. In: IEEE Workshop on Statistical Signal Processing (SSP'14). IEEE, pp. 13–16.

Nadakuditi, R. R., 2014. OptShrink: An algorithm for improved low-rank signal matrix denoising by optimal, data-driven singular value shrinkage. IEEE Transactions on Information Theory 60 (5), 3002–3018.

Paul, D., 2007. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. Statistica Sinica, 17, 1617–1642.

Paul, D., Johnstone, I. M., 2012. Augmented sparse principal component analysis for high dimensional data. arXiv preprint arXiv:1202.1242.

Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L., Guhr, T., Stanley, H., 2002. Random matrix approach to cross correlations in financial data. Phys. Rev. E. 65:1–18.

Quadeer, A. A., Louie, R. H. Y., Shekhar, K., Chakraborty, A. K., Hsing, I.-M., McKay, M. R., 2014. Statistical linkage of mutations in the non-structural proteins of hepatitis C virus exposes targets for immunogen design. Journal of Virology 88 (13), 7628–7644.

Rubio, F., Mestre, X., 2009. Consistent reduced-rank LMMSE estimation with a limited number of samples per observation dimension. IEEE Transactions on Signal Processing 57 (8), 2889–2902.

Rubio, F., Mestre, X., Palomar, D. P., 2012. Performance analysis and optimal selection of large minimum variance portfolios under estimation risk. IEEE Journal of Selected Topics in Signal Processing 6 (4), 337–350.

Shen, D., Shen, H., Marron, J. S., 2013. Consistency of sparse PCA in high dimension, low sample size contexts. Journal of Multivariate Analysis 115, 317–333.

Shen, H., Huang, J. Z., 2008. Sparse principal component analysis via regularized low rank matrix approximation. Journal of Multivariate Analysis 99 (6), 1015–1034.

Silverstein, J. W., Bai, Z. D., 1995. On the empirical distribution of eigenvalues of a class of large dimensional random matrices. Journal of Multivariate Analysis 54 (2), 175–192.

Ulfarsson, M. O., Solo, V., 2008. Sparse variable PCA using geodesic steepest descent. IEEE Transactions on Signal Processing 56 (12), 5823–5832.

Vinogradova, J., Couillet, R., Hachem, W., 2013. Statistical inference in large antenna arrays under unknown noise pattern. IEEE Transactions on Signal Processing 61 (22), 5633–5645.

Vinogradova, J., Couillet, R., Hachem, W., 2015. Estimation of Toeplitz covariance matrices in large dimensional regime with application to source detection. IEEE Transactions on Signal Processing 63 (18), 4903–4913.

Wang, Z., Lu, H., Liu, H., 2014. Tighten after relax: Minimax-optimal Sparse PCA in polynomial time. Adv. Neural Inf. Process. Syst., 3383–3391.

Witten, D., Tibshirani, R., Hastie, T., 2009. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics 10 (3), 515–534.

Yang, L., Couillet, R., McKay, M., 2014. Minimum variance portfolio optimization with robust shrinkage covariance estimation. In: Proc. IEEE Asilomar Conference on Signals, Systems, and Computers. Pacific Grove, CA, USA.

Yuan, X.-T., Zhang, T., 2013. Truncated power method for sparse eigenvalue problems. The Journal of Machine Learning Research 14 (1), 899–925.

Zeng, Y., Liang, Y. C., 2009. Eigenvalue based spectrum sensing algorithms for cognitive radio. IEEE Transactions on Communications 57 (6), 1784–1793.

Zou, H., Hastie, T., Tibshirani, R., 2006. Sparse principal component analysis. Journal of Computational and Graphical Statistics 15 (2), 265–286.