Robust high dimensional factor models with applications to statistical machine learning

Jianqing Fan[†∗]**, Kaizheng Wang**[‡∗]**, Yiqiao Zhong**[⋆∗] **and Ziwei Zhu**[♯∗]

[†]Princeton University, [‡]Columbia University, [⋆]Stanford University, [♯]University of Michigan

Abstract. Factor models are a class of powerful statistical models that have been widely used to deal with dependent measurements that arise frequently from various applications from genomics and neuroscience to economics and finance. As data are collected at an ever-growing scale, statistical machine learning faces some new challenges: high dimensionality, strong dependence among observed variables, heavy-tailed variables and heterogeneity. High-dimensional robust factor analysis serves as a powerful toolkit to conquer these challenges.

This paper gives a selective overview on recent advance on high-dimensional factor models and their applications to statistics including Factor-Adjusted Robust Model selection (FarmSelect) and Factor-Adjusted Robust Multiple testing (FarmTest). We show that classical methods, especially principal component analysis (PCA), can be tailored to many new problems and provide powerful tools for statistical estimation and inference. We highlight PCA and its connections to matrix perturbation theory, robust statistics, random projection, false discovery rate, etc., and illustrate through several applications how insights from these fields yield solutions to modern challenges. We also present far-reaching connections between factor models and popular statistical learning problems, including network analysis and low-rank matrix recovery.

Key words and phrases: Factor model, PCA, covariance estimation, perturbation bounds, robustness, random sketch, FarmSelect, FarmTest.

Department of Operations Research and Financial Engineering, Princeton University, Princeton, 08540, NJ, USA (e-mail: jqfan@princeton.edu)
Department of Industrial Engineering and Operations Research, Columbia University, New York, NY (e-mail: kaizheng@princeton.edu)
Packard 239, Stanford University, Stanford, 94305, CA, USA (e-mail: yiqiaoz@stanford.edu)
Department of Statistics, University of Michigan, Ann Arbor, 48109, MI, USA (e-mail: ziweiz@umich.edu)

1. INTRODUCTION

In modern data analytics, dependence across high-dimensional outcomes or measurements is ubiquitous. For example, stocks within the same industry exhibit significantly correlated returns, housing prices of a country depend on various economic factors, gene expressions can be stimulated by cytokines. Ignoring such dependence structure can produce significant systematic bias and yields inefficient statistical results and misleading insights. The problems are more severe for high-dimensional big data, where dependence, non-Gaussianity and heterogeneity of measurements are common.

Factor models aim to capture such dependence by assuming several variates or "factors", usually much fewer than the outcomes, that drive the dependence of the entire outcomes (Lawley and Maxwell, 1962; Stock and Watson, 2002). Stemming from the early works on measuring human abilities (Spearman, 1927), factor models have become one of the most popular and powerful tools in multivariate analysis and have made profound impact in the past century on psychology (Bartlett, 1938; McCrae and John, 1992), economics and finance (Chamberlain and Rothschild, 1982; Fama and French, 1993; Stock and Watson, 2002; Bai and Ng, 2002), biology (Hirzel et al., 2002; Hochreiter et al., 2006; Leek and Storey, 2008), etc. Suppose $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are n i.i.d. p-dimensional random vectors, which may represent financial returns, housing prices, gene expressions, etc. The generic factor model assumes that

$$(1.1) \qquad \mathbf{x}_i = \boldsymbol{\mu} + \mathbf{B}\mathbf{f}_i + \mathbf{u}_i, \quad \text{or in matrix form,} \quad \mathbf{X} = \boldsymbol{\mu}\mathbf{1}_n^\top + \mathbf{B}\mathbf{F}^\top + \mathbf{U},$$

where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathbb{R}^{p \times n}$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^\top$ is the mean vector, $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_p)^\top \in \mathbb{R}^{p \times K}$ is the matrix of factor loadings, $\mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_n)^\top \in \mathbb{R}^{n \times K}$ stores K-dimensional vectors of common factors with $\mathbb{E}\mathbf{f}_i = \mathbf{0}$, and $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_n) \in \mathbb{R}^{p \times n}$ represents the error terms (a.k.a. idiosyncratic components), which has mean zero and is uncorrelated with or independent of $\mathbf{F}$. We emphasize that, for most of our discussions in the paper (except Section 3.1), only $\{\mathbf{x}_i\}_{i=1}^n$ are observable, and the goal is to infer $\mathbf{B}$ and $\{\mathbf{f}_i\}_{i=1}^n$ through $\{\mathbf{x}_i\}_{i=1}^n$. Here we use the name "factor model" to refer to a general concept where the idiosyncratic components $\mathbf{u}_i$ are allowed to be weakly correlated. This is also known as the "approximate factor model" in the literature, in contrast to the "strict factor model" where the idiosyncratic components are assumed to be uncorrelated.

Note that the model (1.1) has identifiability issues: given any invertible matrix $\mathbf{R} \in \mathbb{R}^{K \times K}$, simultaneously replacing $\mathbf{B}$ with $\mathbf{BR}$ and $\mathbf{f}_i$ with $\mathbf{R}^{-1}\mathbf{f}_i$ does not change the observation $\mathbf{x}_i$. To resolve this ambiguity issue, the following identifiability assumption is usually imposed:

ASSUMPTION 1.1 (Identifiability). $\mathbf{B}^\top\mathbf{B}$ *is diagonal and* $\operatorname{cov}(\mathbf{f}_i) = \mathbf{I}_p$.

Other identifiability assumptions as well as detailed discussions can be found in Bai and Li (2012) and Fan et al. (2013).

Factor analysis is closely related to principal component analysis (PCA), which breaks down the covariance matrix into a set of orthogonal components and identifies the subspace that explains the most variation of the data (Pearson, 1901; Hotelling, 1933). In this selective review, we will mainly leverage PCA, or

more generally, spectral methods, to estimate the factors $\{\mathbf{f}_i\}_{i=1}^n$ and the loading matrix $\mathbf{B}$ in (1.1). Other popular estimators, mostly based on the maximum likelihood principle, can be found in Lawley and Maxwell (1962); Anderson and Amemiya (1988); Bai and Li (2012), etc. The covariance matrix of $\mathbf{x}_i$ consists of two components: $\mathrm{cov}(\mathbf{Bf}_i)$ and $\mathrm{cov}(\mathbf{u}_i)$. Intuitively, when the contribution of the covariance from the error term $\mathbf{u}_i$ is negligible compared with those from the factor term $\mathbf{Bf}_i$, the top-K eigenspace (namely, the space spanned by top K eigenvectors) of the sample covariance of $\{\mathbf{x}_i\}_{i=1}^n$ should be well aligned with the column space of $\mathbf{B}$. This can be seen from the assumption that $\mathrm{cov}(\mathbf{x}_i) = \mathbf{BB}^\top + \mathrm{cov}(\mathbf{u}_i) \approx \mathbf{BB}^\top$, which occurs frequently in high-dimensional statistics (Fan et al., 2013).

Here is our main message: applying PCA to well-crafted covariance matrices (including vanilla sample covariance matrices and their robust version) consistently estimates the factors and loadings, as long as the signal-to-noise ratio is large enough. The core theoretical challenge is to characterize how idiosyncratic covariance $\mathrm{cov}(\mathbf{u}_i)$ perturb the eigenstructure of the factor covariance $\mathbf{BB}^\top$. In addition, the situation is more complicated with the presence of heavy-tailed data, missing data, computational constraints, heterogeneity, etc.

The rest of the paper is devoted to solutions to these challenges and a wide range of applications to statistical machine learning problems. In Section 2, we will elucidate the relationship between factor models and PCA and present several useful deterministic perturbation bounds for eigenspaces. We will also discuss robust covariance inputs for the PCA procedure to guard against corruption from heavy-tailed data. Exploiting the factor structure of the data helps solve many statistical and machine learning problems. In Section 3, we will see how the factor models and PCA can be applied to high-dimensional covariance estimation, regression, multiple testing and model selection. In Section 4, we demonstrate the connection between PCA and a wide range of machine learning problems including Gaussian mixture models, community detection, matrix completion, etc. We will develop useful tools and establish strong theoretical guarantees for our proposed methods.

Here we collect all the notations for future convenience. We use $[m]$ to refer to $\{1, 2, \ldots, m\}$. We adopt the convention of using regular letters for scalars and using bold-face letters for vectors or matrices. For $\mathbf{x} = (x_1, \ldots, x_p)^\top \in \mathbb{R}^p$, and $1 \le q < \infty$, we define $\|\mathbf{x}\|_q = \left(\sum_{j=1}^p |x_j|^q\right)^{1/q}$, $\|\mathbf{x}\|_0 = |\mathrm{supp}(\mathbf{x})|$, where $\mathrm{supp}(\mathbf{x}) = \{j : x_j \ne 0\}$, and $\|\mathbf{x}\|_\infty = \max_{1 \le j \le p} |x_j|$. For a matrix $\mathbf{M}$, we use $\|\mathbf{M}\|_2, \|\mathbf{M}\|_F, \|\mathbf{M}\|_{\max}$ and $\|\mathbf{M}\|_1$ to denote its operator norm (spectral norm), Frobenius norm, entry-wise (element-wise) max-norm, and vector ℓ_1 norm, respectively. To be more specific, the last two norms are defined by $\|\mathbf{M}\|_{\max} = \max_{j,k} |M_{jk}|$ and $\|\mathbf{M}\|_1 = \sum_{j,k} |M_{jk}|$. Let $\mathbf{I}_p$ denote the $p \times p$ identity matrix, $\mathbf{1}_p$ denote the p-dimensional all-one vector, and $\mathbb{1}_A$ denote the indicator of event A, i.e., $\mathbb{1}_A = 1$ if A happens, and 0 otherwise. We use $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to refer to the normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. For two nonnegative numbers a and b that possibly depend on n and p, we use the notation $a = O(b)$ and $a \lesssim b$ to mean $a \le C_1 b$ for some constant $C_1 > 0$, and the notation $a = \Omega(b)$ and $a \gtrsim b$ to mean $a \ge C_2 b$ for some constant $C_2 > 0$. We write $a \asymp b$ if both $a = O(b)$ and $a = \Omega(b)$ hold. For a sequence of random variables $\{X_n\}_{n=1}^\infty$ and a sequence of nonnegative deterministic numbers $\{a_n\}_{n=1}^\infty$,

we write $X_n = O_{\mathbb{P}}(a_n)$ if for any $\varepsilon > 0$, there exists $C > 0$ and $N > 0$ such that $\mathbb{P}(|X_n| \geq Ca_n) \leq \varepsilon$ holds for all $n > N$; and we write $X_n = o_{\mathbb{P}}(a_n)$ if for any $\varepsilon > 0$ and $C > 0$, there exists $N > 0$ such that $\mathbb{P}(|X_n| \geq Ca_n) \leq \varepsilon$ holds for all $n > N$. We omit the subscripts when it does not cause confusion.

2. FACTOR MODELS AND PCA

2.1 Relationship between PCA and factor models in high dimensions

Under model (1.1) with the identifiability condition, $\boldsymbol{\Sigma} = \mathrm{cov}(\mathbf{x}_i)$ is given by

$$(2.1) \qquad \boldsymbol{\Sigma} = \mathbf{B}\mathbf{B}^\top + \boldsymbol{\Sigma}_u, \quad \boldsymbol{\Sigma}_u = (\sigma_{u,jk})_{1 \leq j,k \leq p} = \mathrm{cov}(\mathbf{u}_i).$$

Intuitively, if the magnitude of $\mathbf{B}\mathbf{B}^\top$ dominates $\boldsymbol{\Sigma}_u$, the top-K eigenspace of $\boldsymbol{\Sigma}$ should be approximately aligned with the column space of $\mathbf{B}$. Naturally we expect a large gap between the eigenvalues of $\mathbf{B}\mathbf{B}^\top$ and $\boldsymbol{\Sigma}_u$ to be important for estimating the column space of $\mathbf{B}$ through PCA (see Figure 1). On the other hand, if this gap is small compared with the eigenvalues of $\boldsymbol{\Sigma}_u$, it is known that PCA leads to inconsistent estimation (Johnstone and Lu, 2009). The above discussion motivates a simple vanilla PCA-based method for estimating $\mathbf{B}$ and $\mathbf{F}$ as follows (assuming the Identifiability Assumption).

Step 1. Obtain an estimator $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Sigma}}$ of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, e.g., the sample mean and covariance matrix or their robust versions.

Step 2. Compute the eigen-decomposition of $\widehat{\boldsymbol{\Sigma}} = \sum_{j=1}^{p} \widehat{\lambda}_j \widehat{\mathbf{v}}_j \widehat{\mathbf{v}}_j^\top$. Let $\{\widehat{\lambda}_k\}_{k=1}^{K}$ be the top K eigenvalues and $\{\widehat{\mathbf{v}}_k\}_{k=1}^{K}$ be their corresponding eigenvectors. Set $\widehat{\mathbf{V}} = (\widehat{\mathbf{v}}_1, \ldots, \widehat{\mathbf{v}}_K) \in \mathbb{R}^{p \times K}$ and $\widehat{\boldsymbol{\Lambda}} = \mathrm{diag}(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_K) \in \mathbb{R}^{K \times K}$.

Step 3. Obtain PCA estimators $\widehat{\mathbf{B}} = \widehat{\mathbf{V}}\widehat{\boldsymbol{\Lambda}}^{1/2}$ and $\widehat{\mathbf{F}} = (\mathbf{X} - \widehat{\boldsymbol{\mu}}\mathbf{1}^\top)^\top \widehat{\mathbf{V}}\widehat{\boldsymbol{\Lambda}}^{-1/2}$, namely, $\widehat{\mathbf{B}}$ consists of the top-K rescaled eigenvectors of $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\mathbf{f}}_i$ is just the rescaled projection of $\mathbf{x}_i - \widehat{\boldsymbol{\mu}}$ onto the space spanned by the eigen-space: $\widehat{\mathbf{f}}_i = \widehat{\boldsymbol{\Lambda}}^{-1/2}\widehat{\mathbf{V}}^T(\mathbf{x}_i - \widehat{\boldsymbol{\mu}})$.
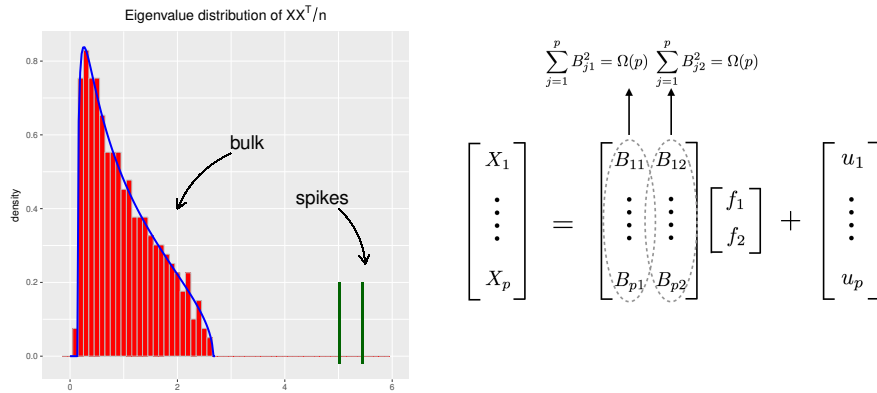

FIG 1. *The **left panel** is the histogram of the eigenvalue distribution from a synthetic dataset. Fix $n = 1000$, $p = 400$ and $K = 2$ and let all the entries of $\mathbf{B}$ be i.i.d. Gaussian $\mathcal{N}(0, 1/4)$. Each entry of $\mathbf{F}$ and $\mathbf{U}$ is generated from i.i.d. $\mathcal{N}(0,1)$ and i.i.d. $\mathcal{N}(0, 5^2)$ respectively. The data matrix $\mathbf{X}$ is formed according to the factor model (1.1). The **right diagram** illustrates the Pervasiveness Assumption.*

Let us provide some intuitions for the estimators in Step 3. Recall that $\mathbf{b}_j$ is the jth column of $\mathbf{B}$. Then, by model (1.1), $\mathbf{B}^\top(\mathbf{x}_i - \boldsymbol{\mu}) = \mathbf{B}^\top\mathbf{B}\mathbf{f}_i + \mathbf{B}^\top\mathbf{u}_i$. In the high-dimensional setting, the second term is averaged out when $\mathbf{u}_i$ is weakly dependent across its component. This along with the identifiability condition delivers that

$$(2.2) \quad \mathbf{f}_i \approx \mathrm{diag}(\mathbf{B}^\top\mathbf{B})^{-1}\mathbf{B}^\top(\mathbf{x}_i - \boldsymbol{\mu}) = \mathrm{diag}(\|\mathbf{b}_1\|^2, \cdots, \|\mathbf{b}_K\|^2)^{-1}\mathbf{B}^\top(\mathbf{x}_i - \boldsymbol{\mu}).$$

Now, we estimate $\mathbf{B}\mathbf{B}^\top$ by $\sum_{j=1}^{K} \widehat{\lambda}_j \widehat{\mathbf{v}}_j \widehat{\mathbf{v}}_j^\top$ and hence $\mathbf{b}_j$ by $\widehat{\lambda}_j^{1/2}\widehat{\mathbf{v}}_j$ and $\|\mathbf{b}_j\|^2$ by $\widehat{\lambda}_j$. Using the substitution method, we obtain the estimators in Step 3.

The above heuristic also reveals that the PCA-based methods work well if the effect of the factors outweighs the noise. To quantify this, we introduce a form of Pervasiveness Assumption from the factor model literature. While this assumption is strong[1], it simplifies our discussion and captures the above intuition well: it holds when the factor loadings $\{\mathbf{b}_j\}_{j=1}^{p}$ are random samples from a nondegenerate population (Fan et al., 2013).

ASSUMPTION 2.1 (Pervasiveness). *The first K eigenvalues of $\mathbf{B}\mathbf{B}^\top$ have order $\Omega(p)$, whereas $\|\boldsymbol{\Sigma}_u\|_2 = O(1)$.*

Note that $\mathrm{cov}(\mathbf{f}_i) = \mathbf{I}_K$ under the Identifiability Assumption 1.1. The first part of this assumption holds when each factor influences a non-vanishing proportion of outcomes. Mathematically speaking, it means that for any $k \in [K] := \{1, 2, \ldots, K\}$, the average of squared loadings of the kth factor satisfies $p^{-1}\sum_{j=1}^{p} B_{jk}^2 = \Omega(1)$ (right panel of Figure 1). This holds with high probability if, for example, $\{B_{jk}\}_{j=1}^{p}$ are i.i.d. realizations from a non-degenerate distribution, but we will not make such assumption in this paper. The second part of the assumption is reasonable, as cross-sectional correlation becomes weak after we take out the common factors. Typically, if $\boldsymbol{\Sigma}_u$ is a sparse matrix, the norm bound $\|\boldsymbol{\Sigma}_u\|_2 = O(1)$ holds; see Section 3.1 for details. Under this Pervasiveness Assumption, the first K eigenvalues of $\boldsymbol{\Sigma}$ will be well separated with the rest of eigenvalues. By the Davis-Kahan theorem (Davis and Kahan, 1970), which we present as Theorem 2.1, we can consistently estimate the column space of $\mathbf{B}$ through the top-K eigenspace of $\boldsymbol{\Sigma}$. This explains why we can apply PCA to factor model analysis (Fan et al., 2013).

Though factor models and PCA are not identical (see Jolliffe, 1986), they are approximately the same for high-dimensional problems with the pervasiveness assumption(Fan et al., 2013). Thus, PCA-based ideas are important components of estimation and inference for factor models. In later sections (especially Section 4), we discuss statistical and machine learning problems with factor-model-type structures. There PCA is able to achieve consistent estimation even when the Pervasiveness Assumption is weakened—and somewhat surprisingly—PCA can work well down to the information limit. For perspectives from random matrix theory, see Baik et al. (2005); Paul (2007); Johnstone and Lu (2009); Benaych-Georges and Nadakuditi (2011); O'Rourke et al. (2016); Wang and Fan (2017), among others.

---

[1]There is a weaker assumption, under which (1.1) is usually called the weak factor model; see Onatski (2012).

2.2 Estimating the number of factors

In high-dimensional factor models, if the factors are unobserved, we need to choose the number of factors K before estimating the loading matrix, factors, etc. The number K can be usually estimated from the eigenvalues of the the sample covariance matrix or its robust version. With certain conditions such as separation of the top K eigenvalues from the others, the estimation is consistent. Classical methods include likelihood ratio tests (Bartlett, 1950), the scree plot (Cattell, 1966), parallel analysis (Horn, 1965), etc. Here, we introduce a few recent methods: the first one is based on the eigenvalue ratio, the second on eigenvalue differences, and the third on the eigenvalue magnitude.

For simplicity, let us use the sample covariance and arrange its eigenvalues in descending order: $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{n \wedge p}$, where $n \wedge p = \min\{n, p\}$ (the remaining eigenvalues, if any, are zero). Lam and Yao (2012) and Ahn and Horenstein (2013) proposed an estimator $\widehat{K}_1$ based on ratios of consecutive eigenvalues. For a pre-determined $k_{\max}$, the eigenvalue ratio estimator is

$$\widehat{K}_1 = \operatorname*{argmax}_{i \leq k_{\max}} \frac{\lambda_i}{\lambda_{i+1}}.$$

Intuitively, when the signal eigenvalues are well separated from the other eigenvalues, the ratio at $k = K$ should be large. Under some conditions, the consistency of this estimator, which does not involve complicated tuning parameters, is established.

In an earlier work, Onatski (2010) proposed to use the differences of consecutive eigenvalues. For a given $\delta > 0$ and pre-determined integer $k_{\max}$, define

$$\widehat{K}_2(\delta) = \max\{i \leq k_{\max} : \lambda_i - \lambda_{i+1} \geq \delta\}.$$

Using a result on eigenvalue empirical distribution from random matrix theory, Onatski (2010) proved consistency of $\widehat{K}_2(\delta)$ under the Pervasiveness Assumption. The intuition is that, the Pervasiveness Assumption implies that $\lambda_K - \lambda_{K+1}$ tends on ∞ in probability as $n \to \infty$; whereas $\lambda_i - \lambda_{i+1} \to 0$ almost surely for $K < i < k_{\max}$ because these λ_i-s converge to the same limit, which can be determined using random matrix theory. Onatski (2010) also proposed a data-driven way to determine δ from the empirical eigenvalue distribution of the sample covariance matrix.

A third possibility is to use an information criterion. Define

$$V(k) = \frac{1}{np} \min_{\widehat{\mathbf{B}} \in \mathbb{R}^{p \times k}, \widehat{\mathbf{F}} \in \mathbb{R}^{n \times k}} \|\mathbf{X} - \widehat{\boldsymbol{\mu}} \mathbf{1}_n^\top - \widehat{\mathbf{B}} \widehat{\mathbf{F}}^\top\|_F^2 = p^{-1} \sum_{j>k} \lambda_j,$$

where $\widehat{\boldsymbol{\mu}}$ is the sample mean, and the equivalence (second equality) is well known. For a given k, $V(k)$ is interpreted as the scaled sum of squared residuals, which measures how well k factors fit the data. A very natural estimator $\widehat{K}_3$ is to find the best $k \leq k_{\max}$ such that the following penalized version of $V(k)$ is minimized (Bai and Ng, 2002):

$$PC(k) = V(k) + k \widehat{\sigma}^2 g(n, p), \qquad \text{where} \quad g(n, p) := \frac{n+p}{np} \log\left(\frac{np}{n+p}\right),$$

and $\widehat{\sigma}^2$ is any consistent estimate of $(np)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{d}\mathbb{E}u_{ji}^2$. The upper limit $k_{\max}$ is assumed to be no smaller than K, and is typically chosen as 8 or 15 in empirical studies in Bai and Ng (2002). Consistency results are established under more general choices of $g(n, p)$.

We conclude this section by remarking that in general, it is impossible to consistently estimate K if the smallest nonzero eigenvalue $\mathbf{B}^\top\mathbf{B}$ is much smaller than $\|\mathbf{\Sigma}_u\|_2$, because the 'signals' (eigenvalues of $\mathbf{B}^\top\mathbf{B}$) would not be distinguishable from the the noise (eigenvalues of $\mathbf{U}\mathbf{U}^\top$). As mentioned before, consistency of PCA is well studied in the random matrix theory literature. See Dobriban (2017) for a recent work that justifies parallel analysis using random matrix theory.

2.3 Robust covariance inputs

To extract latent factors and their factor loadings, we need an initial covariance estimator. Given independent observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ with mean zero, the sample covariance matrix, namely $\widehat{\mathbf{\Sigma}}_{\mathrm{sam}} := n^{-1}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top$, is a natural choice to estimate $\mathbf{\Sigma} \in \mathbb{R}^{p\times p}$. The finite sample bound on $\|\widehat{\mathbf{\Sigma}}_{\mathrm{sam}} - \mathbf{\Sigma}\|_2$ has been well studied in the literature (Vershynin, 2010; Tropp, 2012; Koltchinskii and Lounici, 2017). Before presenting the result from Vershynin (2010), let us review the definition of sub-Gaussian variables.

A random variable ξ is called sub-Gaussian if $\|\xi\|_{\psi_2} \equiv \sup_{q\geq 1} q^{-1/2}(\mathbb{E}|\xi|^q)^{1/q}$ is finite, in which case this quantity defines a norm $\|\cdot\|_{\psi_2}$ called the sub-Gaussian norm. Sub-Gaussian variables include as special cases Gaussian variables, bounded variables, and other variables with tails similar to or lighter than Gaussian tails. For a random vector $\boldsymbol{\xi}$, we define $\|\boldsymbol{\xi}\|_{\psi_2} := \sup_{\|\mathbf{v}\|_2=1} \|\boldsymbol{\xi}^\top\mathbf{v}\|_{\psi_2}$; we call $\boldsymbol{\xi}$ sub-Gaussian if $\|\boldsymbol{\xi}\|_{\psi_2}$ is finite.

THEOREM 2.1. *Let $\mathbf{\Sigma}$ be the covariance matrix of $\mathbf{x}_i$. Assume that $\{\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{x}_i\}_{i=1}^{n}$ are i.i.d. sub-Gaussian random vectors, and denote $\kappa = \sup_{\|\mathbf{v}\|_2=1} \|\mathbf{x}_i^\top\mathbf{v}\|_{\psi_2}$. Then for any $t \geq 0$, there exist constants C and c only depending on κ such that*

$$(2.3) \qquad \mathbb{P}\Big(\|\widehat{\mathbf{\Sigma}}_{\mathrm{sam}} - \mathbf{\Sigma}\|_2 \geq \max(\delta, \delta^2)\|\mathbf{\Sigma}\|_2\Big) \leq 2\exp(-ct^2),$$

where $\delta = C\sqrt{p/n} + t/\sqrt{n}$.

REMARK 2.1. *The spectral-norm bound above depends on the ambient dimension p, which can be large in high-dimensional scenarios. Interested readers can refer to Koltchinskii and Lounici (2017) for a refined result that only depends on the intrinsic dimension (or effective rank) of $\mathbf{\Sigma}$.*

An important asepect of the above result is the sub-Gaussian concentration in (2.3), but this depends heavily on the sub-Gaussian or sub-exponential behaviors of observed random vectors. This condition can not be validated in high dimensions when tens of thousands of variables are collected. See Fan et al. (2016b). When the distribution is heavy-tailed[2], one cannot expect sub-Gaussian or sub-exponential behaviors of the sample covariance in the spectral norm (Catoni,

---

[2]Here, we mean it has second bounded moment when estimating the mean and has bounded fourth moment when estimating the variance.

2012). See also Vershynin (2012) and Srivastava and Vershynin (2013). There-fore, to perform PCA for heavy-tailed data, the sample covariance is not a good choice to begin with. Alternative robust estimators have been constructed to achieve better finite sample performance.

Catoni (2012), Fan et al. (2017b) and Fan et al. (2016b) approached the problem by first considering estimation of a univariate mean μ from a sequence of i.i.d random variables $X_1, \cdots, X_n$ with variance σ^2. In this case, the sample mean $\bar{X}$ provides an estimator but without exponential concentration. Indeed, by Markov inequality, we have $\mathbb{P}(|\bar{X} - \mu| \geq t\sigma/\sqrt{n}) \leq t^{-2}$, which is tight in general and has a Cauchy tail (in terms of t). On the other hand, if we truncate the data $\widetilde{X}_i = \text{sign}(X_i)\min(|X_i|, \tau)$ with $\tau \asymp \sigma\sqrt{n}$ and compute the mean of the truncated data, then we have (Fan et al., 2016b)

$$\mathbb{P}\Big(\big|\frac{1}{n}\sum_{i=1}^{n}\widetilde{X}_i - \mu\big| \geq t\frac{\sigma}{\sqrt{n}}\Big) \leq 2\exp(-ct^2),$$

for a universal constant $c > 0$. In other words, the mean of truncated data with only a finite second moment behaves very much the same as the sample mean from the normal data: both estimators have Gaussian tails (in terms of t). This sub-Gaussian concentration is fundamental in high-dimensional statistics as the sample mean is computed tens of thousands or even millions of times.

As an example, estimating the high-dimensional covariance matrix $\mathbf{\Sigma} = (\sigma_{ij})$ involves $O(p^2)$ univariate mean estimation, since the covariance can be expressed as an expectation: as $\sigma_{ij} = \mathbb{E}(X_iX_j) - \mathbb{E}(X_i)\mathbb{E}(X_j)$. Estimating each component by the truncated mean yields a covariance matrix $\widetilde{\mathbf{\Sigma}}$. Assuming the fourth moment is bounded (as the covariance itself are second moments), by using the union bound and the above concentration inequality, we can easily obtain

$$\mathbb{P}\Big(\|\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\max} \geq \sqrt{\frac{a\log p}{c'n}}\Big) \lesssim p^{2-a}$$

for any $a > 0$ and a constant $c' > 0$. In other words, with truncation, when the data have merely bounded fourth moments, we can achieve the same estimation rate as the sample covariance matrix under the Gaussian data.

Fan et al. (2016b) and Minsker (2016) independently proposed shrinkage variants of the sample covariance with sub-Gaussian behavior under the spectral norm, as long as the fourth moments of $\mathbf{X}$ are finite. For any $\tau \in \mathbb{R}^+$, Fan et al. (2016b) proposed the following shrinkage sample covariance matrix

$$(2.4) \qquad \widehat{\mathbf{\Sigma}}_s(\tau) = \frac{1}{n}\sum_{i=1}^{n}\widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top, \qquad \widetilde{\mathbf{x}}_i := (\|\mathbf{x}_i\|_4 \wedge \tau)\mathbf{x}_i/\|\mathbf{x}_i\|_4,$$

to estimate $\mathbf{\Sigma}$, where $\|\cdot\|_4$ is the ℓ_4-norm. The following theorem establishes the statistical error rate of $\widetilde{\mathbf{\Sigma}}_s(\tau)$ in terms of the spectral norm.

THEOREM 2.2.   *Suppose* $\mathbb{E}(\mathbf{v}^\top\mathbf{x}_i)^4 \leq R$ *for any unit vector* $\mathbf{v} \in \mathcal{S}^{p-1}$. *Then it holds that for any* $\delta > 0$,

$$(2.5) \qquad \mathbb{P}\Big(\|\widehat{\mathbf{\Sigma}}_s(\tau) - \mathbf{\Sigma}\|_2 \geq \sqrt{\frac{\delta Rp\log p}{n}}\Big) \leq p^{1-C\delta},$$

where $\tau \asymp \big(nR/(\delta\log p)\big)^{1/4}$ *and* C *is a universal constant.*

Applying PCA to the robust covariance estimators as described above leads to more reliable estimation of principal eigenspaces in the presence of heavy-tailed data.

In Theorem 2.2, we assume that the mean of $\mathbf{x}_i$ is zero. When this does not hold, a natural estimator of $\boldsymbol{\Sigma} = \frac{1}{2}\mathbb{E}(\mathbf{x}_1 - \mathbf{x}_2)(\mathbf{x}_1 - \mathbf{x}_2)^\top$ is to use the shrunk U-statistic (Fan et al., 2017a):

$$
\begin{aligned}
\widehat{\boldsymbol{\Sigma}}_U(\tau) &= \frac{1}{2\binom{n}{2}} \sum_{j \neq k} \frac{\psi_\tau(\|\mathbf{x}_j - \mathbf{x}_k\|_2^2)}{\|\mathbf{x}_j - \mathbf{x}_k\|_2^2} (\mathbf{x}_j - \mathbf{x}_k)(\mathbf{x}_j - \mathbf{x}_k)^\top \\
&= \frac{1}{2\binom{n}{2}} \sum_{j \neq k} \min\big(1, \tau/\|\mathbf{x}_j - \mathbf{x}_k\|_2^2\big)(\mathbf{x}_j - \mathbf{x}_k)(\mathbf{x}_j - \mathbf{x}_k)^\top,
\end{aligned}
$$

where $\psi_\tau(u) = (|u| \wedge \tau)\mathrm{sign}(u)$. When $\tau = \infty$, it reduces to the usual U-statistics. It possesses a similar concentration property to that in Theorem 2.2 with a proper choice of τ.

2.4 Perturbation bounds

In this section, we introduce several perturbation results on eigenspaces, which serve as fundamental technical tools in factor models and related learning problems. For example, in relating the factor loading matrix $\mathbf{B}$ to the principal components of covariance matrix $\boldsymbol{\Sigma}$ in (2.1), one can regard $\boldsymbol{\Sigma}$ as a perturbation of $\mathbf{B}\mathbf{B}^\top$ by an amount of $\boldsymbol{\Sigma}_u$ and take $\mathbf{A} = \mathbf{B}\mathbf{B}^\top$ and $\widetilde{\mathbf{A}} = \boldsymbol{\Sigma}$ in Theorem 2.3 below. Similarly, we can also regard a covariance matrix estimator $\widehat{\boldsymbol{\Sigma}}$ as a perturbation of $\boldsymbol{\Sigma}$ by an amount of $\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}$.

We will begin with a review of the Davis-Kahan theorem (Davis and Kahan, 1970), which is usually useful for deriving ℓ_2-type bounds (which includes spectral norm bounds) for symmetric matrices. Then, based on this classical result, we introduce entry-wise (ℓ_∞) bounds, which typically give refined results under structural assumptions. We also derive bounds for rectangular matrices that are similar to Wedin's theorem (Wedin, 1972). Several recent works on this topic can be found in Yu et al. (2014); Fan et al. (2018b); Koltchinskii and Xia (2016); Abbe et al. (2017); Zhong (2017); Cape et al. (2017); Eldridge et al. (2017).

First, for any two subspaces $\mathcal{S}$ and $\widetilde{\mathcal{S}}$ of the same dimension K in $\mathbb{R}^p$, we choose any $\mathbf{V}, \widetilde{\mathbf{V}} \in \mathbb{R}^{p \times K}$ with orthonormal columns that span $\mathcal{S}$ and $\widetilde{\mathcal{S}}$, respectively. We can measure the closeness between two subspaces though the difference between their projectors:

$$
d_2(\mathcal{S}, \widetilde{\mathcal{S}}) = \|\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^\top - \mathbf{V}\mathbf{V}^\top\|_2 \quad \text{or} \quad d_F(\mathcal{S}, \widetilde{\mathcal{S}}) = \|\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^\top - \mathbf{V}\mathbf{V}^\top\|_F.
$$

The above definitions are both proper metrics (or distances) for subspaces $\mathbf{S}$ and $\widetilde{\mathbf{S}}$ and do not depend on the specific choice of $\mathbf{V}$ and $\widetilde{\mathbf{V}}$, since $\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^\top$ and $\mathbf{V}\mathbf{V}^\top$ are projection operators. Importantly, these two metrics are connected to the well-studied notion of *canonical angles* (or principal angles). Formally, let the singular values of $\widetilde{\mathbf{V}}^\top\mathbf{V}$ be $\{\sigma_k\}_{k=1}^K$, and define the canonical angles $\theta_k = \cos^{-1}\sigma_k$ for $k = 1, \ldots, K$. It is often useful to denote the sine of the canonical (principal) angles by $\sin\boldsymbol{\Theta}(\widehat{\mathbf{V}}, \mathbf{V}) := \mathrm{diag}(\sin\theta_1, \ldots, \sin\theta_K) \in \mathbb{R}^{K \times K}$, which can be interpreted as a generalization of sine of angles between two vectors. The following identities are well known (Stewart and Sun, 1990).

$$
\|\sin\boldsymbol{\Theta}(\widetilde{\mathbf{V}}, \mathbf{V})\|_2 = d_2(\mathcal{S}, \widetilde{\mathcal{S}}), \qquad \sqrt{2}\|\sin\boldsymbol{\Theta}(\widetilde{\mathbf{V}}, \mathbf{V})\|_F = d_F(\mathcal{S}, \widetilde{\mathcal{S}}).
$$

In some cases, it is convenient to fix a specific choice of $\widetilde{\mathbf{V}}$ and $\mathbf{V}$. It is known that for both Frobenius norm and spectral norm,

$$\|\sin\mathbf{\Theta}(\widetilde{\mathbf{V}},\mathbf{V})\| \leq \min_{\mathbf{R}\in\mathcal{O}(K)} \|\widetilde{\mathbf{V}}\mathbf{R} - \mathbf{V}\| \leq \sqrt{2}\,\|\sin\mathbf{\Theta}(\widetilde{\mathbf{V}},\mathbf{V})\|,$$

where $\mathcal{O}(K)$ is the space of orthogonal matrices of size $K \times K$. The minimizer (best rotation of basis) can be given by the singular value decomposition (SVD) of $\widetilde{\mathbf{V}}^\top\mathbf{V}$. For details, see Cape et al. (2017) for example.

Now, we present the Davis-Kahan $\sin\theta$ theorem (Davis and Kahan, 1970).

THEOREM 2.3. *Suppose* $\mathbf{A}, \widetilde{\mathbf{A}} \in \mathbb{R}^{n\times n}$ *are symmetric, and that* $\mathbf{V}, \widetilde{\mathbf{V}} \in \mathbb{R}^{n\times K}$ *have orthonormal column vectors which are eigenvectors of* $\mathbf{A}$ *and* $\widetilde{\mathbf{A}}$ *respectively. Let* $\mathcal{L}(\mathbf{V})$ *be the set of eigenvalues corresponding to the eigenvectors given in* $\mathbf{V}$, *and let* $\mathcal{L}(\mathbf{V}^\perp)$ *(respectively* $\mathcal{L}(\widetilde{\mathbf{V}}^\perp)$*) be the set of eigenvalues corresponding to the eigenvectors not given in* $\mathbf{V}$ *(respectively* $\widetilde{\mathbf{V}}$*). If there exists an interval* $[\alpha,\beta]$ *and* $\delta > 0$ *such that* $\mathcal{L}(\mathbf{V}) \subset [\alpha,\beta]$ *and* $\mathcal{L}(\widetilde{\mathbf{V}}^\perp) \subset (-\infty, \alpha-\delta] \cup [\beta+\delta, +\infty)$, *then for any orthogonal-invariant norm*[3]

$$\|\sin\mathbf{\Theta}(\widetilde{\mathbf{V}},\mathbf{V})\| \leq \delta^{-1}\,\|(\widetilde{\mathbf{A}} - \mathbf{A})\mathbf{V}\|.$$

This theorem can be generalized to singular vector perturbation for rectangular matrices; see Wedin (1972). A slightly unpleasant feature of this theorem is that δ depends on the eigenvalues of both $\mathbf{A}$ and $\widetilde{\mathbf{A}}$. However, with the help of Weyl's inequality, we can immediately obtain a corollary that does not involve the eigenvalues of $\widetilde{\mathbf{A}}$. Let $\lambda_j(\cdot)$ denote the jth largest eigenvalue of a real symmetric matrix. Recall that Weyl's inequality bounds the differences between the eigenvalues of $\mathbf{A}$ and $\widetilde{\mathbf{A}}$:

$$(2.6) \qquad \max_{1\leq j\leq n} \left|\lambda_j(\widetilde{\mathbf{A}}) - \lambda_j(\mathbf{A})\right| \leq \|\widetilde{\mathbf{A}} - \mathbf{A}\|_2.$$

This inequality suggests that, if the eigenvalues in $\mathcal{L}(\widetilde{\mathbf{V}}^\perp)$ have the same ranks (in descending order) as those in $\mathcal{L}(\mathbf{V}^\perp)$, then $\mathcal{L}(\widetilde{\mathbf{V}}^\perp)$ and $\mathcal{L}(\mathbf{V}^\perp)$ are similar. Below we state our corollary, whose proof is in the appendix.

COROLLARY 2.1. *Assume the setup of the above theorem, and suppose the eigenvalues in* $\mathcal{L}(\widetilde{\mathbf{V}})$ *have the same ranks as those in* $\mathcal{L}(\mathbf{V})$. *If* $\mathcal{L}(\mathbf{V}) \subset [\alpha,\beta]$ *and* $\mathcal{L}(\mathbf{V}^\perp) \subset (-\infty, \alpha-\delta_0] \cup [\beta+\delta_0, +\infty)$ *for some* $\delta_0 > 0$, *then*

$$\|\sin\mathbf{\Theta}(\widetilde{\mathbf{V}},\mathbf{V})\|_2 \leq 2\delta_0^{-1}\|\widetilde{\mathbf{A}} - \mathbf{A}\|_2.$$

We can then use $\|\sin\mathbf{\Theta}(\widetilde{\mathbf{V}},\mathbf{V})\|_F \leq \sqrt{K}\,\|\sin\mathbf{\Theta}(\widetilde{\mathbf{V}},\mathbf{V})\|_2$ to obtain a bound under the Frobenius norm. In the special case where $\mathcal{L}(\mathbf{V}) = \{\lambda\}$ and $\mathbf{V} = \mathbf{v}$, $\widetilde{\mathbf{V}} = \widetilde{\mathbf{v}}$ reduce to vectors, we can choose $\alpha = \beta = \lambda$, and the above corollary translates into

$$(2.7) \qquad \min_{s\in\{\pm 1\}} \|\widehat{\mathbf{v}} - s\mathbf{v}\|_2 \leq \sqrt{2}\sin\theta(\widehat{\mathbf{v}},\mathbf{v}) \leq 2\sqrt{2}\,\delta_0^{-1}\|\widetilde{\mathbf{A}} - \mathbf{A}\|_2.$$

---

[3]A norm $\|\cdot\|$ is orthogonal-invariant if $\|\mathbf{U}^\top\mathbf{B}\mathbf{V}\| = \|\mathbf{B}\|$ for any matrix B and any orthogonal matrices $\mathbf{U}$ and $\mathbf{V}$.
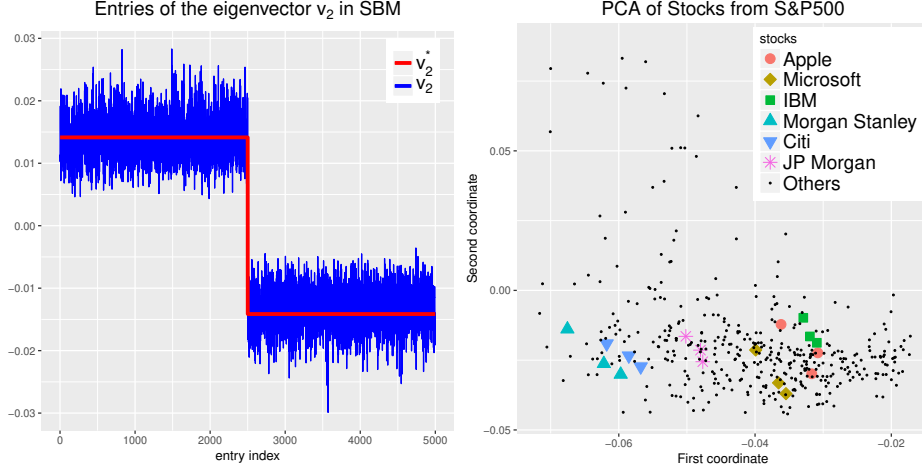
FIG 2. *The **left plot** shows the entries (coordinates) of the second eigenvectors $\mathbf{v}_2$ computed from the adjacency matrix from the SBM with two equal-sized blocks ($n = 5000, K = 2$). The plot also shows the expectation counterpart $\mathbf{v}_2^*$, whose entries have the same magnitude $O(1/\sqrt{n})$. The deviation of $\mathbf{v}_2$ from $\mathbf{v}_2^*$ is quite uniform, which is a phenomenon not captured by the Davis-Kahan's theorem. The **right plot** shows the coordinates of two leading eigenvectors of the sample covariance matrix calculated from 2012–2017 daily return data of 484 stocks (tiny black dots). We also highlight six stocks during three time windows (2012–2015, 2013–2016, 2014–2017) with big markers, so that the fluctuation/perturbation is shown. The magnitude of these coordinates is typically small, and the fluctuation is also small.*

We can now see that the factor model and PCA are approximately the same with sufficiently large eigen-gap. Indeed, under Identifiability Assumption 1.1, we have $\mathbf{\Sigma} = \mathbf{B}\mathbf{B}^\top + \mathbf{\Sigma}_u$. Applying Weyl's inequality and Corollary 2.1 to $\mathbf{B}\mathbf{B}^\top$ (as $\mathbf{A}$) and $\mathbf{\Sigma}$ (as $\widetilde{\mathbf{A}}$), we can easily control the eigenvalue/eigenvector differences by $\|\mathbf{\Sigma}_u\|_2$ and the eigengap, which is comparably small under Pervasiveness Assumption 2.1. This difference can be interpreted as the bias incurred by PCA on approximating factor models.

Furthermore, given any covariance estimator $\widehat{\mathbf{\Sigma}}$, we can similarly apply the above results by setting $\mathbf{A} = \mathbf{\Sigma}$ and $\widetilde{\mathbf{A}} = \widehat{\mathbf{\Sigma}}$ to bound the difference between the estimated eigenvalues/eigenvectors and the population counterparts. Note that the above corollary gives us an upper bound on the subspace estimation error in terms of the ratio $\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_2/\delta_0$.

Next, we consider entry-wise bounds on the eigenvectors. For simplicity, here we only consider eigenvectors corresponding to unique eigenvalues rather than the general eigenspace. Often, we want to have a bound on each entry of the eigenvector difference $\widetilde{\mathbf{v}} - \mathbf{v}$, instead of an ℓ_2 norm bound, which is an average-type result. In many cases, none of these entries has dominant perturbation, but the Davis-Kahan's theorem falls short of providing a reasonable bound (the naïve bound $\|\cdot\|_\infty \leq \|\cdot\|_2$ gives a suboptimal result).

Some recent papers (Abbe et al., 2017) have addressed this problem, and in particular, entry-wise bounds of the following form are established.

$$|[\widetilde{\mathbf{v}} - \mathbf{v}]_m| \lesssim \mu \frac{\|\widetilde{\mathbf{A}} - \mathbf{A}\|_2}{\delta_0} + \text{small term}, \qquad \forall\, m \in [n],$$

where $\mu \in [0, 1]$ is related to the structure of the statistical problem and typically

can be as small as $O(1/\sqrt{n})$, which is very desirable in high-dimensional setting. The small term is often related to independence pattern of the data, which is typically small under mild independence conditions.

We illustrate this idea in Figure 2 through a simulated data example (left) and a real data example (right), both of which have factor-type structure. For the left plot, we generated a network data according to the stochastic block model with $K = 2$ blocks (communities), each having nodes $n/2 = 2500$: the adjacency matrix that represents the links between nodes is a symmetric matrix, with upper triangular elements generated independently from Bernoulli trials (diagonal elements are taken as 0), with the edge probability $5\log n/n$ for two nodes within blocks and $\log n/(4n)$ otherwise. Our task is to classify (cluster) these two communities based on the adjacency matrix. We used the second eigenvector $\mathbf{v}_2 \in \mathbb{R}^{5000}$ (that is, corresponding to the second largest eigenvalue) of the adjacency matrix as a classifier. The left panel of Figure 2 represents the values of the 5000 coordinates (or entries) $[\mathbf{v}_2]_i$ in the y-axis against the indices $i = 1, \ldots, 5000$ in the x-axis. For comparison, the second eigenvector $\mathbf{v}_2^* \in \mathbb{R}^{5000}$ of the expectation of the adjacency matrix—which is of interest but unknown—have entries taking values only in $\{\pm 1/\sqrt{5000}\}$, depending on the unknown nature of which block a vertex belongs to (this statement is not hard to verify). We used the horizontal line to represent these ideal values: they indicate exactly the membership of each vertex. Clearly, the magnitude of entry-wise perturbation is $O(1/\sqrt{n})$. Therefore, we can use $\text{sign}(\mathbf{v}_2)/\sqrt{5000}$ as an estimate of $\mathbf{v}^*$ and classify all nodes with the same sign as the same community. See Section 4.2 for more details.

For the right plot, we used daily return data of stocks that are constituents of S&P 500 index from 2012.1.1–2017.12.31. We considered stocks with exactly $n = 1509$ records and excluded stocks with incomplete/missing values, which resulted in $p = 484$ stocks. Then, we calculated the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}_{\text{sam}} \in \mathbb{R}^{p\times p}$ using the data in the entire period, and computed two leading eigenvectors (note that they span the column space of $\mathbf{B}$) and plotted the coordinates (entries) using small dots. Stocks with an coordinate smaller than 5% quantile or larger than 95% quantile are potentially outlying values and are not shown in the plot. In addition, we also highlighted the fluctuation of six stocks during three time windows: 2012.1–2015.12, 2013.1–2016.12 and 2014.1–2017.12, with different big markers. That is, for each of the three time windows, we re-computed the covariance matrices and the two leading eigenvectors, and then highlighted coordinates that correspond to the six major stocks. Clearly, the magnitude for these stocks is small, which is roughly $O(1/\sqrt{p})$, and the fluctuation of coordinates is also very small. Both plots suggest an interesting phenomenon of eigenvectors in high dimensions: *entry-wise behavior of eigenvectors can be benign under factor model structure.*

To state our results rigorously, let us suppose that $\mathbf{A}, \widetilde{\mathbf{A}}, \mathbf{W} \in \mathbb{R}^{n\times n}$ are symmetric matrices, with $\widetilde{\mathbf{A}} = \mathbf{A} + \mathbf{W}$ and $\text{rank}(\mathbf{A}) = K < n$. Let the eigendecomposition of $\mathbf{A}$ and $\widetilde{\mathbf{A}}$ be

$$(2.8) \qquad \mathbf{A} = \sum_{k=1}^{K} \lambda_k \mathbf{v}_k \mathbf{v}_k^\top, \quad \text{and} \quad \widetilde{\mathbf{A}} = \sum_{k=1}^{K} \widetilde{\lambda}_k \widetilde{\mathbf{v}}_k \widetilde{\mathbf{v}}_k^\top + \sum_{k=K+1}^{n} \widetilde{\lambda}_k \widetilde{\mathbf{v}}_k \widetilde{\mathbf{v}}_k^\top.$$

Here the eigenvalues $\{\lambda_k\}_{k=1}^{K}$ and $\{\widetilde{\lambda}_k\}_{k=1}^{K}$ are the K largest ones of $\mathbf{A}$ and $\widetilde{\mathbf{A}}$, respectively, in terms of absolute values. Both sequences are sorted in descending

order. $\{\widetilde{\lambda}_k\}_{k=K+1}^n$ are eigenvalues of $\widetilde{\mathbf{A}}$ whose absolute values are smaller than $|\widetilde{\lambda}_K|$. The eigenvectors $\{\mathbf{v}_k\}_{k=1}^K$ and $\{\widetilde{\mathbf{v}}_k\}_{k=1}^n$ are normalized to have unit norms.

Here $\{\lambda_k\}_{k=1}^K$ are allowed to take negative values. Thanks to Weyl's inequality, $\{\widetilde{\lambda}_k\}_{k=1}^K$ and $\{\widetilde{\lambda}_k\}_{k=K+1}^n$ are well-separated when the size of perturbation $\mathbf{W}$ is not too large. In addition, we have the freedom to choose signs for eigenvectors, since they are not uniquely defined. Later, we will use 'up to sign' to signify that our statement is true for at least one choice of sign. With the conventions $\lambda_0 = +\infty$ and $\lambda_{K+1} = -\infty$, we define the eigen-gap as

$$(2.9) \qquad \delta_k = \min\{\lambda_{k-1} - \lambda_k, \lambda_k - \lambda_{k+1}, |\lambda_k|\}, \quad \forall k \in [K],$$

which is the smallest distance between λ_k and other eigenvalues (including 0). This definition coincides with the (usual) eigen-gap in Corollary 2.1 in the special case $\mathcal{L}(\mathbf{v}_k) = \{\lambda_k\}$ where we are interested in a single eigenvalue and its associated eigenvector.

We now present an entry-wise perturbation result. Let us first look at only one eigenvector. In this case, when $\|\widetilde{\mathbf{A}} - \mathbf{A}\|$ is small, heuristically,

$$\widetilde{\mathbf{v}}_k = \frac{\widetilde{\mathbf{A}}\widetilde{\mathbf{v}}_k}{\widetilde{\lambda}_k} \approx \frac{\widetilde{\mathbf{A}}\mathbf{v}_k}{\lambda_k} = \mathbf{v}_k + \frac{(\widetilde{\mathbf{A}} - \mathbf{A})\mathbf{v}_k}{\lambda_k}$$

holds uniformly for each entry. When $\mathbf{A} = \mathbb{E}\widetilde{\mathbf{A}}$, that is, $\widetilde{\mathbf{A}}$ is unbiased, this gives the first-order approximation (rather than bounds on the difference $\widetilde{\mathbf{v}}_k - \mathbf{v}_k$) of the random vector $\widetilde{\mathbf{v}}_k$. Abbe et al. (2017) proves rigorously this result and generalizes to eigenspaces. The key technique for the proof is similar to Theorem 2.4 below, which simplifies the one in Abbe et al. (2017) in various ways but holds under more general conditions. It is stated in a deterministic way, and can be powerful if there is certain structural independence in the perturbation matrix $\mathbf{W}$. A self-contained proof can be found in the appendix.

For each $m \in [n]$, let $\mathbf{W}^{(m)} \in \mathbb{R}^{n \times n}$ be a modification of $\mathbf{W}$ with the mth row and mth column zeroed out, i.e.,

$$W_{ij}^{(m)} = W_{ij}\mathbb{1}_{\{i \neq m\}}\mathbb{1}_{\{j \neq m\}}, \quad \forall i, j \in [n].$$

We also define $\widetilde{\mathbf{A}}^{(m)} = \mathbf{A} + \mathbf{W}^{(m)}$, and denote its eigenvalues and eigenvectors by $\{\widetilde{\lambda}_k^{(m)}\}_{k=1}^n$ and $\{\widetilde{\mathbf{v}}_k^{(m)}\}_{k=1}^n$, respectively. This construction is related to the leave-one-out technique in probability and statistics. For recent papers using this technique, see Bean et al. (2013); Zhong and Boumal (2018); Abbe et al. (2017) for example.

THEOREM 2.4. *Fix any $\ell \in [K]$. Suppose that $|\lambda_\ell| \asymp \max_{k \in [K]} |\lambda_k|$, and that the eigen-gap δ_ℓ as defined in (2.9) satisfies $\delta_\ell \geq 5\|\mathbf{W}\|_2$. Then, up to sign,*

$$(2.10) \qquad |[\widetilde{\mathbf{v}}_\ell - \mathbf{v}_\ell]_m| \lesssim \frac{\|\mathbf{W}\|_2}{\delta_\ell}\left(\sum_{k=1}^K [\mathbf{v}_k]_m^2\right)^{1/2} + \frac{|\langle \mathbf{w}_m, \widetilde{\mathbf{v}}_\ell^{(m)}\rangle|}{\delta_\ell}, \quad \forall m \in [n],$$

where $\mathbf{w}_m$ is the mth column of $\mathbf{W}$.

To understand this theorem, let us compare it with the standard ℓ_2 bound (Theorem 2.3) , which implies $\|\widetilde{\mathbf{v}}_\ell - \mathbf{v}_\ell\|_2 \lesssim \|\mathbf{W}\|_2/\delta_\ell$. The first term of the upper bound in (2.10) says the perturbation on the mth entry can be much smaller, because the factor $(\sum_{k=1}^{K}[\mathbf{v}_k]_m^2)^{1/2}$, always bounded by 1, can be usually much smaller. For example, if $\mathbf{v}_k$'s are uniformly distributed on the unit sphere, then this factor is typically of order $O(\sqrt{K \log n/n})$. This factor is related to the notion of incoherence in Candès and Recht (2009); Candès et al. (2011), etc.

The second term of the upper bound in (2.10) is typically much smaller than $\|\mathbf{W}\|_2/\delta_\ell$, especially under certain independence assumption. For example, if $\mathbf{w}_m$ is independent of other entries, then, by construction, $\widetilde{\mathbf{v}}_\ell^{(m)}$ and $\mathbf{w}_m$ are independent. If, moreover, entries of $\mathbf{w}_m$ are i.i.d. standard Gaussian, $|\langle \mathbf{w}_m, \widetilde{\mathbf{v}}_\ell^{(m)}\rangle|$ is of order $O_{\mathbb{P}}(1)$, whereas $\|\mathbf{W}\|_2$ typically scales with $\sqrt{n}$. This gives a bound for the mth entry, and can be extended to an ℓ_∞ bound if we are willing to make independence assumption for all $m \in [n]$ (which is typical for random graphs for example).

We remark that this result can be generalized to perturbation bounds for eigenspaces (Abbe et al., 2017), and the conditions on eigenvalues can be relaxed using certain random matrix assumptions (Koltchinskii and Xia, 2016; O'Rourke et al., 2017; Zhong, 2017).

Now, we extend this perturbation result to singular vectors of rectangular matrices. Suppose $\mathbf{L}, \widetilde{\mathbf{L}}, \mathbf{E} \in \mathbb{R}^{n \times p}$ satisfy $\widetilde{\mathbf{L}} = \mathbf{L} + \mathbf{E}$ and $\mathrm{rank}(\mathbf{L}) = K < \min\{n, p\}$. Let the SVD of $\mathbf{L}$ and $\widetilde{\mathbf{L}}$ be[4]

$$\mathbf{L} = \sum_{k=1}^{K} \sigma_k \mathbf{u}_k \mathbf{v}_k^\top \quad \text{and} \quad \widetilde{\mathbf{L}} = \sum_{k=1}^{K} \widetilde{\sigma}_k \widetilde{\mathbf{u}}_k \widetilde{\mathbf{v}}_k^\top + \sum_{k=K+1}^{\min\{n,p\}} \widetilde{\sigma}_k \widetilde{\mathbf{u}}_k \widetilde{\mathbf{v}}_k^\top,$$

where σ_k and $\widetilde{\sigma}_k$ are respectively non-increasing in k, and $\mathbf{u}_k$ and $\mathbf{v}_k$ are all normalized to have unit ℓ_2 norm. As before, let $\{\widetilde{\sigma}_k\}_{k=1}^{K}$ have K largest absolute values. Similar to (2.9), we adopt the conventions $\sigma_0 = +\infty$, $\sigma_{K+1} = 0$ and define the eigen-gap as

(2.11) $$\gamma_k = \min\{\sigma_{k-1} - \sigma_k, \sigma_k - \sigma_{k+1}\}, \quad \forall\, k \in [K].$$

For $j \in [p]$ and $i \in [n]$, we define unit vectors $\{\widetilde{\mathbf{u}}_k^{(j)}\}_{k=1}^{\min\{n,p\}} \subseteq \mathbb{R}^n$ and $\{\widetilde{\mathbf{v}}_k^{(i)}\}_{k=1}^{\min\{n,p\}} \subseteq \mathbb{R}^p$ by replacing certain row or column of $\mathbf{E}$ with zeros. To be specific, in our expression $\widetilde{\mathbf{L}} = \mathbf{L} + \mathbf{E}$, if we replace the ith row of $\mathbf{E}$ by zeros, then the normalized right singular vectors of the resulting perturbed matrix are denoted by $\{\widetilde{\mathbf{v}}_k^{(i)}\}_{k=1}^{\min\{n,p\}}$; and if we replace the jth column of $\mathbf{E}$ by zeros, then the normalized left singular vectors of the resulting perturbed matrix are denoted by $\{\widetilde{\mathbf{u}}_k^{(j)}\}_{k=1}^{\min\{n,p\}}$.

COROLLARY 2.2.   *Fix any $\ell \in [K]$. Suppose that $\sigma_\ell \asymp \max_{k \in [K]} \sigma_k$, and that*

---

[4]Here, we prefer using $\mathbf{u}_k$ to refer to the singular vectors (not to be confused with the noise term in factor models). The same applies to Section 4.

$\gamma_\ell \geq 5\|\mathbf{E}\|_2$. *Then, up to sign,*

$$|[\widetilde{\mathbf{u}}_\ell - \mathbf{u}_\ell]_i| \lesssim \frac{\|\mathbf{E}\|_2}{\gamma_\ell} \left(\sum_{k=1}^{K}[\mathbf{u}_k]_i^2\right)^{1/2} + \frac{|\langle(\mathbf{e}_i^{\mathrm{row}})^\top, \widetilde{\mathbf{v}}_\ell^{(i)}\rangle|}{\gamma_\ell}, \quad \forall i \in [n], \quad and$$

$$|[\widetilde{\mathbf{v}}_\ell - \mathbf{v}_\ell]_j| \lesssim \frac{\|\mathbf{E}\|_2}{\gamma_\ell} \left(\sum_{k=1}^{K}[\mathbf{v}_k]_j^2\right)^{1/2} + \frac{|\langle\mathbf{e}_j^{\mathrm{col}}, \widetilde{\mathbf{u}}_\ell^{(j)}\rangle|}{\gamma_\ell}, \quad \forall j \in [p],$$

where $\mathbf{e}_i^{\mathrm{row}} \in \mathbb{R}^p$ *is the ith row vector of* $\mathbf{E}$, *and* $\mathbf{e}_j^{\mathrm{col}} \in \mathbb{R}^n$ *is the jth column vector of* $\mathbf{E}$.

If we view $\widetilde{\mathbf{L}}$ as the data matrix (or observation) $\mathbf{X}$, then, the low rank matrix $\mathbf{L}$ can be interpreted as $\mathbf{B}\mathbf{F}^\top$. The above result provides a tool of studying estimation errors of the singular subspace of this low rank matrix. Note that $\widetilde{\mathbf{v}}_\ell^{(i)}$ can be interpreted as the result of removing the idiosyncratic error of the ith observation, and $\widetilde{\mathbf{u}}_\ell^{(j)}$ as the result of removing the jth covariate of the idiosyncratic error.

To better understand this result, let us consider a very simple case: $K = 1$ and each row of $\mathbf{E}$ is i.i.d. $\mathcal{N}(\mathbf{0}, \mathbf{I}_p)$. We are interested in bounding the singular vector difference between the rank-1 matrix $\mathbf{L} = \sigma_1 \mathbf{u}\mathbf{v}^\top$ and its noisy observation $\widetilde{\mathbf{L}} = \mathbf{L} + \mathbf{E}$. This is a spiked matrix model with a single spike. By independence between $\mathbf{e}_i^{\mathrm{row}}$ and $\widetilde{\mathbf{v}}_\ell^{(i)}$ as well as elementary properties of Gaussian variables, Corollary 2.2 implies that with probability $1 - o(1)$, up to sign,

$$(2.12) \qquad \|\widetilde{\mathbf{u}}_1 - \mathbf{u}_1\|_\infty \leq \frac{\|\mathbf{E}\|_2}{\sigma_1}\|\mathbf{u}_1\|_\infty + \frac{O(\sqrt{\log n})}{\sigma_1}.$$

Random matrix theory gives $\|\mathbf{E}\|_2 \asymp \sqrt{n} + \sqrt{p}$ with high probability. Our ℓ_2 perturbation inequality (Corollary 2.1) implies that $\|\widetilde{\mathbf{u}}_1 - \mathbf{u}_1\|_2 \leq \|\mathbf{E}\|_2/\sigma_1$. This upper bound is much larger than the two terms in (2.12), as $\|\mathbf{u}_1\|_\infty$ is typically much smaller than 1 in high dimensions. Thus, (2.12) gives a better entry-wise control over the ℓ_2 counterpart.

Beyond this simple case, there are many desirable features of Corollary 2.2. First of all, we allow K to be moderately large, in which case, as mentioned before, the factor $(\sum_{k=1}^{K}[\mathbf{u}_k]_i^2)^{1/2}$ is related to the incoherence structure in the matrix completion and robust PCA literature. Secondly, the result holds deterministically, so random matrices are also applicable. Finally, the result holds for each $i \in [n]$ and $j \in [p]$, and thus it is useful even if the entries of $\mathbf{E}$ are not independent, e.g. when a subset of covariates are dependent.

To sum up, our results Theorem 2.4 and Corollary 2.2 provide flexible tools of studying entry-wise perturbation of eigenvectors and singular vectors. It is also easy to adapt to other problems since their proofs are not complicated (see the appendix).

3. APPLICATIONS TO HIGH-DIMENSIONAL STATISTICS

3.1 Covariance estimation

Estimation of high-dimensional covariance matrices has wide applications in modern data analysis. When the dimensionality p exceeds the sample size n, the

sample covariance matrix becomes singular. Structural assumptions are necessary in order to obtain a consistent estimator in this challenging scenario. One typical assumption in the literature is that the population covariance matrix is sparse, with a large fraction of entries being (close to) zero, see Bickel and Levina (2008) and Cai and Liu (2011). In this setting, most variables are nearly uncorrelated. In financial and genetic data, however, the presence of common factors leads to strong dependencies among variables (Fan et al., 2008). The approximate factor model (1.1) better characterizes this structure and helps construct valid estimates. Under this model, the covariance matrix $\mathbf{\Sigma}$ has decomposition (2.1), where $\mathbf{\Sigma}_u = \mathrm{cov}(\mathbf{u}_i) = (\sigma_{u,jk})_{1 \leq j,k \leq p}$ is assumed to be sparse (Fan et al., 2013). Intuitively, we may assume that $\mathbf{\Sigma}_u$ only has a small number of nonzero entries. Formally, we require the sparsity parameter

$$m_0 := \max_{j \in [p]} \sum_{k=1}^{p} \mathbb{1} \left\{ \sigma_{u,jk} \neq 0 \right\}$$

to be small. This definition can be generalized to a weaker sense of sparsity, which is characterized by $m_q = \max_{j \in [p]} \sum_{k=1}^{p} |\sigma_{u,jk}|^q$, where $q \in (0,1)$ is a parameter. Note that small m_q forces $\mathbf{\Sigma}_u$ to have few large entries. However, for simplicity, we choose not to use this more general definition when presenting theoretical results below.

The approximate factor model has the following two important special cases, under which the parameter estimation has been well studied.

- The sparse covariance model is (2.1) without factor structure, i.e. $\mathbf{\Sigma} = \mathbf{\Sigma}_u$; typically, entry-wise thresholding is employed for estimation.
- The strict factor model corresponds to (2.1) with $\mathbf{\Sigma}_u$ being diagonal; usually, PCA-based methods are used.

The approximate factor model is a combination of the above two models, as it comprises both a low-rank component and a sparse component. A natural idea is to fuse methodologies for the two models into one, by estimating the two components using their corresponding methods. This motivated our high-level idea for estimation under the approximate factor model: (1) estimating the low-rank component (factors and loadings) using regression (when factors are observable) or PCA (when factors are latent); (2) after eliminating it from $\mathbf{\Sigma}$, employing standard techniques such as thresholding in the sparse covariance matrix literature to estimate $\mathbf{\Sigma}_u$; (3) adding the two estimated components together.

First, let us consider the scenario where the factors $\{\mathbf{f}_i\}_{i=1}^n$ are observable. In this setting, we do not need the Identifiability Assumption 1.1. Fan et al. (2008) focused on the strict factor model where the $\mathbf{\Sigma}_u$ in (2.1) is diagonal. It is then extended to the approximate factor model (1.1) by Fan et al. (2011). Later, Fan et al. (2018b) relaxed the sub-Gaussian assumption on the data to moment condition, and proposed a robust estimator. We are going to present the main idea of these methods using the one in Fan et al. (2011).

Step 1. Estimate $\mathbf{B}$ using the ordinary least-squares: $\widehat{\mathbf{B}} = (\widehat{\mathbf{b}}_1, \ldots, \widehat{\mathbf{b}}_p)^\top$ where

$$(\widehat{a}_j, \widehat{\mathbf{b}}_j) = \mathrm{argmin}_{a, \widehat{\mathbf{b}}} \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - a - \mathbf{b}^\top \mathbf{f}_i)^2.$$

Step 2. Let $\widehat{\mathbf{a}} = (\widehat{a}_1, \cdots, \widehat{a}_p)^\top$ be the vector of intercepts, $\widehat{\mathbf{u}}_i = \mathbf{x}_i - \widehat{\mathbf{a}} - \widehat{\mathbf{B}}\mathbf{f}_i$ be the vector of residual for $i \in [n]$, and $\mathbf{S}_u = \frac{1}{n}\sum_{i=1}^n \widehat{\mathbf{u}}_i\widehat{\mathbf{u}}^\top$ be the sample covariance. Apply thresholding to $\mathbf{S}_u$ and obtain a regularized estimator $\widehat{\boldsymbol{\Sigma}}_u$.

Step 3. Estimate $\mathrm{cov}(\mathbf{f}_i)$ by $\widehat{\mathrm{cov}}(\mathbf{f}_i) = \frac{1}{n}\sum_{i=1}^n (\mathbf{f}_i - \bar{\mathbf{f}})(\mathbf{f}_i - \bar{\mathbf{f}})^\top$.

Step 4. The final estimator is $\widehat{\boldsymbol{\Sigma}} = \widehat{\mathbf{B}}\widehat{\mathrm{cov}}(\mathbf{f}_i)\widehat{\mathbf{B}}^\top + \widehat{\boldsymbol{\Sigma}}_u$.

We remark that in Step 2, there are many thresholding rules for estimating sparse covariance matrices. Two popular choices are the *t*-statistic-based adaptive thresholding (Cai and Liu, 2011) and correlation-based adaptive thresholding (Fan et al., 2013), with the entry-wise thresholding level chosen to be $\omega \asymp K\sqrt{\frac{\log p}{n}}$. As the sparsity pattern of correlation and covariance are the same and the correlation matrix is scale-invariant, one typically applies the thresholding on the correlation and then scales it back to the covariance. Except for the number of factors K, this coincides with the commonly-used threshold for estimating sparse covariance matrices.

While it is not possible to achieve better convergence of $\boldsymbol{\Sigma}$ in terms of the operator norm or the Frobenius norm, Fan et al. (2011) considered two other important norms. Under regularity conditions, it is shown that

$$(3.1) \qquad \begin{aligned} \|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\boldsymbol{\Sigma}} &= O_{\mathbb{P}}\Big(m_0 K\sqrt{\frac{\log p}{n}} + \frac{K\sqrt{p}\log p}{n}\Big), \\ \|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\max} &= O_{\mathbb{P}}\Big(K\sqrt{\frac{\log p}{n}} + K^2\sqrt{\frac{\log n}{n}}\Big). \end{aligned}$$

Here for $\mathbf{A} \in \mathbb{R}^{p \times p}$, $\|\mathbf{A}\|_{\boldsymbol{\Sigma}}$ and $\|\mathbf{A}\|_{\max}$ refer to its entropy-loss norm $p^{-1/2}\|\boldsymbol{\Sigma}^{-1/2}\mathbf{A}\boldsymbol{\Sigma}^{-1/2}\|_F$ and entry-wise max-norm $\max_{i,j}|A_{ij}|$. As is pointed out by Fan et al. (2011) and Wang and Fan (2017), they are relevant to portfolio selection and risk management. In addition, convergence rates for $\|\widehat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1}\|_2$, $\|\widehat{\boldsymbol{\Sigma}}_u - \boldsymbol{\Sigma}_u\|_2$ and $\|\widehat{\boldsymbol{\Sigma}}_u^{-1} - \boldsymbol{\Sigma}_u^{-1}\|_2$ are also established.

Now we come to covariance estimation with latent factors. As is mentioned in Section 2.1, the Pervasiveness Assumption 2.1 helps separate the low-rank part $\mathbf{BB}^\top$ from the sparse part $\boldsymbol{\Sigma}_u$ in (2.1). Fan et al. (2013) proposed a Principal Orthogonal complEment Thresholding (POET) estimator, motivated by the relationship between PCA and factor model, and the estimation of sparse covariance matrix $\boldsymbol{\Sigma}_u$ in Fan et al. (2011). The procedure is described as follows.

Step 1. Let $\mathbf{S} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top$ be the sample covariance matrix, $\{\widehat{\lambda}_j\}_{j=1}^p$ be the eigenvalues of $\mathbf{S}$ in non-ascending order, $\{\widehat{\boldsymbol{\xi}}_j\}_{j=1}^p$ be their corresponding eigenvectors.

Step 2. Apply thresholding to $\mathbf{S}_u = \mathbf{S} - \sum_{j=1}^K \widehat{\lambda}_j\widehat{\boldsymbol{\xi}}_j\widehat{\boldsymbol{\xi}}_j^\top$ and obtain a regularized estimator $\widehat{\boldsymbol{\Sigma}}_u$.

Step 3. The final estimator is $\widehat{\boldsymbol{\Sigma}} = \sum_{j=1}^K \widehat{\lambda}_j\widehat{\boldsymbol{\xi}}_j\widehat{\boldsymbol{\xi}}_j^\top + \widehat{\boldsymbol{\Sigma}}_u$.

Here K is assumed to be known and bounded to simplify presentation and emphasize the main ideas. The methodology and theory in Fan et al. (2013) also allow using a data-driven estimate $\widehat{K}$ of K. In Step 2 above we can choose from

a large class of thresholding rules, and it is recommended to use the correlation-based adaptive thresholding. However, the thresholding level should be set to $\widetilde{\omega} \asymp \sqrt{\frac{\log p}{n}} + \frac{1}{\sqrt{p}}$. Compared to the level $\sqrt{\frac{\log p}{n}}$ we use in covariance estimation with observed factors, the extra term $1/\sqrt{p}$ here is the price we pay for not knowing the latent factors. It can be negligible when p grows much faster than n. Intuitively, thanks to the Pervasiveness Assumption, the latent factors can be estimated accurately in high dimensions. Fan et al. (2013) obtained theoretical guarantees for the POET that are similar to (3.1). The analysis allows for general sparsity patterns of $\mathbf{\Sigma}_u$ by considering m_q as the measure of sparsity for $q \in [0, 1)$.

Robust procedures handling heavy-tailed data are proposed and analyzed by Fan et al. (2018a,b). In another line of research, Li et al. (2017) considered estimation of the covariance matrix of a set of targeted variables, when additional data beyond the variables of interest are available. By assuming a factor model structure, they constructed an estimator taking advantage of all the data and justified the information gain theoretically.

The Pervasiveness Assumption rules out the case where factors are weak and the leading eigenvalues of $\mathbf{\Sigma}$ are not as large as $O(p)$. Shrinkage of eigenvalues is a powerful technique in this scenario. Donoho et al. (2013) systematically studied the optimal shrinkage in spiked covariance model where all the eigenvalues except several largest ones are assumed to be the same. Wang and Fan (2017) considered the approximate factor model, which is more general, and proposed a new version of POET with shrinkage for covariance estimation.

3.2 Principal component regression with random sketch

Principal component regression (PCR), first proposed by Hotelling (1933) and Kendall (1965), is one of the most popular methods of dimension reduction in linear regression. It employs the principal components of the predictors $\mathbf{x}_i$ to explain or predict the response y_i. Why do principal components, not other components, have more prediction power? Here we offer an insight from the perspective of high-dimensional factor models.

The basic assumption is that the unobserved latent factors $\mathbf{f}_i \in \mathbb{R}^K$ drive simultaneously the covariates via (1.1) and responses, as shown in Figure 3. As a specific example, we assume

$$y_i = \boldsymbol{\theta}^{*\top} \mathbf{f}_i + \varepsilon_i, \quad i = 1, \ldots, n, \quad \text{or in matrix form}, \quad \mathbf{y} = \mathbf{F}\boldsymbol{\theta}^* + \boldsymbol{\varepsilon},$$

where $\mathbf{y} = (y_1, \ldots, y_n)^\top$ and the noise $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ has zero means. Since $\mathbf{f}_i$ is latent and the covariate vector is high dimensional, we naturally infer the latent factors from the observed covariates via PCA. This yields the PCR.

By (2.2) (assume $\boldsymbol{\mu} = 0$ for simplicity), $y_i \approx (\boldsymbol{\beta}^\dagger)^\top \mathbf{x}_i + \varepsilon_i$, where $\boldsymbol{\beta}^\dagger := \mathbf{B}(\mathbf{B}^\top \mathbf{B})^{-1} \boldsymbol{\theta}^* \in \mathbb{R}^p$. This suggests that if we directly regress y_i over $\mathbf{x}_i$, then the regression coefficient $\boldsymbol{\beta}^\dagger$ should lie in the column space spanned by $\mathbf{B}$. This inspires the core idea of PCR, i.e., instead of seeking the least square estimator in the entire $\mathbb{R}^p$ space, we restrict our search scope to be the left leading singular space of $\mathbf{X}$, which is approximately the column space of $\mathbf{B}$ under the Pervasiveness Assumption.

Let us discuss PCR more rigorously. To be consistent with the rest of this paper, we let $\mathbf{X} \in \mathbb{R}^{p \times n}$, which is different from conventions, and

(3.2) $y_i = \mathbf{x}_i^\top \boldsymbol{\beta}^* + \varepsilon_i, \quad i = 1, \ldots, n, \quad \text{or in matrix form}, \quad \mathbf{y} = \mathbf{X}^\top \boldsymbol{\beta}^* + \boldsymbol{\varepsilon}.$
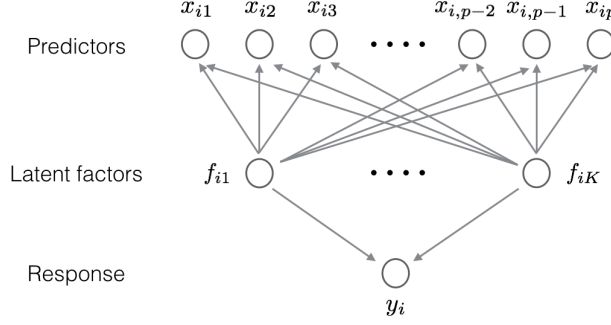
FIG 3. *Illustration of the data generation mechanism in PCR. Both predictors $\mathbf{x}_i$ and responses y_i are driven by the latent factors $\mathbf{f}_i$. PCR extracts latent factors via the principal components from $\mathbf{X}$, and uses the resulting estimate $\widehat{\mathbf{F}}$ as the new predictor. Regressing $\mathbf{y}$ against $\widehat{\mathbf{F}}$ leads to the PCR estimator $\widehat{\theta} \in \mathbb{R}^K$, which typically enjoys a smaller variance due to its reduced dimension, though it introduces bias.*

Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) = \mathbf{P}\mathbf{\Sigma}\mathbf{Q}^\top$ be the SVD of $\mathbf{X}$, where $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_{\min(n,p)})$ with non-increasing singular values. For some integer K satisfying $1 \leq K \leq \min(n,p)$, write $\mathbf{P} = (\mathbf{P}_K, \mathbf{P}_{K+})$ and $\mathbf{Q} = (\mathbf{Q}_K, \mathbf{Q}_{K+})$. The PCR estimator $\widehat{\boldsymbol{\beta}}_K$ solves the following optimization problem:

$$(3.3) \qquad \widehat{\boldsymbol{\beta}}_K := \mathrm{argmin}_{\mathbf{P}_{K+}^\top \boldsymbol{\beta} = \mathbf{0}} \ \|\mathbf{y} - \mathbf{X}^\top \boldsymbol{\beta}\|_2.$$

It is easy to verify that

$$(3.4) \qquad \widehat{\boldsymbol{\beta}}_K = \mathbf{P}_K \mathbf{\Sigma}_K^{-1} \mathbf{Q}_K^\top \mathbf{y} = \mathbf{P}_K \mathbf{P}_K^\top \boldsymbol{\beta}^* + \mathbf{P}_K \mathbf{\Sigma}_K^{-1} \mathbf{Q}_K^\top \boldsymbol{\varepsilon},$$

where $\mathbf{\Sigma}_K \in \mathbb{R}^{K \times K}$ is the top left submatrix of $\mathbf{\Sigma}$. The following lemma calculates the excess risk of $\widehat{\boldsymbol{\beta}}_K$, i.e., $\mathcal{E}(\widehat{\boldsymbol{\beta}}_K) := \mathbb{E}_{\boldsymbol{\varepsilon}}[\|\mathbf{X}^\top \widehat{\boldsymbol{\beta}}_K - \mathbf{X}^\top \boldsymbol{\beta}^*\|_2^2 / n]$, treating $\mathbf{X}$ as fixed. The proof is relegated to the appendix.

LEMMA 3.1. *Let $\mathbf{p}_1, \ldots, \mathbf{p}_{\min\{n,p\}} \in \mathbb{R}^p$ be the column vectors of $\mathbf{P}$. For $j = 1, \ldots, p$, denote $\alpha_j = (\boldsymbol{\beta}^*)^\top \mathbf{p}_j$. We have*

$$\mathcal{E}(\widehat{\boldsymbol{\beta}}_K) = \frac{K\sigma^2}{n} + \sum_{j=K+1}^{p} \lambda_j^2 \alpha_j^2.$$

Define the ordinary least squares (OLS) estimator $\widehat{\boldsymbol{\beta}} := (\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\mathbf{y}$. Note that $\mathcal{E}(\widehat{\boldsymbol{\beta}}) = \mathbb{E}_{\boldsymbol{\varepsilon}}[\|\mathbf{X}^\top \widehat{\boldsymbol{\beta}} - \mathbf{X}^\top \boldsymbol{\beta}^*\|_2^2 / n] = \min(n,p)\sigma^2 / n$. Comparing $\mathcal{E}(\widehat{\boldsymbol{\beta}}_K)$ and $\mathcal{E}(\widehat{\boldsymbol{\beta}})$, one can clearly see a variance-bias tradeoff: PCR reduces the variance by introducing a bias term $\sum_j \lambda_j^2 \alpha_j^2$, which is typically small and vanishes in the ideal case $\mathbf{P}_{K+}^\top \boldsymbol{\beta}^* = \mathbf{0}$ —this is the bias incurred by imposing the constraint in (3.3).

In the high-dimensional setting where p is large, calculating $\mathbf{P}_K$ using SVD is computationally expensive. Recently, sketching has gained growing attention in statistics community and is used for downscaling and accelerating inference tasks with massive data. See recent surveys by Woodruff (2014) and Yang et al. (2016). The essential idea is to multiply the data matrix by a sketch matrix to reduce

its dimension while still preserving the statistical performance of the procedure, since random projection reduces the strength of the idiosyncratic noise. To apply sketching to PCR, we first multiply the design matrix $\mathbf{X}$ by an appropriately chosen matrix $\mathbf{R} \in \mathbb{R}^{p \times m}$ with $K \leq m < p$:

$$\widetilde{\mathbf{X}} := \mathbf{R}^\top \mathbf{X}, \tag{3.5}$$

where $\mathbf{R}$ is called the "sketching matrix". This creates m indices based on $\mathbf{X}$. From the factor model perspective (assuming $\boldsymbol{\mu} = 0$), with a proper choice of $\mathbf{R}$, we have $\widetilde{\mathbf{X}} \approx \mathbf{R}^\top \mathbf{B} \mathbf{F}^\top$, since the idiosyncratic components in (1.1) is averaged out due to weak dependence of $\mathbf{u}$. Hence, the indices in $\widetilde{\mathbf{X}}$ are approximately linear combinations of the factors $\{\mathbf{f}_i\}_{i=1}^n$. At the same time, since $m \geq K$ and $\mathbf{R}$ is nondegenerate, the row space of $\widetilde{\mathbf{X}}$ is approximately the same as that spanned by $\mathbf{F}^\top$. This shows running linear regression on $\widetilde{\mathbf{X}}$ is approximately the same as running it on $\mathbf{F}^\top$, without using the computationally expensive PCA.

We now examine the property of sketching approach beyond the factor models. Let $\widetilde{\mathbf{X}} = \widetilde{\mathbf{P}} \widetilde{\boldsymbol{\Sigma}} \widetilde{\mathbf{Q}}^\top$ be the SVD of $\widetilde{\mathbf{X}}$, and write $\widetilde{\mathbf{P}} = (\widetilde{\mathbf{P}}_K, \widetilde{\mathbf{P}}_{K+})$ and $\widetilde{\mathbf{Q}} = (\widetilde{\mathbf{Q}}_K, \widetilde{\mathbf{Q}}_{K+})$. Imitating the form of (3.4), we consider the following sketched PCR estimator:

$$\widetilde{\boldsymbol{\beta}}_K := \mathbf{R} \widetilde{\mathbf{P}}_K \widetilde{\boldsymbol{\Sigma}}_K^{-1} \widetilde{\mathbf{Q}}_K^\top \mathbf{y}, \tag{3.6}$$

where $\widetilde{\boldsymbol{\Sigma}}_K \in \mathbb{R}^{K \times K}$ is the top left submatrix of $\widetilde{\boldsymbol{\Sigma}}$.

We now explain the above construction for $\widetilde{\boldsymbol{\beta}}_K$. It is easy to derive from (3.4) that given $\mathbf{R}^\top \mathbf{X}$ and $\mathbf{y}$ as the design matrix and response vector, the PCR estimator should be $\widetilde{\boldsymbol{\beta}}_K^0 := \widetilde{\mathbf{P}}_K \widetilde{\boldsymbol{\Sigma}}_K^{-1} \widetilde{\mathbf{Q}}_K^\top \mathbf{y}$. Then the corresponding PCR projection of $\mathbf{y}$ onto $\mathbf{R}^\top \mathbf{X}$ should be $\mathbf{X}^\top \mathbf{R} \widetilde{\boldsymbol{\beta}}_K^0 = \mathbf{X}^\top \mathbf{R} \widetilde{\mathbf{P}}_K \widetilde{\boldsymbol{\Sigma}}_K^{-1} \widetilde{\mathbf{Q}}_K^\top \mathbf{y} = \mathbf{X}^\top \widetilde{\boldsymbol{\beta}}_K$. This leads to the construction of $\widetilde{\boldsymbol{\beta}}_K$ in (3.6). Theorem 4 in Mor-Yosef and Avron (2018) gives the excess risk of $\widetilde{\boldsymbol{\beta}}_K$, which holds for any $\mathbf{R}$ satisfying the conditions of the theorem.

THEOREM 3.1.   *Assume $m \geq K$ and $\mathrm{rank}(\mathbf{R}^\top \mathbf{X}) \geq K$. If $\|\sin \Theta(\widetilde{\mathbf{P}}_K, \mathbf{P}_K)\|_2 \leq \nu < 1$, then*

$$\mathcal{E}(\widetilde{\boldsymbol{\beta}}_K) \leq \mathcal{E}(\widehat{\boldsymbol{\beta}}_K) + \frac{(2\nu + \nu^2)\|\mathbf{X}^\top \boldsymbol{\beta}^*\|_2^2}{n}. \tag{3.7}$$

This theorem shows that the extra bias induced by sketching is $(2\nu + \nu^2)\|\mathbf{X}^\top \boldsymbol{\beta}^*\|_2^2 / n$. Given the bound of $\mathcal{E}(\widehat{\boldsymbol{\beta}}_K)$ in Lemma 3.1, we can deduce that

$$\mathcal{E}(\widetilde{\boldsymbol{\beta}}_K) \leq \frac{K\sigma^2}{n} + \sum_{j=K+1}^p \alpha_j^2 \sigma_j^2 + \frac{(2\nu + \nu^2)\|\mathbf{X}^\top \boldsymbol{\beta}^*\|_2^2}{n}.$$

As we will see below, a smaller ν requires a larger m, and thus more computation. Therefore, we observe a tradeoff between statistical accuracy and computational resources: if we have more computational resources, we can allow a large dimension of sketched matrix $\widetilde{\mathbf{X}}$, and the sketched PCR is more accurate, and vice versa.

One natural question thus arises: which $\mathbf{R}$ should we choose to guarantee a small ν to retain the statistical rate of $\widehat{\boldsymbol{\beta}}_K$? Recent results (Cohen et al., 2015) on

approximate matrix multiplication (AMM) suggest several candidate sketching matrices for $\mathbf{R}$. Define the stable rank $\mathrm{sr}(\mathbf{X}) := \|\mathbf{X}\|_F^2/\|\mathbf{X}\|_2^2$, which can be interpreted as a soft version of the usual rank—indeed, $\mathrm{sr}(\mathbf{X}) \leq \mathrm{rank}(\mathbf{X})$ always holds, and $\mathrm{sr}(\mathbf{X})$ can be small if $\mathbf{X}$ is approximately low-rank. An example of candidate sketching matrices for $\mathbf{R}$ is a random matrix with independent and suitably scaled sub-Gaussian entries. As long as the sketch size $m = \Omega(\mathrm{sr}(\mathbf{X}) + \log(1/\delta)/\varepsilon^2)$, it will hold for any $\varepsilon, \delta \in (0, 1/2)$ that

$$(3.8) \qquad \mathbb{P}(\|\mathbf{X}^\top \mathbf{R}\mathbf{R}^\top \mathbf{X} - \mathbf{X}^\top \mathbf{X}\|_2^2 \geq \varepsilon \|\mathbf{X}\|_2^2) \leq \delta.$$

Combining this with the Davis-Kahan Theorem (Corollary 2.1), we can deduce that $\|\sin\Theta(\widetilde{\mathbf{P}}_K, \mathbf{P}_K)\|_2$ is small with certain eigen-gap condition. We summarize our argument by presenting a corollary of Theorem 9 in Mor-Yosef and Avron (2018) below. Readers can find more candidate sketching matrices in the examples after Theorem 1 in Cohen et al. (2015).

COROLLARY 3.1. *For any* $\nu, \delta \in (0, 1/2)$, *let*

$$\varepsilon = \nu(1+\nu)^{-1}(\sigma_K^2 - \sigma_{K+1}^2)/\sigma_1^2.$$

Let $\mathbf{R} \in \mathbb{R}^{p \times m}$ *a random matrix with i.i.d.* $\mathcal{N}(0, 1/m)$ *entries. Then there exists a universal constant* $C > 0$ *such that for any* $\delta > 0$, *if* $m \geq C(\mathrm{sr}(\mathbf{X}) + \log(1/\delta)/\varepsilon^2)$, *it holds with probability at least* $1 - \delta$ *that*

$$(3.9) \qquad \mathcal{E}(\widetilde{\boldsymbol{\beta}}_K) \leq \mathcal{E}(\widehat{\boldsymbol{\beta}}_K) + \frac{(2\nu + \nu^2)\|\mathbf{X}^\top \boldsymbol{\beta}^*\|_2^2}{n}.$$

REMARK 3.1. *Note that* $\varepsilon \leq \nu(\sigma_K^2 - \sigma_{K+1}^2)/\sigma_1^2$, *and this bound is tight with a small* ν. *Some algebra yields that* (3.9) *holds when*

$$m = \Omega\Big(\mathrm{sr}(\mathbf{X}) + \frac{\sigma_1^2 \log(1/\delta)}{\nu^2(\sigma_K^2 - \sigma_{K+1}^2)^2}\Big).$$

One can see that reducing ν *requires a larger sketch size* m. *Besides, a large eigengap of the design matrix* $\mathbf{X}$ *helps reduce the required sketch size.*

3.3 Factor-Adjust Robust Multiple (FARM) tests

Large-scale multiple testing is a fundamental problem in high-dimensional inference. In genome-wide association studies and many other applications, tens of thousands of hypotheses are tested simultaneously. Standard approaches such as Benjamini and Hochberg (1995) and Storey (2002) can not control well both false and missed discovery rates in the presence of strong correlations among test statistics. Important efforts on dependence adjustment include Efron (2007), Friguet et al. (2009), Efron (2010), and Desai and Storey (2012). Fan et al. (2012) and Fan and Han (2017) considered FDP estimation under the approximate factor model. Wang et al. (2017) studied a more complicated model with both observed variables and latent factors. All these existing papers heavily rely on the joint normality assumption of the data, which is easily violated in real applications. A recent paper (Fan et al., 2017a) developed a factor-adjusted robust procedure that can handle heavy-tailed data while controlling FDP. We are going to introduce this method in this subsection.

Suppose our i.i.d. observations $\{\mathbf{x}_i\}_{i=1}^n$ satisfy the approximate factor model (1.1) where $\boldsymbol{\mu} \in \mathbb{R}^p$ is an unknown mean vector. To make the model identifiable, we use the Identifiability Assumption 1.1. We are interested in simultaneously testing

$$H_{0j} : \mu_j = 0 \quad \text{versus} \quad H_{1j} : \mu_j \neq 0, \quad \text{for } j \in [p].$$

Let T_j be a generic test statistic for H_{0j}. For a pre-specified level $z > 0$, we reject H_{0j} whenever $|T_j| \geq z$. The numbers of total discoveries $R(z)$ and false discoveries $V(z)$ are defined as

$$R(z) = \#\{j : |T_j| \geq z\} \quad \text{and} \quad V(z) = \#\{j : |T_j| \geq z, \ \mu_j = 0\}.$$

Note that $R(z)$ is observable while $V(z)$ needs to be estimated. Our goal is to control the false discovery proportion $\mathrm{FDP}(z) = V(z)/R(z)$ with the convention $0/0 = 0$.

Naïve tests based on sample averages $\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$ suffer from size distortion of FDP control due to dependence of common factors in (1.1). On the other hand, the factor-adjusted test based on the sample averages of $\mathbf{x}_i - \mathbf{B}\mathbf{f}_i$ ($\mathbf{B}$ and $\mathbf{f}_i$ need to be estimated) has two advantages: the noise $\mathbf{u}_i$ is now weakly dependent so that FDP can be controlled with high accuracy, and the variance of $\mathbf{u}_i$ is smaller than that of $\mathbf{B}\mathbf{f}_i + \mathbf{u}_i$ in model (1.1), so that it is more powerful. This will be convincingly demonstrated in Figure 5 below. The factor-adjusted robust multiple test (FarmTest) is a robust implementation of the above idea (Fan et al., 2017a), which replaces the sample mean by its adaptive Huber estimation and extracts latent factors from a robust covariance input.

To begin with, we consider the Huber loss (Huber, 1964) with the robustification parameter $\tau \geq 0$:

$$\ell_\tau(u) = \begin{cases} u^2/2, & \text{if } |u| \leq \tau, \\ \tau|u| - \tau^2/2, & \text{if } |u| > \tau \end{cases},$$

and use $\widehat{\mu}_j = \operatorname{argmax}_{\theta \in \mathbb{R}} \sum_{i=1}^n \ell_\tau(x_{ij} - \theta)$ as a robust M-estimator of μ_j. Fan et al. (2017a) suggested choosing $\tau \asymp \sqrt{n/\log(np)}$ to deal with possible asymmetric distribution and called it adaptive Huber estimator. They showed, assuming bounded fourth moments only, that

$$(3.10) \qquad \sqrt{n}(\widehat{\mu}_j - \mu_j - \mathbf{b}_j^\top \bar{\mathbf{f}}) = \mathcal{N}(0, \sigma_{u,jj}) + o_{\mathbb{P}}(1) \text{ uniformly over } j \in [p],$$

where $\bar{\mathbf{f}} = \frac{1}{n} \sum_{i=1}^n \mathbf{f}_i$, and $\sigma_{u,jj}$ is the (j,j)th entry of $\boldsymbol{\Sigma}_u$ as is defined in (2.1). Assuming for now that $\{\mathbf{b}_j\}_{j=1}^p$, $\bar{\mathbf{f}}$ and $\{\sigma_{u,jj}\}_{j=1}^p$ are all observable, then the factor-adjusted test statistic $T_j = \sqrt{n/\sigma_{u,jj}}(\widehat{\mu}_j - \mathbf{b}_j^\top \bar{\mathbf{f}})$ is asymptotically $\mathcal{N}(0,1)$. The law of large numbers implies that $V(z)$ should be close to $2p_0\Phi(-z)$ for $z \geq 0$, where $\Phi(\cdot)$ is the cumulative distribution function of $\mathcal{N}(0,1)$, and $p_0 = \#\{j : \mu_j = 0\}$ is the number of true nulls. Hence

$$\mathrm{FDP}(z) = \frac{V(z)}{R(z)} \approx \frac{2p_0\Phi(-z)}{R(z)} \leq \frac{2p\Phi(-z)}{R(z)} =: \mathrm{FDP}^{\mathrm{A}}(z).$$

Note that in the high-dimensional and sparse regime, we have $p_0 = p - o(p)$ and thus $\mathrm{FDP}^{\mathrm{A}}(z)$ is only a slightly conservative surrogate. However, we can also

estimate the proportion $\pi_0 = p_0/p$ and use less conservative estimate $\mathrm{FDP}^{\mathrm{A}}(z) = 2p\widehat{\pi}_0\Phi(-z)/R(z)$ instead, where $\widehat{\pi}_0$ is an estimate of π_0 whose idea is depicted in Figure 4; see Storey (2002). Finally, we define the critical value $z_\alpha = \inf\{z \geq 0 : \mathrm{FDP}^{\mathrm{A}}(z) \leq \alpha\}$ and reject H_{0j} whenever $|T_j| \geq z_\alpha$.
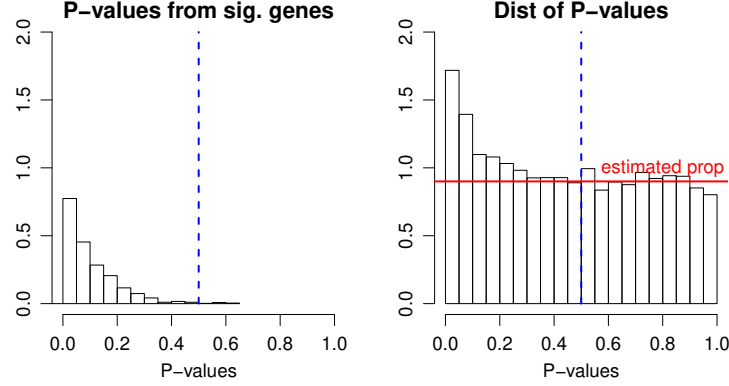

FIG 4. *Estimation of proportion of true nulls. The observed P-values (right panel) consist of those from significant variables (genes), which are usually small, and those from insignificant variables, which are uniformly distributed. Assuming the P-values for significant variables are mostly less than λ (taken to be 0.5 in this illustration, left panel), the contributions of observed P-values $> \lambda$ are mostly from true nulls and this yields a natural estimator $\widehat{\pi}_0(\lambda) = \frac{1}{(1-\lambda)p}\sum_{j=1}^{p} 1(\widehat{P}_j > \lambda)$, which is the average height of the histogram with P-values $> \lambda$ (red line). Note that the histograms above the red line estimates the distributions of P-values from the significant variavles (genes) in the left panel.*

In practice, we have no access to $\{\mathbf{b}_j\}_{j=1}^{p}$, $\bar{\mathbf{f}}$ or $\{\sigma_{u,jj}\}_{j=1}^{p}$ in (3.10) and need to use their estimates. This results in the Factor-Adjusted Robust Multiple test (FarmTest) in Fan et al. (2017a). The inputs include $\{\mathbf{x}_i\}_{i=1}^{n}$, a generic robust co-variance matrix estimator $\widehat{\mathbf{\Sigma}} \in \mathbb{R}^{p \times p}$ from the data, a pre-specified level $\alpha \in (0,1)$ for FDP control , the number of factors K, and the robustification parameters γ and $\{\tau_j\}_{j=1}^{p}$. Note that K can be estimated by the methods in Section 2.2, and overestimating K has little impact on final outputs.

Step 1. Denote by $\widehat{\mathbf{\Sigma}} \in \mathbb{R}^{p \times p}$ a generic robust covariance matrix estimator. Compute the eigen-decomposition of $\widehat{\mathbf{\Sigma}}$, set $\{\widehat{\lambda}_j\}_{j=1}^{K}$ to be its top K eigenvalues in descending order, and $\{\widehat{\mathbf{v}}_j\}_{j=1}^{K}$ to be their corresponding eigenvectors. Let $\widehat{\mathbf{B}} = (\widetilde{\lambda}_1^{1/2}\widehat{\mathbf{v}}_1, \ldots, \widetilde{\lambda}_K^{1/2}\widehat{\mathbf{v}}_K) \in \mathbb{R}^{p \times K}$ where $\widetilde{\lambda}_j = \max\{\widehat{\lambda}_j, 0\}$, and denote its rows by $\{\widehat{\mathbf{b}}_j\}_{j=1}^{p}$.

Step 2. Let $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ for $j \in [p]$ and $\widehat{\mathbf{f}} = \mathrm{argmax}_{\mathbf{f} \in \mathbb{R}^K} \sum_{j=1}^{p} \ell_\gamma(\bar{x}_j - \widehat{\mathbf{b}}_j^\top \mathbf{f})$. Construct factor-adjusted test statistics

$$(3.11) \qquad T_j = \sqrt{n/\widehat{\sigma}_{u,jj}}(\widehat{\mu}_j - \widehat{\mathbf{b}}_j^\top \widehat{\mathbf{f}}) \quad \text{for } j \in [p],$$

where $\widehat{\sigma}_{u,jj} = \widehat{\theta}_j - \widehat{\mu}_j^2 - \|\widehat{\mathbf{b}}_j\|_2^2$, $\widehat{\theta}_j = \mathrm{argmin}_{\theta \geq \widehat{\mu}_j^2 + \|\widehat{\mathbf{b}}_j\|_2^2} \ell_{\tau_j}(x_{ij}^2 - \theta)$.

Step 3. Calculate the critical value $z_\alpha = \inf\{z \geq 0 : \mathrm{FDP}^{\mathrm{A}}(z) \leq \alpha\}$, where $\mathrm{FDP}^{\mathrm{A}}(z) = 2\widehat{\pi}_0 p\Phi(-z)/R(z)$, and reject H_{0j} whenever $|T_j| \geq z_\alpha$.

In Step 2, we estimate $\bar{\mathbf{f}}$ based on $\bar{x}_j = \mu_j + \widehat{\mathbf{b}}_j^\top \bar{\mathbf{f}} + \bar{u}_j$, which is implied by the
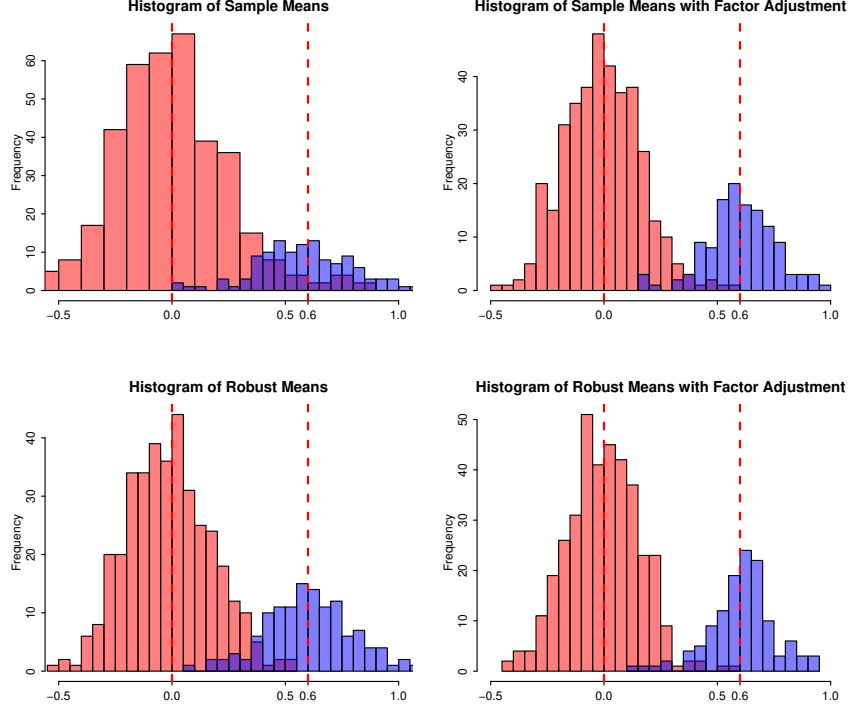
FIG 5. *Histograms of four different mean estimators for simultaneous inference. Fix $n = 100$, $p = 500$ and $K = 3$, and data are generated i.i.d. from t_3, which is heavy-tailed. Dashes lines correspond to $\mu_j = 0$ and $\mu_j = 0.6$, which is unknown. Robustification and factor adjustment help distinguish nulls and alternatives.*

factor model (1.1), and regard non-vanishing μ_j as an outlier. In the estimation of $\sigma_{u,jj}$, we used the identity $\theta_j := \mathbb{E}x_{ij}^2 = \mu_j^2 + \|\mathbf{b}_j\|^2 + \sigma_{u,jj}$ and robustly estimated the second moment θ_j.

Figure 5 is borrowed from Figure 1 in Fan et al. (2017a) that illustrates the effectiveness of this procedure. Here $n = 100$, $p = 500$, $K = 3$, $\mathbf{f}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_3)$ and the entries of $\mathbf{u}_i$ are generated independently from the t-distribution with 3 degrees of freedom. It is known that $t-$distributions are not sub-Guassian variables and are often used to model heavy-tailed data. The unknown means $\mu \in \mathbb{R}^p$ are fixed as $\mu_j = 0.6$ for $j \leq 125$ and $\mu_j = 0$ otherwise. We plot the histograms of sample means, robust mean estimators, and their counterparts with factor-adjustment. The latent factors and heavy-tailed errors make it difficult to distinguish $\mu_j = 0.6$ from $\mu_j = 0$, and that explains why the sample means behave poorly. As is shown in Figure 5, better separation can be obtained by factor adjustment and robustification.

While existing literature usually imposes the joint normal assumption on $\{\mathbf{f}_i, \mathbf{u}_i\}_{i=1}^n$, the FarmTest only requires the coordinates of $\{\mathbf{u}_i\}_{i=1}^n$ to have bounded fourth-order moments, and $\{\mathbf{f}_i\}_{i=1}^n$ to be sub-Gaussian. Under standard regularity conditions for the approximate factor model, it is proved by Fan et al. (2017a) that

$$\mathrm{FDP}^{\mathrm{A}}(z) - \mathrm{FDP}(z) = o_{\mathbb{P}}(1).$$

We see that $\mathrm{FDP}^{\mathrm{A}}$ is a valid approximation of FDP, which is therefore faithfully

controlled by the FARM-Test.

3.4 Factor-Adjusted Robust Model (FARM) selection

Model selection is one of the central tasks in high dimensional data analysis. Parsimonious models enjoy interpretability, stability and oftentimes, better prediction accuracy. Numerous methods for model selection have been proposed in the past two decades, including, Lasso (Tibshirani, 1996), SCAD (Fan and Li, 2001), the elastic net Zou and Hastie (2005), the Dantzig selector (Candes and Tao, 2007), among others. However, these methods work only when the covariates are weakly dependent or statisfy certain regularity conditions (Zhao and Yu, 2006; Bickel et al., 2009). When covariates are strongly correlated, Paul et al. (2008); Kneip and Sarda (2011); Wang (2012); Fan et al. (2016a) used factor model to eliminate the dependencies caused by pervasive factors, and to conduct model selection using the resulting weakly correlated variables.

Assume that $\{\mathbf{x}_i\}_{i=1}^n$ follow the approximate factor model (1.1). As a standard assumption, the coordinates of $\mathbf{w}_i = (\mathbf{f}_i^\top, \mathbf{u}_i^\top)^\top \in \mathbb{R}^{K+p}$ are weakly dependent. Thanks to this condition and the decomposition

$$(3.12) \qquad \mathbf{x}_i^\top \boldsymbol{\beta} = (\boldsymbol{\mu} + \mathbf{B}\mathbf{f}_i + \mathbf{u}_i)^\top \boldsymbol{\beta} = \alpha + \mathbf{u}_i^\top \boldsymbol{\beta} + \mathbf{f}_i^\top \boldsymbol{\gamma}$$

where $\alpha = \boldsymbol{\mu}^T \boldsymbol{\beta}$ and $\boldsymbol{\gamma} = \mathbf{B}^\top \boldsymbol{\beta}$. we may treat $\mathbf{w}_i$ as the new predictors. In other words, by lifting the number of variables from p to $p + K$, the covariates of $\mathbf{w}_i$ are now weakly dependent. The usual regularized estimation can now be applied to this new set of variables. Note that we regard the coefficients $\mathbf{B}^\top \boldsymbol{\beta}$ as free parameters to facilitate the implementation (ignoring the relation $\gamma = \mathbf{B}^\top \beta$) and this requires an additional assumption to make this valid (Fan et al., 2016a).

Suppose we wish to fit a model $y_i = g(\mathbf{x}_i^\top \boldsymbol{\beta}, \varepsilon_i)$ via a loss function $L(y_i, \mathbf{x}_i^\top \boldsymbol{\beta})$. The above idea suggests the following two-step approach, which is called Factor-Adjusted Regularized (or Robust when so implemented) Model selection (Farm-Select) (Fan et al., 2016a).

Step 1: Factor estimation. Fit the approximate factor model (1.1) to get $\widehat{\mathbf{B}}, \widehat{\mathbf{f}}_i$ and $\widehat{\mathbf{u}}_i = \mathbf{x}_i - \widehat{\mathbf{B}}\widehat{\mathbf{f}}_i$.

Step 2: Augmented regularization. Find α, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ to minimize

$$\sum_{i=1}^n L(y_i, \alpha + \widehat{\mathbf{u}}_i^\top \boldsymbol{\beta} + \widehat{\mathbf{f}}_i^\top \boldsymbol{\gamma}) + \sum_{j=1}^p p_\lambda(|\beta_j|),$$

where $p_\lambda(\cdot)$ is a folded concave penalty (Fan and Li, 2001) with parameter λ.

In Step 1, standard estimation procedures such as POET (Fan et al., 2013) and S-POET (Wang and Fan, 2017) can be applied, as long as they produce consistent estimators of $\mathbf{B}$, $\{\mathbf{f}_i\}_{i=1}^n$ and $\{\mathbf{u}_i\}_{i=1}^n$. Step 2 is carried out using usual regularization methods with new covariates.

Figure 6, borrowed from Figure 3 (a) in Fan et al. (2016a), shows that the proposed method outperforms other popular ones for model selection including Lasso (Tibshirani, 1996), SCAD (Fan and Li, 2001) and elastic net (Zou and Hastie, 2005), in the presence of correlated covariates. The basic setting is sparse linear regression $y = \mathbf{x}^\top \boldsymbol{\beta}^* + \varepsilon$ with $p = 500$ and n growing from 50 to 160.

The true coefficients are $\boldsymbol{\beta}^* = (\beta_1, \cdots, \beta_{10}, \mathbf{0}_{p-10})^\top$, where $\{\boldsymbol{\beta}_j\}_{j=1}^{10}$ are drawn uniformly at random from $[2, 5]$, and $\varepsilon \sim \mathcal{N}(0, 0.3)$. The correlation structure of covariates $\mathbf{x}$ is calibrated from S&P 500 monthly excess returns between 1980 and 2012.
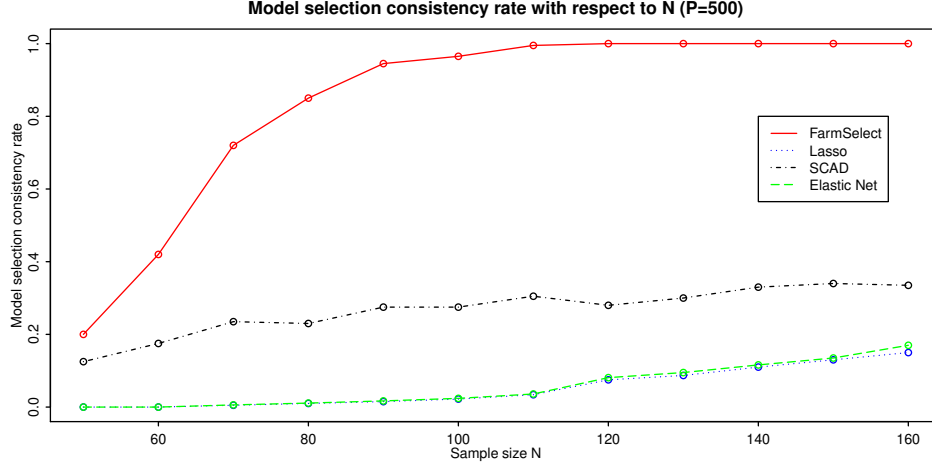
Model selection consistency rate with respect to N (P=500)

FIG 6. *Model selection consistency rate, i.e., the proportion of simulations that the selected model is identical to the true one, with $p = 500$ and n varying from 50 to 160. With moderate sample size, the proposed method faithfully identifies the correct model while other methods cannot.*

Under the generalized linear model, $L(y, z) = -yz + b(z)$ and $b(\cdot)$ is a convex function. Fan et al. (2016a) analyzed theoretical properties of the above procedure. As long as the coordinates of $\mathbf{w}_i$ (rather than $\mathbf{x}_i$) are not too strongly dependent and the factor model is estimated to enough precision, $\widehat{\boldsymbol{\beta}}$ enjoys optimal rates of convergence $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_q = O_{\mathbb{P}}(|S|^{1/q}\sqrt{\log p/n})$, where $q = 1$, 2 or ∞. When the minimum entry of $|\boldsymbol{\beta}^*|$ is at least $\Omega(\sqrt{\log p/n})$, the model selection consistency is achieved.

When we use the square loss, this method reduces to the one in Kneip and Sarda (2011). By using the square loss and replacing the penalized multiple regression in Step 2 with marginal regression, we recover the factor-profiled variable screening method in Wang (2012). While these papers aim at modeling and then eliminating the dependencies in $\mathbf{x}_i$ via (1.1), Paul et al. (2008) used a factor model to characterize the joint distribution of $(y_i, \mathbf{x}_i^\top)^\top$ and develops a related but different approach.

4. RELATED LEARNING PROBLEMS

4.1 Gaussian mixture model

PCA, or more generally, spectral decomposition, can be also applied to learn mixture models for heterogeneous data. A thread of recent papers (Hsu and Kakade, 2013; Anandkumar et al., 2014; Yi et al., 2016; Sedghi et al., 2016) apply spectral decomposition to lower-order moments of the data to recover the parameters of interest in a wide class of latent variable models. Here we use the Gaussian mixture model to illustrate their idea. Consider a mixture of K Gaussian distributions with spherical covariances. Let $w_k \in (0, 1)$ be the probability of

choosing component $k \in \{1, \ldots, K\}$, and $\{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k\} \subseteq \mathbb{R}^p$ be the component mean vectors, and $\{\sigma_k^2 \mathbf{I}_p\}_{k=1}^K$ be the component covariance matrices, which is required by Hsu and Kakade (2013) and Anandkumar et al. (2014). Each data vector $\mathbf{x} \sim w_1 \mathcal{N}(\boldsymbol{\mu}_1, \sigma_1^2 \mathbf{I}_p) + \cdots + w_K \mathcal{N}(\boldsymbol{\mu}_K, \sigma_K^2 \mathbf{I}_p)$ follows the mixture of the Gaussian distribution. The parameters of interest are $\{w_k, \boldsymbol{\mu}_k, \sigma_k^2\}_{k=1}^K$.

Hsu and Kakade (2013) and Anandkumar et al. (2014) shed lights on the close connection between the lower-order moments of the data and the parameters of interest, which motivates the use of Method of Moments (MoM). Denote the population covariance $\mathbb{E}[(\mathbf{x} - \mathbb{E}\mathbf{x})(\mathbf{x} - \mathbb{E}\mathbf{x})^\top]$ by $\boldsymbol{\Sigma}$. Below we present Theorem 1 in Hsu and Kakade (2013) to elucidate the moment structure of the problem.

THEOREM 4.1. *Suppose that $\{\boldsymbol{\mu}_k\}_{k=1}^K$ are linearly independent. Then the average variance $\sigma_{\text{ave}}^2 := K^{-1} \sum_{k=1}^K \sigma_k^2$ is the smallest eigenvalue of $\boldsymbol{\Sigma}$. Let $\mathbf{v}$ be any eigenvector of $\boldsymbol{\Sigma}$ that is associated with the eigenvalue σ_{ave}^2. Define the following quantities:*

$\mathbf{M}_1 = \mathbb{E}[(\mathbf{v}^\top(\mathbf{x} - \mathbb{E}\mathbf{x}))^2 \mathbf{x}] \in \mathbb{R}^p,$

$\mathbf{M}_2 = \mathbb{E}[\mathbf{x} \otimes \mathbf{x}] - \sigma_{\text{ave}}^2 \cdot \mathbf{I}_p \in \mathbb{R}^{p \times p},$

$\mathbf{M}_3 = \mathbb{E}[\mathbf{x} \otimes \mathbf{x} \otimes \mathbf{x}] - \sum_{j=1}^p (\mathbf{M}_1 \otimes \mathbf{e}_j \otimes \mathbf{e}_j + \mathbf{e}_j \otimes \mathbf{M}_1 \otimes \mathbf{e}_j + \mathbf{e}_j \otimes \mathbf{e}_j \otimes \mathbf{M}_1) \in \mathbb{R}^{p \times p \times p}.$

Then we have

$$(4.1) \quad \mathbf{M}_1 = \sum_{k=1}^K w_k \sigma_k^2 \boldsymbol{\mu}_k, \quad \mathbf{M}_2 = \sum_{k=1}^K w_k \boldsymbol{\mu}_k \otimes \boldsymbol{\mu}_k, \quad \mathbf{M}_3 = \sum_{k=1}^K w_k \boldsymbol{\mu}_k \otimes \boldsymbol{\mu}_k \otimes \boldsymbol{\mu}_k,$$

where the notation $\otimes$ represents the tensor product.

Theorem 4.1 gives the relationship between the moments of the first three orders of $\mathbf{x}$ and the parameters of interest. With $\{\mathbf{M}_i\}_{i=1}^3$ replaced by their empirical versions, the remaining task is to solve for all the parameters of interest via (4.1). Hsu and Kakade (2013) and Anandkumar et al. (2014) proposed a fast method called *robust tensor power method* to compute the estimators. The crux therein is to construct an estimable third-order tensor $\widetilde{\mathbf{M}}_3$ that can be decomposed as the sum of orthogonal tensors based on $\boldsymbol{\mu}_k$. This orthogonal tensor decomposition can be regarded as an extension of spectral decomposition to third-order tensors (simply speaking, three-dimensional arrays). Then the power iteration method is applied to the estimate of $\widetilde{\mathbf{M}}_3$ to recover each $\boldsymbol{\mu}_k$, as well as other parameters.

Specifically, consider first the following linear transformation of $\boldsymbol{\mu}_k$:

$$(4.2) \qquad\qquad \widetilde{\boldsymbol{\mu}}_k := \sqrt{\omega_k} \mathbf{W}^\top \boldsymbol{\mu}_k$$

for $k \in [K]$, where $\mathbf{W} \in \mathbb{R}^{p \times K}$. The key is to use the *whitening* transformation by setting $\mathbf{W}$ to be a square root of $\mathbf{M}_2$. This ensures that $\{\widetilde{\boldsymbol{\mu}}_k\}_{k=1}^K$ are orthogonal to each other. Denoting $\mathbf{a}^{\otimes 3} = \mathbf{a} \otimes \mathbf{a} \otimes \mathbf{a}$,

$$(4.3) \qquad\qquad \widetilde{\mathbf{M}}_3 := \sum_{k=1}^K \omega_k (\mathbf{W}^\top \boldsymbol{\mu}_k)^{\otimes 3} = \sum_{k=1}^K \frac{1}{\sqrt{\omega_k}} \widetilde{\boldsymbol{\mu}}_k^{\otimes 3} \in \mathbb{R}^{K \times K \times K}$$

is an orthogonal tensor decomposition; that is, it satisfies orthogonality of $\{\widetilde{\boldsymbol{\mu}}_k\}_{k=1}^K$. The following theorem from Anandkumar et al. (2014) summarizes the above argument, and more importantly, it shows how to obtain $\boldsymbol{\mu}_k$ back from $\widetilde{\boldsymbol{\mu}}_k$.

THEOREM 4.2.   *Suppose the vectors $\{\boldsymbol{\mu}_k\}_{k=1}^K$ are linearly independent, and the scalars $\{\omega_k\}_{k=1}^K$ are strictly positive. Let $\mathbf{M}_2 = \mathbf{U}\mathbf{D}\mathbf{U}^\top$ be the spectral decomposition of $\mathbf{M}_2$ and let $\mathbf{W} = \mathbf{U}\mathbf{D}^{-1/2}$. Then $\{\widetilde{\boldsymbol{\mu}}_k\}_{k=1}^K$ in (4.2) are orthogonal to each other. Furthermore, the Moore-Penrose pseudo-inverse of $\mathbf{W}$ is $\mathbf{W}^\dagger := \mathbf{D}^{1/2}\mathbf{U}^\top \in \mathbb{R}^{K \times p}$, and we have $\boldsymbol{\mu}_k = (\mathbf{W}^\dagger)^\top \widetilde{\boldsymbol{\mu}}_k / \sqrt{\omega_k}$ for $k \in [K]$.*

As promised, the orthogonal tensor $\widetilde{\mathbf{M}}_3$ can be estimated from empirical moments. We will make use of the following identity, which is similar to Theorem 4.1.

$$(4.4) \qquad \widetilde{\mathbf{M}}_3 = \mathbb{E}[(\mathbf{W}^\top \mathbf{x})^{\otimes 3}] - \sum_{j=1}^p \sum_{\mathrm{cyc}} (\mathbf{W}^\top \mathbf{M}_1) \otimes (\mathbf{W}^\top \mathbf{e}_j) \otimes (\mathbf{W}^\top \mathbf{e}_j),$$

where we used the cyclic sum notation

$$\sum_{\mathrm{cyc}} \mathbf{a} \otimes \mathbf{b} \otimes \mathbf{c} := \mathbf{a} \otimes \mathbf{b} \otimes \mathbf{c} + \mathbf{b} \otimes \mathbf{c} \otimes \mathbf{a} + \mathbf{c} \otimes \mathbf{a} \otimes \mathbf{b}.$$

Note that $\mathbf{W}^\top \mathbf{e}_j \in \mathbb{R}^K$ is simply the jth row of $\mathbf{W}$. To obtain an estimate of $\widetilde{\mathbf{M}}_3$, we replace the expectation $\mathbb{E}$ by the empirical average, and substitute $\mathbf{W}$ and $\mathbf{M}_1$ by their plug-in estimates. It is worth mentioning that, because $\widetilde{\mathbf{M}}_3$ has a smaller size than $\mathbf{M}_3$, computations involving $\widetilde{\mathbf{M}}_3$ can be implemented more efficiently.

Once we obtain an estimate of $\widetilde{\mathbf{M}}_3$, which we denote by $\overline{\mathbf{M}}_3$, to recover $\{\boldsymbol{\mu}_k\}_{k=1}^K, \{\omega_k\}_{k=1}^K$ and $\{\sigma_k^2\}_{k=1}^K$, the only task left is computing the orthogonal tensor decomposition (4.3) for $\overline{\mathbf{M}}_3$. The tensor power method in Anandkumar et al. (2014) is shown to solve this problem with provable computational guarantees. We omit the details of the algorithm here. Interested readers are referred to Section 5 of Anandkumar et al. (2014) for the introduction and analysis of this algorithm.

To conclude this subsection, we summarize the entire procedure of estimating $\{\boldsymbol{\mu}_k, \sigma_k, \omega_k\}_{k=1}^K$ as below.

Step 1. Calculate the sample covariance matrix $\widehat{\boldsymbol{\Sigma}} := n^{-1} \sum_{i=1}^n (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^\top$, its minimum eigenvalue $\widehat{\sigma}_{\mathrm{ave}}^2$ and its associated eigenvector $\widehat{\mathbf{v}}$.

Step 2. Derive the estimators $\widehat{\mathbf{M}}_1, \widehat{\mathbf{M}}_2, \widehat{\mathbf{M}}_3$ based on Theorem 4.1 by plug-in of empirical moments of $\mathbf{x}$, $\widehat{\mathbf{v}}$ and $\widehat{\sigma}_{\mathrm{ave}}^2$.

Step 3. Calculate the spectral decomposition $\widehat{\mathbf{M}}_2 = \widehat{\mathbf{U}}\widehat{\mathbf{D}}\widehat{\mathbf{U}}^\top$. Let $\widehat{\mathbf{W}} = \widehat{\mathbf{U}}\widehat{\mathbf{D}}^{-1/2}$. Construct an estimator of $\widetilde{\mathbf{M}}_3$, denoted by $\overline{\mathbf{M}}_3$, based on (4.4) by plug-in of empirical moments of $\widehat{\mathbf{W}}^\top \mathbf{x}$, $\widehat{\mathbf{W}}$ and $\widehat{\mathbf{M}}_1$. Apply the robust tensor power method in Anandkumar et al. (2014) to $\overline{\mathbf{M}}_3$ and obtain $\{\overline{\boldsymbol{\mu}}_k\}_{k=1}^K$ and $\{\widehat{\omega}_k\}_{k=1}^K$.

Step 4. Set $\widehat{\mathbf{W}}^\dagger = \widehat{\mathbf{D}}^{1/2}\widehat{\mathbf{U}}^\top$ and $\widehat{\boldsymbol{\mu}}_k = (\widehat{\mathbf{W}}^\dagger)^\top \overline{\boldsymbol{\mu}}_k / \sqrt{\widehat{\omega}_k}$. Solve the linear equation $\widehat{\mathbf{M}}_1 = \sum_{k=1}^K \widehat{\omega}_k \widehat{\sigma}_k^2 \widehat{\boldsymbol{\mu}}_k$ for $\{\widehat{\sigma}_k^2\}_{k=1}^K$.
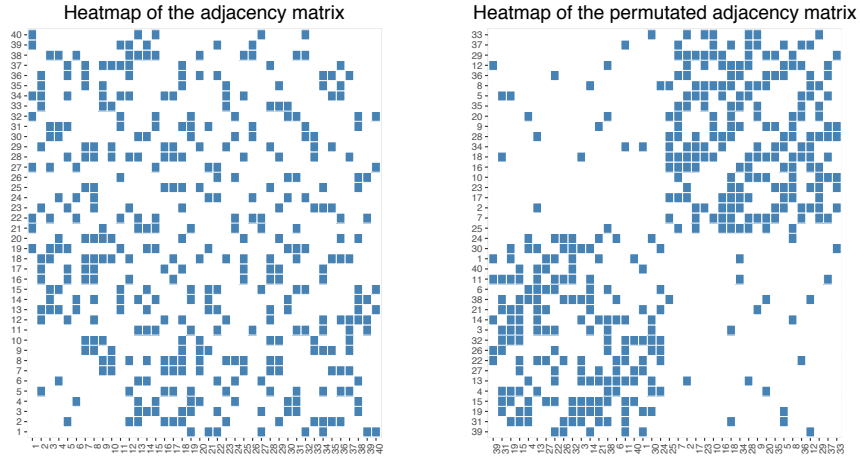
FIG 7. *In both heatmaps, a dark pixel represents an entry with value* 1 *in a matrix, and a white pixel represents an entry with value* 0. *The **left heatmap** shows the (observed) adjacency matrix* $\mathbf{A}$ *of size* $n = 40$ *generated from the SBM with two equal-sized blocks* $(K = 2)$, *with edge probabilities* $5 \log n/n$ *(within blocks) and* $\log n/(4n)$ *(between blocks). The **right heatmap** shows the same matrix with its row indices and column indices suitably permutated based on unobserved* z_i. *Clearly, we observe an approximate rank-*2 *structure in the right heatmap. This motivates estimating* z_i *via the second eigenvector.*

4.2 Community detection

In statistical modeling of networks, the stochastic block model (SBM), first proposed by Holland et al. (1983), has gained much attention in recent years (see Abbe, 2017 for a recent survey). Suppose our observation is a graph of n vertices, each of which belongs to one of K communities (or blocks). Let the vertices be indexed by $[n]$, and the community that vertex i belongs to is indicated by an unknown $\theta_i \in [K]$. In SBM, the probability of an edge between two vertices depends entirely on the membership of the communities. To be specific, let $\mathbf{W} \in \mathbb{R}^{K \times K}$ be a symmetric matrix where each entry takes value in $[0, 1]$, and let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be the adjacency matrix, i.e., $A_{ij} = 1$ if there is an edge between vertex i and j, and $A_{ij} = 0$ otherwise. Then, the SBM assumes

$$\mathbb{P}(A_{ij} = 1) = W_{k\ell} \quad \text{with} \quad \theta_i = k, \theta_j = \ell$$

and $\{A_{ij}\}_{i>j}$ are independent. Here, for ease of presentation, we allow self-connecting edges. Figure 7 gives one realization of the network with two communities.

Though seemingly different, this problem shares a close connection with PCA and spectral methods. Let $\mathbf{z}_i = \mathbf{e}_k$ (namely, the kth canonical basis in $\mathbb{R}^K$) if $\theta_i = k$, indicating the membership of ith node, and define $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n]^\top \in \mathbb{R}^{n \times K}$. The expectation of $\mathbf{A}$ has a low-rank decomposition $\mathbb{E}\mathbf{A} = \mathbf{ZWZ}^\top$ and

(4.5) $$\mathbf{A} = \mathbf{ZWZ}^\top + (\mathbf{A} - \mathbb{E}\mathbf{A}).$$

Loosely speaking, the matrix $\mathbf{Z}$ plays a similar role as factors or loading matrices (unnormalized), and $\mathbf{A} - \mathbb{E}\mathbf{A}$ is similar to the noise (idiosyncratic component). In the ideal situation, the adjacency matrix $\mathbf{A}$ and its expectation are close, and naturally we expect the eigenvectors of $\mathbf{A}$ to be useful for estimating θ_i. Indeed, this

observation is the underpinning of many methods (Rohe et al., 2011; Gao et al., 2015; Abbe and Sandon, 2015). The vanilla spectral method for network/graph data is as follows:

Step 1. Construct the adjacency matrix $\mathbf{A}$ or other similarity-based matrices;

Step 2. Compute eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_L$ corresponding to the largest eigenvalues, and form a matrix $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_\ell] \in \mathbb{R}^{n \times L}$;

Step 3. Run a clustering algorithm on the row vectors of $\mathbf{V}$.

There are many variants and improvements of this vanilla spectral method. For example, in Step 1, very often the graph Laplacian $\mathbf{D} - \mathbf{A}$ or normalized Laplacian $\mathbf{D}^{-1/2}(\mathbf{D} - \mathbf{A})\mathbf{D}^{-1/2}$ is used in place of the adjacency matrix, where $\mathbf{D} = \mathrm{diag}(d_1, \ldots, d_n)$, and $d_i = \sum_j A_{ij}$ is the degree of vertex i. If real-valued similarities or distances between vertices are available, weighted graphs are usually constructed. In Step 2, there are many other refinements over raw eigenvectors in the construction of $\mathbf{V}$, for example, projecting row vectors of $\mathbf{V}$ onto the unit sphere (Ng et al., 2002), and calculating scores based on eigenvector ratios (Jin, 2015), etc. In Step 3, a very popular algorithm for clustering is the K-means algorithm.

We will look at the vanilla spectral algorithm in its simplest form. Our goal is *exact recovery*, which means finding an estimator $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^\top$ such that as $n \to \infty$,

$$\mathbb{P}(\text{there exists a permutation } \pi \text{ of } [K] \text{ s.t. } \widehat{\theta}_i = \pi(\theta_i), \forall i \in [n]) = 1 - o(1).$$

Note that we can only determine $\boldsymbol{\theta}$ up to a permutation since the distribution of our observation is invariant to permutations of $[K]$. There are nice theoretical results, including information limits for exact recovery in Abbe et al. (2016).

Despite its simplicity, spectral methods can be quite sharp for exact recovery in SBM, which succeed in a regime that matches the information limit. The next theorem from Abbe et al. (2017) will make this point clear. Consider the SBM with two balanced blocks, i.e., $K = 2$ and $\{i : \theta_i = 1\} = \{i : \theta_i = 2\} = n/2$, and suppose $W_{11} = W_{22} = a \log n/n$, $W_{12} = b \log n/n$ where $a > b > 0$. In this case, one can easily see that the second eigenvector of $\mathbb{E}\mathbf{A}$ is given by $\mathbf{v}_2^*$ whose ith entry is given by $1/\sqrt{n}$ if $\theta_i = 1$ and -1 otherwise. In other words, $\mathrm{sgn}(\mathbf{v}_2^*)$ classifies the two communities, where $\mathrm{sgn}(\cdot)$ is the sign function applied to each entry of a vector. This is shown in Figure 2 for the case that $\#\{i : \theta_i = 1\} = 2500$ (red curve, left panel), where the second eigenvector $\mathbf{v}_2$ of $\mathbf{A}$ is also depicted (blue curve). The entrywise closeness between these two quantities is guaranteed by the perturbation theory under ℓ_∞-norm (Abbe et al., 2017).

THEOREM 4.3. *Let $\mathbf{v}_2$ be the normalized second eigenvector of $\mathbf{A}$. If $\sqrt{a} - \sqrt{b} < \sqrt{2}$, then no estimator achieves exact recovery; if $\sqrt{a} - \sqrt{b} > \sqrt{2}$, then both the maximum likelihood estimator and the eigenvector estimator $sgn(\mathbf{v}_2)$ achieves exact recovery.*

The proof of this result is based on entry-wise analysis of eigenvectors in a spirit similar to Theorem 2.4, together with a probability tail bound for differences of binomial variables.

4.3 Matrix completion

In recommendation systems, an important problem is to estimate users' preferences based on history data. Usually, the available data per user is very small compared with the total number of items (each user sees only a small number of movies and buys only a small fraction of books, comparing to the total). Matrix completion is one formulation of such problem.

The goal of (noisy) matrix completion is to estimate a low-rank matrix $\mathbf{M}^* \in \mathbb{R}^{n_1 \times n_2}$ from noisy observations of some entries (n_1 users and n_2 items). Suppose we know $\mathrm{rank}(\mathbf{M}^*) = K$. For each $i \in [n_1]$ and $j \in [n_2]$, let I_{ij} be i.i.d. Bernoulli variable with $\mathbb{P}(I_{ij} = 1) = p$ that indicates if we have observed information about the entry M_{ij}^*, i.e., $I_{ij} = 1$ if and only if it is observed. Also suppose that our observation is $M_{ij} = M_{ij}^* + \varepsilon_{ij}$ if $I_{ij} = 1$, where ε_{ij} is i.i.d. $\mathcal{N}(0, \sigma^2)$ jointly independent of I_{ij}.

One natural way to estimate $\mathbf{M}^*$ is to solve

$$\min_{\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}} \frac{1}{2} \left\| \mathcal{P}_\Omega(\mathbf{M}) - \mathcal{P}_\Omega(\mathbf{X}) \right\|^2 \quad \text{subject to} \quad \mathrm{rank}(\mathbf{X}) = K,$$

where $\mathcal{P}_\Omega : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{n_1 \times n_2}$ is the sampling operator defined by $[\mathcal{P}_\Omega(\mathbf{X})]_{ij} = I_{ij} X_{ij}, \ \forall i, j$. The minimizer of this problem is essentially the MLE for M^*. Due to the nonconvex constraint $\mathrm{rank}(\mathbf{X}) = K$, it is desirable to relax this optimization into a convex program. A popular way to achieve that is to transform the rank constraint into a penalty term $\lambda \|\mathbf{X}\|_*$ that is added to the quadratic objective function, where λ is a tuning parameter and $\|\cdot\|_*$ is the nuclear norm (that is, the ℓ_1 norm of the vector of all its singular values), which encourages a solution with low rank (number of nonzero components in that vector). A rather surprising conclusion from Candès and Recht (2009) is that in the noiseless setting, solving the relaxed problem yields the same solution as the nonconvex problem with high probability.

We can view this problem from the perspective of factor models. The assumption that $\mathbf{M}^*$ has low rank can be justified by interpreting each M_{ij} as the linear combination of a few latent factors. Indeed, if M_{ij} is the preference score of user i for item j, then it is reasonable to posit $M_{ij} = \mathbf{b}_i^\top \mathbf{f}_j$, where $\mathbf{f}_j \in \mathbb{R}^K$ is the features item j possesses and $\mathbf{b}_i \in \mathbb{R}^K$ is the tendency of user i towards the features. In this regard, $\mathbf{M}^* = \mathbf{B}\mathbf{F}^\top$ can be viewed as the part explained by the factors in the factor models.

This discussion motivates us to write our observation as

$$\mathcal{P}_\Omega(\mathbf{M}) = p\mathbf{M}^* + (\mathcal{P}_\Omega(\mathbf{M}^*) - \mathbb{E}\mathcal{P}_\Omega(\mathbf{M}^*) + \mathcal{P}_\Omega(\mathbf{E})), \quad \text{where } \mathbf{E} := (\varepsilon_{ij})_{i,j} \in \mathbb{R}^{n_1 \times n_2},$$

since $\mathbb{E}\mathcal{P}_\Omega(\mathbf{M}) = p\mathbf{M}^*$. This decomposition gives the familiar "low-rank plus noise" structure. It is natural to conduct PCA on $\mathcal{P}_\Omega(\mathbf{M})$ to extract the low-rank part.

Let the best rank-K approximation of $\mathcal{P}_\Omega(\mathbf{M})$ be given by $\mathbf{U}\mathrm{diag}(\sigma_1, \ldots, \sigma_K)\mathbf{V}^\top$, where $\{\sigma_k\}_{k=1}^K$ are the largest K singular values in descending order, and columns of $\mathbf{U} \in \mathbb{R}^{n_1 \times K}, \mathbf{V} \in \mathbb{R}^{n_2 \times K}$ correspond to their normalized left and right singular vectors, respectively. Similarly, we have singular value decomposition $\mathbf{M}^* = \mathbf{U}^*\mathrm{diag}(\sigma_1^*, \ldots, \sigma_K^*)(\mathbf{V}^*)^\top$. The following result from Abbe et al. (2017) provides entry-wise bounds for our estimates. For

a matrix, denote by $\|\cdot\|_{\max}$ the largest absolute value of all entries, and $\|\cdot\|_{2\to\infty}$ the largest ℓ_2 norm of all row vectors.

THEOREM 4.4. *Let $n = n_1 + n_2$, $\eta = \max\{\|\mathbf{U}\|_{2\to\infty}, \|\mathbf{V}\|_{2\to\infty}\}$ and $\kappa = \sigma_1^*/\sigma_K^*$. There exist constants $C, C' > 0$ and an orthogonal matrix $\mathbf{R} \in \mathbb{R}^{K\times K}$ such that the following holds. If $p \geq 6\log n/n$ and $\kappa\frac{n(\|\mathbf{M}^*\|_{\max}+\sigma)}{\sigma_r^*}\sqrt{\frac{\log n}{np}} \leq 1/C$, then with at least probability $1 - C/n$,*

$$\max\{\|\mathbf{U}\mathbf{R} - \mathbf{U}^*\|_{\max}, \|\mathbf{V}\mathbf{R} - \mathbf{V}^*\|_{\max}\} \leq C'\eta\kappa\frac{n(\|\mathbf{M}^*\|_{\max} + \sigma)}{\sigma_r^*}\sqrt{\frac{\log n}{np}},$$

$$\|\mathbf{U}\mathrm{diag}\{\sigma_1, \ldots, \sigma_K\}\mathbf{V}^\top - \mathbf{M}^*\|_{\max} \leq C'\eta^2\kappa^4(\|\mathbf{M}^*\|_{\max} + \sigma)\sqrt{\frac{n\log n}{p}}.$$

We can simplify the bounds with a few additional assumptions. If $n_1 \asymp n_2$, then η is of order $O(\sqrt{K/n})$ assuming a bounded coherence number. In addition, if κ is also bounded, then

$$\|\mathbf{U}\mathrm{diag}\{\sigma_1, \ldots, \sigma_K\}\mathbf{V}^\top - \mathbf{M}^*\|_{\max} \lesssim (\|\mathbf{M}^*\|_{\max} + \sigma)\sqrt{\frac{\log n}{np}}.$$

We remark that the requirement on the sample ratio $p \gtrsim \log n/n$ is the weakest condition necessary for matrix completion, which ensures each row and column and sampled with high probability. Also, the entry-wise bound above can recover the Frobenius bound (Keshavan et al., 2010) up to a log factor. It is more precise than the Frobenius bound, because the latter only provides control on average error.

4.4 Synchronization problems

Synchronization problems are a class of problems in which one estimates signals from their pairwise comparisons. Consider the phase synchronization problem as an example, that is, estimating n angles $\theta_1, \ldots, \theta_n$ from noisy measurements of their differences. We can express an angle θ_ℓ in the equivalent form of a unit-modulus complex number $z_\ell = \exp(i\theta_\ell)$, and thus, the task is to estimate a complex vector $\mathbf{z} = (z_1, \ldots, z_n)^\top \in \mathbb{C}^n$. Suppose our measurements have the form $C_{\ell k} = \bar{z}_\ell z_k + \sigma w_{\ell k}$, where $\bar{z}_\ell$ denotes the conjugate of z_ℓ, and for all $\ell > k$, $w_{\ell k} \in \mathbb{C}$ is i.i.d. complex Gaussian variable (namely, the real part and imaginary part of $w_{\ell k}$ are $\mathcal{N}(0, 1/2)$ and independent). Then, the phase of $C_{\ell k}$ (namely $\arg(C_{\ell k})$) encodes the noisy difference $\theta_k - \theta_\ell$.

More generally, the goal of a synchronization problem is to estimate n signals from their pairwise measurements, where each signal is an element from a group, e.g., the group of rotations in three dimensions. Synchronization problems are motivated from imaging problems such as cryo-EM (Shkolnisky and Singer, 2012), camera calibration (Tron and Vidal, 2009), etc.

Synchronization problems also admit the "low-rank plus noise" structure. Consider our phase synchronization problem again. If we let $w_{k\ell} = w_{\ell k}$ $(\ell > k)$ and $w_{\ell\ell} = 0$, and write $\mathbf{W} = (w_{\ell k})_{\ell,k=1}^n$, then our measurement matrix $\mathbf{C} = (C_{\ell k})_{\ell,k=1}^n$ has the structure

$$\mathbf{C} = \mathbf{z}\mathbf{z}^* + \sigma\mathbf{W},$$

where $*$ denotes the conjugate transpose. This decomposition has a similar form to (4.5) in community detection. Note that $\mathbf{z}\mathbf{z}^*$ is a complex matrix with a single nonzero eigenvalue n, and $\|\sigma\mathbf{W}\|_2$ is of order $\sigma\sqrt{n}$ with high probability (which is a basic result in random matrix theory). Therefore, we expect that no estimators can do well if $\sigma \gtrsim \sqrt{n}$. Indeed, the information-theoretic limit is established in Lelarge and Miolane (2016). Our next result from Zhong and Boumal (2018) gives estimation guarantees if the reverse inequality is true (up to a log factor).

THEOREM 4.5. *Let* $\mathbf{v} \in \mathbb{C}^n$ *be the leading eigenvector of* $\mathbf{C}$ *such that* $\|\mathbf{v}\|_2 = \sqrt{n}$ *and* $\mathbf{v}^*\mathbf{z} = |\mathbf{v}^*\mathbf{z}|$. *Then, if* $\sigma \lesssim \sqrt{n/\log n}$, *then with probability* $1 - O(n^{-2})$, *the relative errors satisfy*

$$n^{-1/2}\|\mathbf{v} - \mathbf{z}\|_2 \lesssim \sigma/\sqrt{n}, \quad and \quad \|\mathbf{v} - \mathbf{z}\|_\infty \lesssim \sigma\sqrt{\log n/n}.$$

Moreover, the above two inequalities also hold for the maximum likelihood estimator.

Note that the eigenvector of a complex matrix is not unique: for any $\alpha \in \mathbb{R}$, the vector $e^{i\alpha}\mathbf{v}$ is also an eigenvector, so we fix the global phase $e^{i\alpha}$ by restricting $\mathbf{v}^*\mathbf{z} = |\mathbf{v}^*\mathbf{z}|$. Note also that the maximum likelihood estimator is different from $\mathbf{v}$, because the MLE must satisfy the entry-wise constraint $|z_\ell| = 1$ for any $\ell \in [n]$. This result implies consistency of $\mathbf{v}$ in terms of both the ℓ_2 norm and the ℓ_∞ norm if $\sigma \ll \sqrt{n/\log n}$, and thus, provides good evidence that spectral methods (or PCA) are simple, generic, yet powerful.

APPENDIX A: PROOFS

PROOF OF COROLLARY 2.1. Notice that the result is trivial if $\delta_0 \leq 2\|\widetilde{\mathbf{A}} - \mathbf{A}\|_2$, since $\|(\widetilde{\mathbf{A}} - \mathbf{A})\mathbf{V}\|_2 \leq \|\widetilde{\mathbf{A}} - \mathbf{A}\|_2$ and $\|\sin\mathbf{\Theta}(\widetilde{\mathbf{A}}, \mathbf{A})\|_2 \leq 1$ always hold. If $\delta_0 > 2\|\widetilde{\mathbf{A}} - \mathbf{A}\|_2$, then by Weyl's inequality,

$$\mathcal{L}(\widetilde{\mathbf{V}}^\perp) \subset (-\infty, \alpha - \delta_0 + \|\widetilde{\mathbf{A}} - \mathbf{A}\|_2] \cup [\beta + \delta_0 - \|\widetilde{\mathbf{A}} - \mathbf{A}\|_2, +\infty).$$

Thus, we can set $\delta = \delta_0 - \|\widetilde{\mathbf{A}} - \mathbf{A}\|_2$ in Theorem 2.3 and derive

$$\|\sin\mathbf{\Theta}(\widetilde{\mathbf{V}}, \mathbf{V})\|_2 \leq \frac{\|\widetilde{\mathbf{A}} - \mathbf{A}\|_2}{\delta_0 - \|\widetilde{\mathbf{A}} - \mathbf{A}\|_2} \leq \frac{\|\widetilde{\mathbf{A}} - \mathbf{A}\|_2}{\delta_0 - \delta_0/2} = 2\delta_0^{-1}\|\widetilde{\mathbf{A}} - \mathbf{A}\|_2.$$

This proves the spectral norm case. $\square$

PROOF OF THEOREM 2.4. **Step 1:** First, we derive a few elementary inequalities: for any $m \in [n]$,

(A.1) $\quad \|\mathbf{W}^{(m)}\|_2 \leq \|\mathbf{W}\|_2, \quad \|\mathbf{w}_m\|_2 \leq \|\mathbf{W}\|_2, \quad \|\mathbf{W} - \mathbf{W}^{(m)}\|_2 \leq 2\|\mathbf{W}\|_2.$

To prove these inequalities, recall the (equivalent) definition of spectral norm for symmetric matrices:

$$\|\mathbf{W}\|_2 = \max_{\boldsymbol{x}, \boldsymbol{y} \in S^{n-1}} \boldsymbol{x}^\top \mathbf{W} \boldsymbol{y} = \max_{\boldsymbol{x} \in S^{n-1}} \|\mathbf{W}\boldsymbol{x}\|_2,$$

where S^{n-1} is the unit sphere in $\mathbb{R}^n$, and $\boldsymbol{x} = (x_1, \ldots, x_n)^\top, \boldsymbol{y} = (y_1, \ldots, y_n)^\top$. The first and second inequalities follow from

$$\|\mathbf{W}\|_2 \geq \max\{\boldsymbol{x}^\top \mathbf{W} \boldsymbol{y} : \boldsymbol{x}, \boldsymbol{y} \in S^{n-1}, x_m = y_m = 0\} = \|\mathbf{W}^{(m)}\|_2, \quad \text{and}$$

$$\|\mathbf{W}\|_2 \geq \max_{\boldsymbol{x} \in S^{n-1}} |\langle \mathbf{w}_m, \boldsymbol{x} \rangle| = \|\mathbf{w}_m\|_2.$$

The third inequality follows from the first one and the triangle inequality.

Step 2: Next, by the definition of eigenvectors,

$$(A.2) \qquad \widetilde{\mathbf{v}}_\ell - \mathbf{v}_\ell = \frac{\widetilde{\mathbf{A}} \widetilde{\mathbf{v}}_\ell}{\widetilde{\lambda}_\ell} - \mathbf{v}_\ell = \left(\frac{\mathbf{A} \widetilde{\mathbf{v}}_\ell}{\widetilde{\lambda}_\ell} - \mathbf{v}_\ell \right) + \frac{\mathbf{W} \widetilde{\mathbf{v}}_\ell}{\widetilde{\lambda}_\ell}.$$

We first control the entries of the first term on the right-hand side. Using the decomposition (2.8), we have

$$(A.3)$$
$$\left[\frac{\mathbf{A} \widetilde{\mathbf{v}}_\ell}{\widetilde{\lambda}_\ell} - \mathbf{v}_\ell \right]_m = \left(\frac{\lambda_\ell}{\widetilde{\lambda}_\ell} \langle \mathbf{v}_\ell, \widetilde{\mathbf{v}}_\ell \rangle - 1 \right) [\mathbf{v}_\ell]_m + \sum_{k \neq \ell, k \leq K} \frac{\lambda_k}{\widetilde{\lambda}_\ell} \langle \mathbf{v}_k, \widetilde{\mathbf{v}}_\ell \rangle [\mathbf{v}_k]_m, \quad \forall m \in [n].$$

Using the triangle inequality, we have

$$\left| \frac{\lambda_\ell}{\widetilde{\lambda}_\ell} \langle \mathbf{v}_\ell, \widetilde{\mathbf{v}}_\ell \rangle - 1 \right| \leq \left| \frac{\lambda_\ell}{\widetilde{\lambda}_\ell} \langle \mathbf{v}_\ell, \widetilde{\mathbf{v}}_\ell \rangle - \langle \mathbf{v}_\ell, \widetilde{\mathbf{v}}_\ell \rangle \right| + |\langle \mathbf{v}_\ell, \widetilde{\mathbf{v}}_\ell \rangle - 1| \leq \frac{|\lambda_\ell - \widetilde{\lambda}_\ell|}{|\widetilde{\lambda}_\ell|} + \frac{1}{2} \|\widetilde{\mathbf{v}}_\ell - \mathbf{v}_\ell\|^2.$$

By Weyl's inequality, $|\widetilde{\lambda}_\ell - \lambda_\ell| \leq \|\mathbf{W}\|_2$, and thus $|\widetilde{\lambda}_\ell| \geq |\lambda_\ell| - \|\mathbf{W}\|_2 \geq \delta_\ell - \|\mathbf{W}\|_2$. Also, by Corollary 2.1 (simplified Davis-Kahan's theorem) and its following remark, $\|\widetilde{\mathbf{v}}_\ell - \mathbf{v}_\ell\|_2 \leq 2\sqrt{2} \|\mathbf{W}\|_2 / \delta_\ell$. Therefore, under the condition $\delta_\ell \geq 2\|\mathbf{W}\|_2$,

$$\left| \frac{\lambda_\ell}{\widetilde{\lambda}_\ell} \langle \mathbf{v}_\ell, \widetilde{\mathbf{v}}_\ell \rangle - 1 \right| \leq \frac{\|\mathbf{W}\|_2}{\delta_\ell - \|\mathbf{W}\|_2} + \frac{4\|\mathbf{W}\|_2^2}{\delta_\ell^2} \leq \frac{2\|\mathbf{W}\|_2}{\delta_\ell} + \frac{2\|\mathbf{W}\|_2}{\delta_\ell} = \frac{4\|\mathbf{W}\|_2}{\delta_\ell}.$$

Using Corollary 2.1 again, we obtain

$$\sum_{k \neq \ell, k \leq K} \frac{\lambda_k^2}{\widetilde{\lambda}_\ell^2} \langle \mathbf{v}_k, \widetilde{\mathbf{v}}_\ell \rangle^2 \lesssim \sum_{k \neq \ell, k \leq K} \langle \mathbf{v}_k, \widetilde{\mathbf{v}}_\ell \rangle^2 \leq 1 - \langle \mathbf{v}_\ell, \widetilde{\mathbf{v}}_\ell \rangle^2 = \sin^2 \theta(\mathbf{v}_\ell, \widetilde{\mathbf{v}}_\ell) \leq \frac{4\|\mathbf{W}\|_2^2}{\delta_\ell^2},$$

where the first inequality is due to $|\widetilde{\lambda}_\ell| \geq |\lambda_\ell| - \|\mathbf{W}\|_2 \geq 4|\lambda_\ell|/5$ and the condition $|\lambda_\ell| \asymp \max_{k \in [K]} |\lambda_k|$, and the second inequality is due to the fact that $\{\mathbf{v}_k\}_{k=1}^K$ is a subset of orthonormal basis. Now we use the Cauchy-Schwarz inequality to bound the second term on the right-hand side of (A.3) and get

$$(A.4) \qquad \left| \left[\frac{\mathbf{A} \widetilde{\mathbf{v}}_\ell}{\widetilde{\lambda}_\ell} - \mathbf{v}_\ell \right]_m \right| \lesssim \frac{\|\mathbf{W}\|_2}{\delta_\ell} \left(\sum_{k=1}^K [\mathbf{v}_k]_m^2 \right)^{1/2}.$$

Step 3: To bound the entries of the second term in (A.2), we use the leave-one-out idea as follows.
$$(A.5)$$
$$[\mathbf{W} \widetilde{\mathbf{v}}_\ell]_m = [\mathbf{W} \widetilde{\mathbf{v}}_\ell^{(m)}]_m + [\mathbf{W}(\widetilde{\mathbf{v}}_\ell - \widetilde{\mathbf{v}}_\ell^{(m)})]_m = \langle \mathbf{w}_m, \widetilde{\mathbf{v}}_\ell^{(m)} \rangle + \langle \mathbf{w}_m, \widetilde{\mathbf{v}}_\ell - \widetilde{\mathbf{v}}_\ell^{(m)} \rangle, \quad \forall m \in [n].$$

We can bound the second term using the Cauchy-Schwarz inequality: $|\langle \mathbf{w}_m, \widetilde{\mathbf{v}}_\ell - \widetilde{\mathbf{v}}_\ell^{(m)} \rangle| \leq \|\mathbf{w}_m\|_2 \|\widetilde{\mathbf{v}}_\ell - \widetilde{\mathbf{v}}_\ell^{(m)}\|_2$. The crucial observation is that, if we view $\widetilde{\mathbf{v}}_\ell$ as the

perturbed version of $\widetilde{\mathbf{v}}_\ell^{(m)}$, then by Theorem 2.3 (Davis-Kahan's theorem) and Weyl's inequality, for any $\ell \in [K]$,

$$\|\widetilde{\mathbf{v}}_\ell - \widetilde{\mathbf{v}}_\ell^{(m)}\|_2 \leq \frac{\sqrt{2}\|\mathbf{\Delta}^{(m)}\widetilde{\mathbf{v}}_\ell^{(m)}\|_2}{\widetilde{\delta}_\ell^{(m)} - \|\mathbf{\Delta}^{(m)}\|_2}, \quad \text{where } \mathbf{\Delta}^{(m)} := \mathbf{W} - \mathbf{W}^{(m)}.$$

Here, $\widetilde{\delta}_\ell^{(m)}$ is the eigen-gap of $\mathbf{A} + \mathbf{W}^{(m)}$, and it satisfies $\widetilde{\delta}_\ell^{(m)} \geq \delta_\ell - 2\|\mathbf{W}^{(m)}\|_2$ since $|\widetilde{\lambda}_i^{(m)} - \lambda_i| \leq \|\mathbf{W}^{(m)}\|_2$ for all $i \in [n]$, by Weyl's inequality. By (A.1), we have $\widetilde{\delta}_\ell^{(m)} - \|\mathbf{\Delta}^{(m)}\|_2 \geq \delta_\ell - 4\|\mathbf{W}\|_2$. Thus, under the condition $\delta_\ell \geq 5\|\mathbf{W}\|_2$, we have

$$\|\widetilde{\mathbf{v}}_\ell - \widetilde{\mathbf{v}}_\ell^{(m)}\|_2 \lesssim \frac{\|\mathbf{\Delta}^{(m)}\widetilde{\mathbf{v}}_\ell^{(m)}\|_2}{\delta_\ell}.$$

Note that the mth entry of the vector $\mathbf{\Delta}^{(m)}\widetilde{\mathbf{v}}_\ell^{(m)}$ is exactly $\langle \mathbf{w}_m, \widetilde{\mathbf{v}}_\ell^{(m)}\rangle$, and other entries are $W_{im}[\widetilde{\mathbf{v}}_\ell^{(m)}]_m$ where $i \neq m$. Thus,

$$\|\widetilde{\mathbf{v}}_\ell - \widetilde{\mathbf{v}}_\ell^{(m)}\|_2 \lesssim \frac{1}{\delta_\ell}\left(\langle \mathbf{w}_m, \widetilde{\mathbf{v}}_\ell^{(m)}\rangle^2 + \sum_{i \neq m} W_{im}^2[\widetilde{\mathbf{v}}_\ell^{(m)}]_m^2\right)^{1/2} \leq \frac{1}{\delta_\ell}\left(|\langle \mathbf{w}_m, \widetilde{\mathbf{v}}_\ell^{(m)}\rangle| + \|\mathbf{w}_m\|_2|[\widetilde{\mathbf{v}}_\ell^{(m)}]_m|\right),$$

where we used $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ $(a, b \geq 0)$. The above inequality, together with $|\langle \mathbf{w}_m, \widetilde{\mathbf{v}}_\ell - \widetilde{\mathbf{v}}_\ell^{(m)}\rangle| \leq \|\mathbf{w}_m\|_2\|\widetilde{\mathbf{v}}_\ell - \widetilde{\mathbf{v}}_\ell^{(m)}\|_2$, leads to a bound on $[\mathbf{W}\widetilde{\mathbf{v}}_\ell]_m$ in (A.5).

$$|[\mathbf{W}\widetilde{\mathbf{v}}_\ell]_m| \lesssim |\langle \mathbf{w}_m, \widetilde{\mathbf{v}}_\ell^{(m)}\rangle| + \frac{\|\mathbf{w}_m\|_2}{\delta_\ell}\left(|\langle \mathbf{w}_m, \widetilde{\mathbf{v}}_\ell^{(m)}\rangle| + \|\mathbf{w}_m\|_2|[\widetilde{\mathbf{v}}_\ell^{(m)}]_m|\right)$$

(A.6) $$\lesssim |\langle \mathbf{w}_m, \widetilde{\mathbf{v}}_\ell^{(m)}\rangle| + \|\mathbf{w}_m\|_2|[\widetilde{\mathbf{v}}_\ell^{(m)}]_m|$$

where we used $\delta_\ell^{-1}\|\mathbf{w}_m\|_2 \leq \delta_\ell^{-1}\|\mathbf{W}\|_2 < 1$. We claim that $|[\widetilde{\mathbf{v}}_\ell^{(m)}]_m| \lesssim (\sum_{k=1}^K [\mathbf{v}_k]_m^2)^{1/2}$. Once this is proved, combining it with (A.4) and (A.6) yields the desired bound on the entries of $\widetilde{\mathbf{v}}_\ell - \mathbf{v}_\ell$ in (A.2):

$$|[\widetilde{\mathbf{v}}_\ell - \mathbf{v}_\ell]_m| \lesssim \frac{\|\mathbf{W}\|_2}{\delta_\ell}\left(\sum_{k=1}^K [\mathbf{v}_k]_m^2\right)^{1/2} + \frac{1}{\delta_\ell}\left(|\langle \mathbf{w}_m, \widetilde{\mathbf{v}}_\ell^{(m)}\rangle| + \|\mathbf{w}_m\|_2|[\widetilde{\mathbf{v}}_\ell^{(m)}]_m|\right)$$

$$\lesssim \frac{\|\mathbf{W}\|_2}{\delta_\ell}\left(\sum_{k=1}^K [\mathbf{v}_k]_m^2\right)^{1/2} + \frac{|\langle \mathbf{w}_m, \widetilde{\mathbf{v}}_\ell^{(m)}\rangle|}{\delta_\ell},$$

where, in the first inequality, we used $|\widetilde{\lambda}_\ell| \geq |\lambda_\ell| - \|\mathbf{W}\|_2 \geq \delta_\ell - \delta_\ell/5 = 4\delta_\ell/5$, and in the second inequality, we used $\|\mathbf{w}_m\|_2 \leq \|\mathbf{W}\|_2$ and the claim.

Step 4: Finally, we prove our claim that $|[\widetilde{\mathbf{v}}_\ell^{(m)}]_m| \lesssim (\sum_{k=1}^K [\mathbf{v}_k]_m^2)^{1/2}$. By definition, $\widetilde{\lambda}_\ell^{(m)}\widetilde{\mathbf{v}}_\ell^{(m)} = (\mathbf{A} + \mathbf{W}^{(m)})\widetilde{\mathbf{v}}_\ell^{(m)}$. Note that the mth row of $\mathbf{W}^{(m)}\widetilde{\mathbf{v}}_\ell^{(m)}$ is 0, since $\mathbf{W}^{(m)}$ has only zeros in its mth row. Thus,

$$[\widetilde{\mathbf{v}}_\ell^{(m)}]_m = \left([\widetilde{\mathbf{v}}_\ell^{(m)}]_m - [\mathbf{v}_\ell]_m\right) + [\mathbf{v}_\ell]_m = \left[\frac{\mathbf{A}\widetilde{\mathbf{v}}_\ell^{(m)}}{\widetilde{\lambda}_\ell^{(m)}} - \mathbf{v}_\ell\right]_m + [\mathbf{v}_\ell]_m.$$

With an argument similar to the one that leads to (A.4), we can bound the first term on the right-hand side.

$$\left| \left[\frac{\mathbf{A}\widetilde{\mathbf{v}}_\ell^{(m)}}{\widetilde{\lambda}_\ell^{(m)}} - \mathbf{v}_\ell \right]_m \right| \lesssim \frac{\|\mathbf{W}^{(m)}\|_2}{\delta_\ell} \left(\sum_{k=1}^K [\mathbf{v}_k]_m^2 \right)^{1/2} \leq \left(\sum_{k=1}^K [\mathbf{v}_k]_m^2 \right)^{1/2}.$$

Clearly, $|[\mathbf{v}_\ell]_m|$ is also upper bounded by the right-hand side above. This proves our claim and concludes the proof. □

PROOF OF COROLLARY 2.2. Let us construct symmetric matrices $\mathbf{A}, \mathbf{W}, \widetilde{\mathbf{W}}$ of size $n + p$ via a standard *dilation* technique (Paulsen, 2002). Define

$$\mathbf{A} = \begin{pmatrix} \mathbf{0} & \mathbf{L} \\ \mathbf{L}^\top & \mathbf{0} \end{pmatrix}, \quad \mathbf{W} = \begin{pmatrix} \mathbf{0} & \mathbf{E} \\ \mathbf{E}^\top & \mathbf{0} \end{pmatrix}, \quad \text{and} \quad \widetilde{\mathbf{A}} = \mathbf{A} + \mathbf{W}.$$

It can be checked that $\mathrm{rank}(\mathbf{A}) = 2K$, and importantly,

$$(A.7) \quad \mathbf{A} = \frac{1}{2} \sum_{k=1}^K \sigma_k \begin{pmatrix} \mathbf{u}_k \\ \mathbf{v}_k \end{pmatrix} \begin{pmatrix} \mathbf{u}_k^\top & \mathbf{v}_k^\top \end{pmatrix} - \frac{1}{2} \sum_{k=1}^K \sigma_k \begin{pmatrix} \mathbf{u}_k \\ -\mathbf{v}_k \end{pmatrix} \begin{pmatrix} \mathbf{u}_k^\top & -\mathbf{v}_k^\top \end{pmatrix}.$$

Step 1: Check the conditions of Theorem 2.4. The nonzero eigenvalues of $\mathbf{A}$ are $\pm\sigma_k$, ($k \in [K]$), and the corresponding eigenvectors are $(\mathbf{u}_k^\top, \pm\mathbf{v}_k^\top)^\top/\sqrt{2} \in \mathbb{R}^{n+p}$. It is clear that the eigenvalue condition $|\lambda_\ell| \asymp \max_{k\in[K]} |\lambda_k|$ in Theorem 2.4 is satisfied, and the eigen-gap δ_ℓ of $\mathbf{A}$ is exactly γ_ℓ. Since the identity (A.7) holds for any matrix constructed from dilation, by applying it to $\mathbf{W}$ we get $\|\mathbf{W}\|_2 = \|\mathbf{E}\|_2$.

Step 2: Apply the conclusion of Theorem 2.4. Similarly as before, we write $\mathbf{W}^{(m)}$ as the matrix obtained by setting mth row and mth column of $\mathbf{W}$ to zero, where $m \in [n + p]$. We also denote $\widetilde{\mathbf{A}}^{(m)} = \mathbf{A} + \mathbf{W}^{(m)}$. Using a similar argument as Step 1, we find

(1) the eigenvectors of $\widetilde{\mathbf{A}}$ are $\begin{pmatrix} \widetilde{\mathbf{u}}_k \\ \pm\widetilde{\mathbf{v}}_k \end{pmatrix}/\sqrt{2}$,

(2) the eigenvectors of $\widetilde{\mathbf{A}}^{(i)}$ are $\begin{pmatrix} * \\ \pm\widetilde{\mathbf{v}}_k^{(i)} \end{pmatrix}/\sqrt{2}$, $\forall\, i \in [n]$, and

(3) the eigenvectors of $\widetilde{\mathbf{A}}^{(n+j)}$ are $\begin{pmatrix} \widetilde{\mathbf{u}}_k^{(j)} \\ * \end{pmatrix}/\sqrt{2}$, $\forall\, j \in [p]$,

where $*$ means some appropriate vectors we do not need in the proof (we do not bother introducing notations for them). We also observe that

$$\mathbf{w}_m = \begin{cases} \begin{pmatrix} \mathbf{0} & \mathbf{e}_i^{\mathrm{row}} \end{pmatrix}^\top, & m = i \in [n] \\ \begin{pmatrix} (\mathbf{e}_j^{\mathrm{col}})^\top & \mathbf{0} \end{pmatrix}^\top, & m = n + j, \; j \in [p] \end{cases}$$

Note that the inner product between $\mathbf{w}_m$ and the eigenvector of $\widetilde{\mathbf{A}}^{(m)}$ is $\langle(\mathbf{e}_i^{\mathrm{row}})^\top, \pm\widetilde{\mathbf{v}}_k^{(i)}\rangle$ if $m = i \in [n]$, or $\langle\mathbf{e}_j^{\mathrm{col}}, \widetilde{\mathbf{u}}_k^{(j)}\rangle$ if $m = n + j$, $j \in [p]$. Therefore, applying Theorem 2.4 to the first n entries of

$$\frac{1}{\sqrt{2}} \begin{pmatrix} \widetilde{\mathbf{u}}_\ell - \mathbf{u}_\ell \\ \widetilde{\mathbf{v}}_\ell - \mathbf{v}_\ell \end{pmatrix},$$

we obtain the first inequality of Corollary 2.2, and applying Theorem 2.4 to the last p entries leads to the second inequality. □

PROOF OF LEMMA 3.1.

$$\mathbb{E}_{\boldsymbol{\varepsilon}}[\|\mathbf{X}^\top\widehat{\boldsymbol{\beta}}_K - \mathbf{X}^\top\boldsymbol{\beta}^*\|_2^2/n] = \mathbb{E}_{\boldsymbol{\varepsilon}}[\|\mathbf{Q}_K\boldsymbol{\Sigma}_K\mathbf{P}_K^\top\boldsymbol{\beta}^* + \mathbf{Q}_K\mathbf{Q}_K^\top\boldsymbol{\varepsilon} - \mathbf{X}^\top\boldsymbol{\beta}^*\|_2^2/n]$$

$$= \mathbb{E}_{\boldsymbol{\varepsilon}}[\|\mathbf{Q}_K\mathbf{Q}_K^\top\boldsymbol{\varepsilon} - \mathbf{Q}_{K+}\boldsymbol{\Sigma}_{K+}\mathbf{P}_{K+}^\top\boldsymbol{\beta}^*\|_2^2/n]$$

$$= \frac{K\sigma^2}{n} + \underbrace{\boldsymbol{\beta}^{*\top}\mathbf{P}_{K+}}_{\boldsymbol{\alpha}^\top}\boldsymbol{\Sigma}_{K+}^2\underbrace{\mathbf{P}_{K+}^\top\boldsymbol{\beta}^*}_{\boldsymbol{\alpha}}.$$

$$= \frac{K\sigma^2}{n} + \sum_{j=K+1}^{d}\lambda_j^2\alpha_j^2.$$

$\square$

REFERENCES

ABBE, E. (2017). Community detection and stochastic block models: recent developments. *arXiv preprint arXiv:1703.10146* .

ABBE, E., BANDEIRA, A. S. and HALL, G. (2016). Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory* **62** 471–487.

ABBE, E., FAN, J., WANG, K. and ZHONG, Y. (2017). Entrywise eigenvector analysis of random matrices with low expected rank. *arXiv preprint arXiv:1709.09565* .

ABBE, E. and SANDON, C. (2015). Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*. IEEE.

AHN, S. C. and HORENSTEIN, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica* **81** 1203–1227.

ANANDKUMAR, A., GE, R., HSU, D., KAKADE, S. M. and TELGARSKY, M. (2014). Tensor decompositions for learning latent variable models. *The Journal of Machine Learning Research* **15** 2773–2832.

ANDERSON, T. W. and AMEMIYA, Y. (1988). The asymptotic normal distribution of estimators in factor analysis under general conditions. *The Annals of Statistics* **16** 759–771.

BAI, J. and LI, K. (2012). Statistical analysis of factor models of high dimension. *The Annals of Statistics* **40** 436–465.

BAI, J. and NG, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70** 191–221.

BAIK, J., BEN AROUS, G. and PÉCHÉ, S. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Annals of Probability* 1643–1697.

BARTLETT, M. S. (1938). Methods of estimating mental factors. *Nature* **141** 609–610.

BARTLETT, M. S. (1950). Tests of significance in factor analysis. *British Journal of Mathematical and Statistical Psychology* **3** 77–85.

BEAN, D., BICKEL, P. J., EL KAROUI, N. and YU, B. (2013). Optimal M-estimation in high-dimensional regression. *Proceedings of the National Academy of Sciences* **110** 14563–14568.

BENAYCH-GEORGES, F. and NADAKUDITI, R. R. (2011). The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics* **227** 494–521.

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)* 289–300.

BICKEL, P. J. and LEVINA, E. (2008). Covariance regularization by thresholding. *The Annals of Statistics* **36** 2577–2604.

BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics* **37** 1705–1732.

CAI, T. and LIU, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association* **106** 672–684.

CANDES, E. and TAO, T. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics* **35** 2313–2351.

CANDÈS, E. J., LI, X., MA, Y. and WRIGHT, J. (2011). Robust principal component analysis? *Journal of the ACM (JACM)* **58** 11.

CANDÈS, E. J. and RECHT, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational Mathematics* **9** 717.

CAPE, J., TANG, M. and PRIEBE, C. E. (2017). The two-to-infinity norm and singular subspace geometry with applications to high-dimensional statistics. *arXiv preprint arXiv:1705.10735* .

CATONI, O. (2012). Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, vol. 48. Institut Henri Poincaré.

CATTELL, R. B. (1966). The scree test for the number of factors. *Multivariate behavioral research* **1** 245–276.

CHAMBERLAIN, G. and ROTHSCHILD, M. (1982). Arbitrage, factor structure, and mean-variance analysis on large asset markets.

COHEN, M. B., NELSON, J. and WOODRUFF, D. P. (2015). Optimal approximate matrix product in terms of stable rank. *arXiv preprint arXiv:1507.02268* .

DAVIS, C. and KAHAN, W. M. (1970). The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis* **7** 1–46.

DESAI, K. H. and STOREY, J. D. (2012). Cross-dimensional inference of dependent high-dimensional data. *Journal of the American Statistical Association* **107** 135–151.

DOBRIBAN, E. (2017). Factor selection by permutation. *arXiv preprint arXiv:1710.00479* .

DONOHO, D. L., GAVISH, M. and JOHNSTONE, I. M. (2013). Optimal shrinkage of eigenvalues in the spiked covariance model. *arXiv preprint arXiv:1311.0851* .

EFRON, B. (2007). Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association* **102** 93–103.

EFRON, B. (2010). Correlated z-values and the accuracy of large-scale statistical estimates. *Journal of the American Statistical Association* **105** 1042–1055.

ELDRIDGE, J., BELKIN, M. and WANG, Y. (2017). Unperturbed: spectral analysis beyond Davis-Kahan. *arXiv preprint arXiv:1706.06516* .

FAMA, E. F. and FRENCH, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of financial economics* **33** 3–56.

FAN, J., FAN, Y. and LV, J. (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics* **147** 186–197.

FAN, J. and HAN, X. (2017). Estimation of the false discovery proportion with unknown dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79** 1143–1164.

FAN, J., HAN, X. and GU, W. (2012). Estimating false discovery proportion under arbitrary covariance dependence. *Journal of the American Statistical Association* **107** 1019–1035.

FAN, J., KE, Y., SUN, Q. and ZHOU, W.-X. (2017a). Farm-test: Factor-adjusted robust multiple testing with false discovery control. *arXiv preprint arXiv:1711.05386* .

FAN, J., KE, Y. and WANG, K. (2016a). Decorrelation of covariates for high dimensional sparse regression. *arXiv preprint arXiv:1612.08490* .

FAN, J., LI, Q. and WANG, Y. (2017b). Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79** 247–265.

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association* **96** 1348–1360.

FAN, J., LIAO, Y. and MINCHEVA, M. (2011). High-dimensional covariance matrix estimation in approximate factor models. *The Annals of Statistics* **39** 3320–3356.

FAN, J., LIAO, Y. and MINCHEVA, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75** 603–680.

FAN, J., LIU, H. and WANG, W. (2018a). Large covariance estimation through elliptical factor models. *Annals of Statistics* **46** 1383–1414.

FAN, J., WANG, W. and ZHONG, Y. (2018b). An ℓ_∞ eigenvector perturbation bound and its application. *Journal of Machine Learning Research* **18** 1–42.

FAN, J., WANG, W. and ZHU, Z. (2016b). A shrinkage principle for heavy-tailed data: High-dimensional robust low-rank matrix recovery. *arXiv preprint arXiv:1603.08315* .

FRIGUET, C., KLOAREG, M. and CAUSEUR, D. (2009). A factor model approach to multiple testing under dependence. *Journal of the American Statistical Association* **104** 1406–1415.

GAO, C., MA, Z., ZHANG, A. Y. and ZHOU, H. H. (2015). Achieving optimal misclassification

proportion in stochastic block model. *arXiv preprint arXiv:1505.03772* .

HIRZEL, A. H., HAUSSER, J., CHESSEL, D. and PERRIN, N. (2002). Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? *Ecology* **83** 2027–2036.

HOCHREITER, S., CLEVERT, D.-A. and OBERMAYER, K. (2006). A new summarization method for affymetrix probe level data. *Bioinformatics* **22** 943–949.

HOLLAND, P. W., LASKEY, K. B. and LEINHARDT, S. (1983). Stochastic blockmodels: First steps. *Social networks* **5** 109–137.

HORN, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika* **30** 179–185.

HOTELLING, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology* **24** 417.

HSU, D. and KAKADE, S. M. (2013). Learning mixtures of spherical gaussians: moment methods and spectral decompositions. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*. ACM.

HUBER, P. J. (1964). Robust estimation of a location parameter. *The annals of mathematical statistics* 73–101.

JIN, J. (2015). Fast community detection by score. *The Annals of Statistics* **43** 57–89.

JOHNSTONE, I. M. and LU, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association* **104** 682–693.

JOLLIFFE, I. T. (1986). Principal component analysis and factor analysis. In *Principal component analysis*. Springer, 115–128.

KENDALL, M. G. (1965). A course in multivariate analysis .

KESHAVAN, R. H., MONTANARI, A. and OH, S. (2010). Matrix completion from noisy entries. *Journal of Machine Learning Research* **11** 2057–2078.

KNEIP, A. and SARDA, P. (2011). Factor models and variable selection in high-dimensional regression analysis. *The Annals of Statistics* **39** 2410–2447.

KOLTCHINSKII, V. and LOUNICI, K. (2017). Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli* **23** 110–133.

KOLTCHINSKII, V. and XIA, D. (2016). Perturbation of linear forms of singular vectors under gaussian noise. In *High Dimensional Probability VII*. Springer, 397–423.

LAM, C. and YAO, Q. (2012). Factor modeling for high-dimensional time series: inference for the number of factors. *The Annals of Statistics* **40** 694–726.

LAWLEY, D. and MAXWELL, A. (1962). Factor analysis as a statistical method. *Journal of the Royal Statistical Society. Series D (The Statistician)* **12** 209–229.

LEEK, J. T. and STOREY, J. D. (2008). A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences* **105** 18718–18723.

LELARGE, M. and MIOLANE, L. (2016). Fundamental limits of symmetric low-rank matrix estimation. *arXiv preprint arXiv:1611.03888* .

LI, Q., CHENG, G., FAN, J. and WANG, Y. (2017). Embracing the blessing of dimensionality in factor models. *Journal of the American Statistical Association* 1–10.

MCCRAE, R. R. and JOHN, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of personality* **60** 175–215.

MINSKER, S. (2016). Sub-gaussian estimators of the mean of a random matrix with heavy-tailed entries. *arXiv preprint arXiv:1605.07129* .

MOR-YOSEF, L. and AVRON, H. (2018). Sketching for principal component regression. *arXiv preprint arXiv:1803.02661* .

NG, A. Y., JORDAN, M. I. and WEISS, Y. (2002). On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*.

ONATSKI, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics* **92** 1004–1016.

ONATSKI, A. (2012). Asymptotics of the principal components estimator of large factor models with weakly influential factors. *Journal of Econometrics* **168** 244–258.

O'ROURKE, S., VU, V. and WANG, K. (2016). Eigenvectors of random matrices: a survey. *Journal of Combinatorial Theory, Series A* **144** 361–442.

O'ROURKE, S., VU, V. and WANG, K. (2017). Random perturbation of low rank matrices: Improving classical bounds. *Linear Algebra and its Applications* .

PAUL, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica* 1617–1642.

PAUL, D., BAIR, E., HASTIE, T. and TIBSHIRANI, R. (2008). " preconditioning" for feature selection and regression in high-dimensional problems. *The Annals of Statistics* 1595–1618.

PAULSEN, V. (2002). *Completely bounded maps and operator algebras*, vol. 78. Cambridge University Press.

PEARSON, K. (1901). Principal components analysis. *The London, Edinburgh and Dublin Philosophical Magazine and Journal* **6** 566.

ROHE, K., CHATTERJEE, S. and YU, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics* **39** 1878–1915.

SEDGHI, H., JANZAMIN, M. and ANANDKUMAR, A. (2016). Provable tensor methods for learning mixtures of generalized linear models. In *Artificial Intelligence and Statistics*.

SHKOLNISKY, Y. and SINGER, A. (2012). Viewing direction estimation in cryo-EM using synchronization. *SIAM journal on imaging sciences* **5** 1088–1110.

SPEARMAN, C. (1927). The abilities of man. .

SRIVASTAVA, N. and VERSHYNIN, R. (2013). Covariance estimation for distributions with $2 + \varepsilon$ moments. *The Annals of Probability* **41** 3081–3111.

STEWART, G. and SUN, J. (1990). Matrix perturbation theory .

STOCK, J. H. and WATSON, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American statistical association* **97** 1167–1179.

STOREY, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64** 479–498.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 267–288.

TRON, R. and VIDAL, R. (2009). Distributed image-based 3-D localization of camera sensor networks. In *Decision and Control, 2009 held jointly with the 2009 28th Chinese Control Conference. CDC/CCC 2009. Proceedings of the 48th IEEE Conference on*. IEEE.

TROPP, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics* **12** 389–434.

VERSHYNIN, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027* .

VERSHYNIN, R. (2012). How close is the sample covariance matrix to the actual covariance matrix? *Journal of Theoretical Probability* **25** 655–686.

WANG, H. (2012). Factor profiled sure independence screening. *Biometrika* **99** 15–28.

WANG, J., ZHAO, Q., HASTIE, T. and OWEN, A. B. (2017). Confounder adjustment in multiple hypothesis testing. *The Annals of Statistics* **45** 1863–1894.

WANG, W. and FAN, J. (2017). Asymptotics of empirical eigenstructure for high dimensional spiked covariance. *Ann. Statist.* **45** 1342–1374.

WEDIN, P.-A. (1972). Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics* **12** 99–111.

WOODRUFF, D. P. (2014). Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science* **10** 1–157.

YANG, J., MENG, X. and MAHONEY, M. W. (2016). Implementing randomized matrix algorithms in parallel and distributed environments. *Proceedings of the IEEE* **104** 58–92.

YI, X., CARAMANIS, C. and SANGHAVI, S. (2016). Solving a mixture of many random linear equations by tensor decomposition and alternating minimization. *arXiv preprint arXiv:1608.05749* .

YU, Y., WANG, T. and SAMWORTH, R. J. (2014). A useful variant of the davis–kahan theorem for statisticians. *Biometrika* **102** 315–323.

ZHAO, P. and YU, B. (2006). On model selection consistency of lasso. *Journal of Machine learning research* **7** 2541–2563.

ZHONG, Y. (2017). Eigenvector under random perturbation: A nonasymptotic Rayleigh-Schrö dinger theory. *arXiv preprint arXiv:1702.00139* .

ZHONG, Y. and BOUMAL, N. (2018). Near-optimal bounds for phase synchronization. *SIAM Journal on Optimization* **28** 989–1016.

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67** 301–320.