FARM-Test: Factor-Adjusted Robust Multiple Testing with False Discovery Control

Article *in* Journal of the American Statistical Association · November 2017

DOI: 10.1080/01621459.2018.1527700

4 authors:

Jianqing Fan
Princeton University

Yuan Ke
University of Georgia

Qiang Sun
University of Toronto

Wen-Xin Zhou
University of California, San Diego

arXiv:1711.05386v1 [stat.ME] 15 Nov 2017

FARM-Test: Factor-adjusted robust multiple testing with false discovery control*

Jianqing Fan[†], Yuan Ke[‡], Qiang Sun[§] and Wen-Xin Zhou[¶]

Abstract

Large-scale multiple testing with correlated and heavy-tailed data arises in a wide range of research areas from genomics, medical imaging to finance. Conventional methods for estimating the false discovery proportion (FDP) often ignore the effect of heavy-tailedness and the dependence structure among test statistics, and thus may lead to inefficient or even inconsistent estimation. Also, the assumption of joint normality is often imposed, which is too stringent for many applications. To address these challenges, in this paper we propose a factor-adjusted robust procedure for large-scale simultaneous inference with control of the false discovery proportion. We demonstrate that robust factor adjustments are extremely important in both improving the power of the tests and controlling FDP. We identify general conditions under which the proposed method produces consistent estimate of the FDP. As a byproduct that is of independent interest, we establish an exponential-type deviation inequality for a robust U-type covariance estimator under the spectral norm. Extensive numerical experiments demonstrate the advantage of the proposed method over several state-of-the-art methods especially when the data are generated from heavy-tailed distributions. Our proposed procedures are implemented in the R-package FarmTest.

*Supported by NSF Grants DMS-1662139, DMS-1712591, and NIH Grant R01-GM072611-12.

†Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA. E-mail: jqfan@princeton.edu.

‡Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA. E-mail: yuank@princeton.edu.

§Department of Statistical Sciences, University of Toronto, Toronto, ON M5S 3G3. E-mail: qsun.ustc@gmail.com.

¶Department of Mathematics, University of California, San Diego, La Jolla, CA 92093. E-mail: wez243@ucsd.edu.

Keywords: Factor adjustment; False discovery proportion; Huber loss; Large-scale multiple testing; Nonasymptotic robustness.

1 Introduction

Large-scale multiple testing problems with independent test statistics have been extensively explored and is now well understood in both practice and theory (Benjamini and Hochberg, 1995; Storey, 2002; Genovese and Wasserman, 2004; Lehmann and Romano, 2005). Yet, in practice, correlation effects often exist across many observed test statistics. For instance, in neuroscience studies, although the neuroimaging data may appear very high dimensional (with millions of voxels), the effect degrees of freedom are generally much smaller, due to spatial correlation and spatial continuity (Medland et al., 2014). In genomic studies, genes are usually correlated regulatorily or functionally: multiple genes may belong to the same regulatory pathway or there may exist gene-gene interactions. Ignoring these dependence structures will cause loss of statistical power or even lead to inconsistent estimates.

To understand the effect of dependencies on multiple testing problems, validity of standard multiple testing procedures have been studied under weak dependencies, see Benjamini and Yekutieli (2001), Storey (2003), Storey *et al.* (2004), Ferreira and Zwinderman (2006), Wu (2008), Clarke and Hall (2009), Blanchard and Roquain (2009) and Liu and Shao (2014), among others. For example, it has been shown that, the Benjamini-Hochberg procedure or Storey's procedure, is still able to control the false discovery rate (FDR) or false discovery proportion, when only weak dependencies are present. Nevertheless, multiple testing under general and strong dependence structures remains a challenge. Directly applying standard FDR controlling procedures developed for independent test statistics in this case can lead to inaccurate false discovery control and spurious outcomes. Therefore, correlations must be accounted for in the inference procedure; see, for example, Owen (2005), Efron (2007, 2010), Leek and Storey (2008), Sun and Cai (2009), Friguet *et al.* (2009), Schwartzman and Lin (2011), Fan *et al.* (2012), Desai and Storey (2012), Wang *et al.* (2015) and Fan and Han (2017) for an unavoidably incomplete overview.

In this paper, we focus on the case where the dependence structure can be characterized by latent factors, that is, there exist a few unobserved variables that correlate with the outcome. A multi-factor model is an effective tool for modeling dependence, with wide applications in genomics (Kustra *et al.*, 2006), neuroscience (Pournara and Wernish, 2007) and financial economics (Bai, 2003). It relies on the identification of a linear space of random vectors capturing the dependence structure of the data. In Friguet *et al.* (2009) and Desai and Storey (2012), the authors assume a strict factor model with independent idiosyncratic errors, and use the EM algorithm to estimate the factor loadings as well as the realized factors. The FDP is then estimated by subtracting out the realized common factors. Fan *et al.* (2012) considered a general setting for estimating the FDP, where the test statistics follow a multivariate normal distribution with an arbitrary but known covariance structure. Later, Fan and Han (2017) used the POET estimator (Fan *et al.*, 2013) to estimate the unknown covariance matrix, and then proposed a fully data-driven estimate of the FDP. Recently, Wang *et al.* (2015) considered a more complex model with both observed primary variables and unobserved latent factors.

All the methods above assume joint normality of factors and noise, and thus methods based on least square regression, or likelihood generally, can be applied. However, normality is really an idealization of the complex random world. For example, the distribution of the normalized gene expressions is often far from normal, regardless of the normalization methods used (Purdom and Holmes, 2005). Heavy-tailed data also frequently appear in many other scientific fields, such as financial engineering (Cont, 2001) and biomedical imaging (Eklund *et al.*, 2016). In finance, the seminal papers by Mandelbrot (1963) and Fama (1963) discussed the power law behavior of asset returns, and Cont (2001) provided extensive evidence of heavy-tailedness in financial returns. More recently, in functional MRI studies, it has been observed by Eklund *et al.* (2016) that the parametric statistical methods failed to produce valid clusterwise inference, where the principal cause is that the spatial autocorrelation functions do not follow the assumed Gaussian shape. The heavy-tailedness issue may further be amplified by high dimensionality in large scale inference. In the context of multiple testing, as the dimension gets larger, more outliers are likely to appear, and this may lead to significant false discoveries. Therefore, it is imperative to develop inferential procedures that are dependence-adjusted and robust to heavy-tailedness at the same time.

In this paper, we investigate the problem of large-scale multiple testing under dependence via an approximate factor model, where the outcome variables are correlated with each other through latent factors. To simultaneously incorporate the dependencies and tackle with heavy-tailed data, we propose a factor-adjusted robust multiple testing (FARM-Test) procedure. As we proceed, we gradually unveil the whole procedure in four steps. First, we consider an oracle factor-adjusted procedure given the knowledge of the factors and loadings, which provides the key insights into the problem. Next, using the idea of adaptive Huber regression (Sun *et al.*, 2017), we consider estimating the realized factors if the loadings were known and provide a robust control of the FDP. In the third part, we propose two robust covariance matrix estimators, a robust U-type covariance matrix estimator and another one based on elementwise truncation. We then apply spectral decomposition to these estimators and use principal factors to recover the factor loadings. The final part gives a fully data-driven testing procedure based on sample splitting: use part of the data for loading construction and other part for simultaneous inference.

First we illustrate our methodology with a numerical example that consists of observations $\boldsymbol{X}_i$'s generated from a three-factor model:

$$\boldsymbol{X}_i = \boldsymbol{\mu} + \mathbf{B}\boldsymbol{f}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, n,$$

where $\boldsymbol{f}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_3)$ and the entries of $\mathbf{B}$ are a realization of independent and identically distributed (i.i.d.) following a uniform distribution, $\mathcal{U}(-1, 1)$. The idiosyncratic errors, $\boldsymbol{\varepsilon}_i$'s, are independently generated from the t-distribution with 3 degrees of freedom. The sample size n and dimension p are set to be 100 and 500, respectively. We take the true means to be $\mu_j = 0.6$ for $1 \leqslant j \leqslant 0.25 \times p$ and 0 otherwise. In Figure 1, we plot the histograms of sample means, robust mean estimators, and their counterparts with factor-adjustment. Details of robust mean estimation and the related factor-adjusted procedure are specified in Sections 2 and 3. Due to the existence of latent factors and heavy-tailed errors, there is a large overlap between sample means from the null and alternative, which makes it difficult to distinguish them from each other. With the help of either robustification or factor-adjustment, the null and alternative are better separated as shown in the figure. Further, applying both

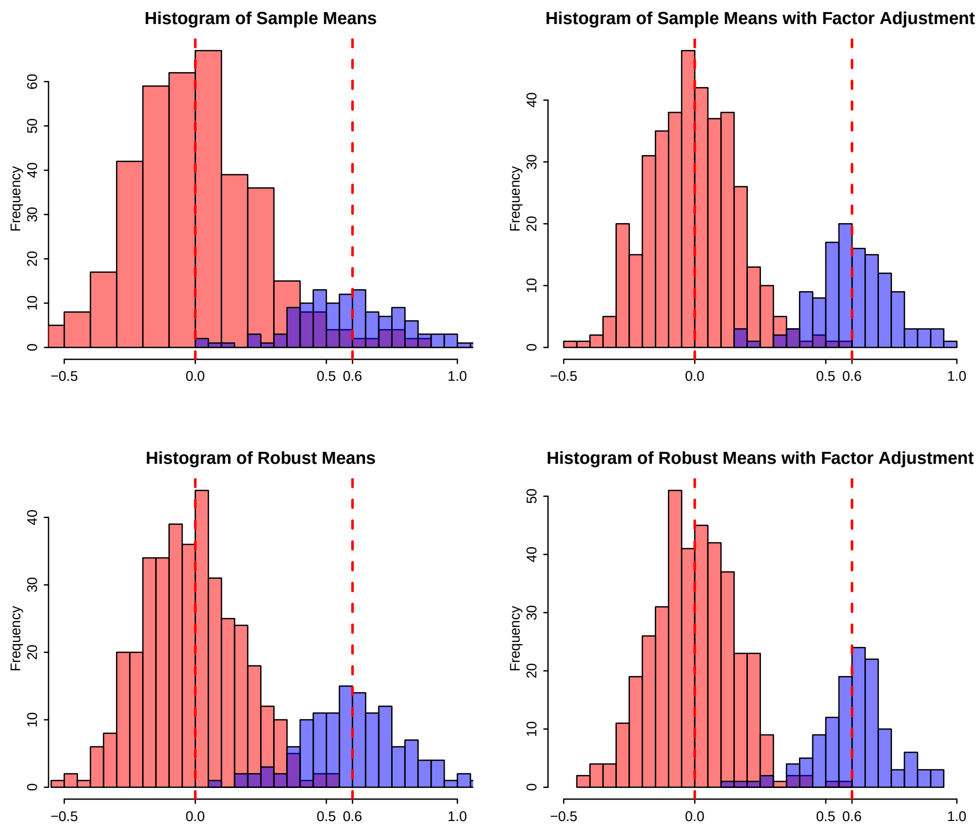

Figure 1: Histograms of four different mean estimators for simultaneous inference.

factor-adjustment and robustification, the resulting mean estimators are well concentrated around the true means and therefore differentiate the signal from the noise more easily. This example demonstrates the effectiveness of the factor-adjusted robust multiple testing procedure.

The rest of the paper proceeds as follows. In Section 2, we describe a generic factor-adjusted robust multiple testing procedure under the approximate factor model. In Section 3, we gradually unfold the proposed method, while we establish its theoretical properties along the way. Section 4 is devoted to simulated numerical studies. Section 5 analyzes an empirical dataset. We conclude the paper in Section 6. Proofs of the main theorems and technical lemmas are provided in the online supplement.

2 FARM-Test

In this section, we described a generic factor-adjusted robust multiple testing procedure under the approximate factor model.

2.1 Problem setup

Let $\boldsymbol{X} = (X_1, \ldots, X_p)^{\mathrm{T}}$ be a p-dimensional random vector with mean $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^{\mathrm{T}}$ and covariance matrix $\boldsymbol{\Sigma} = (\sigma_{jk})_{1 \leqslant j,k \leqslant p}$. We assume the dependence structure in $\boldsymbol{X}$ is captured by a few latent factors such that $\boldsymbol{X} = \boldsymbol{\mu} + \mathbf{B}\boldsymbol{f} + \boldsymbol{\varepsilon}$, where $\mathbf{B} = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_p)^{\mathrm{T}} \in \mathbb{R}^{p \times K}$ is the deterministic factor loading matrix, $\boldsymbol{f} = (f_{i1}, \ldots, f_{iK})^{\mathrm{T}} \in \mathbb{R}^K$ the zero-mean latent random factor, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_p)^{\mathrm{T}} \in \mathbb{R}^p$ the zero-mean idiosyncratic error that is uncorrelated with $\boldsymbol{f}$. Suppose we observe n random samples $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ from $\boldsymbol{X}$, satisfying

$$\boldsymbol{X}_i = \boldsymbol{\mu} + \mathbf{B}\boldsymbol{f}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, n, \tag{2.1}$$

where $\boldsymbol{f}_i$'s and $\boldsymbol{\varepsilon}_i$'s are independent and identically distributed samples of $\boldsymbol{f}$ and $\boldsymbol{\varepsilon}$, respectively. Assume that $\boldsymbol{f}$ and $\boldsymbol{\varepsilon}$ have covariance matrices $\boldsymbol{\Sigma}_f$ and $\boldsymbol{\Sigma}_\varepsilon = (\sigma_{\varepsilon,jk})_{1 \leqslant j,k \leqslant p}$, respectively. To make $\mathbf{B}$ and $\boldsymbol{f}$ identifiable, as in Bai and Li (2012), we assume

$$\boldsymbol{\Sigma}_f = \mathbf{I}_K \quad \text{and} \quad \mathbf{B}^{\mathrm{T}}\mathbf{B} \text{ is diagonal.} \tag{2.2}$$

In this paper, we are interested in simultaneously testing the following hypotheses

$$\mathcal{H}_{0j} : \mu_j = 0 \quad \text{versus} \quad \mathcal{H}_{1j} : \mu_j \neq 0, \quad \text{for } 1 \leqslant j \leqslant p, \tag{2.3}$$

based on the observed data $\{\boldsymbol{X}_i\}_{i=1}^n$. Many existing works (e.g. Fan *et al.*, 2012; Fan and Han, 2017) in the literature assume multivariate normality of the idiosyncratic errors. However, the Gaussian assumption on the sampling distribution is often unrealistic in many practical applications. For each feature, the measurements across different subjects consist of samples from potentially different distributions with quite different scales, and thus can be highly skewed and heavy-tailed. In the big data regime, we are often dealing with thousands or tens of thousands of features simultaneously. Simply by chance, some variables exhibits heavy tails. As a consequence, with the number of variables grows, some outliers may turn out to be so large that they can be mistakenly regarded as discoveries. Therefore, it is imperative to develop robust alternatives that are insensitive to outliers and data contaminations.

For each $1 \leqslant j \leqslant p$, let T_j be a generic test statistic for testing the individual hypothesis $\mathcal{H}_{0j}$. For a prespecified thresholding level $z > 0$, we reject the j-th null hypothesis whenever $|T_j| \geqslant z$. The number of total discoveries $R(z)$ and the number of false discoveries $V(z)$ can be written as

$$R(z) = \sum_{j=1}^{p} 1\big(|T_j| \geqslant z\big) \quad \text{and} \quad V(z) = \sum_{j \in \mathcal{H}_0} 1\big(|T_j| \geqslant z\big), \tag{2.4}$$

respectively, where $\mathcal{H}_0 := \{j : 1 \leqslant j \leqslant p, \mu_j = 0\}$ is the set of the true nulls with cardinality $p_0 = |\mathcal{H}_0| = \sum_{j=1}^{p} 1(\mu_j = 0)$. We are mainly interested in controlling the false discovery proportion, $\mathrm{FDP}(z) = V(z)/R(z)$ with the convention $0/0 = 0$. We remark here that $R(z)$ is observable given the data, while $V(z)$, which depends on the set of true nulls, is an unobserved random quantity that needs to be estimated. Comparing with FDR control, controlling FDP is arguably more relevant as it is directly related to the current experiment.

2.2 A generic procedure

We now propose a Factor Adjusted Robust Multiple Testing procedure, which we call FARM-Test. As the name suggests, this procedure utilizes the dependence structure in $\boldsymbol{X}$ and is robust against heavy tailedness of the error distributions. Recent studies in Fan *et al.* (2017) and Sun *et al.* (2017) show that the Huber estimator (Huber, 1964) with a properly diverging robustification parameter admits a sub-Gaussian-type deviation bound for heavy-tailed data under mild moment conditions. This new perspective motivates new methods, as described below. To begin with, we formally introduce the Huber loss and the robustification parameter.

Definition 2.1 (Huber Loss and Robustification Parameter)**.** The Huber loss $\ell_\tau(\cdot)$ (Huber,

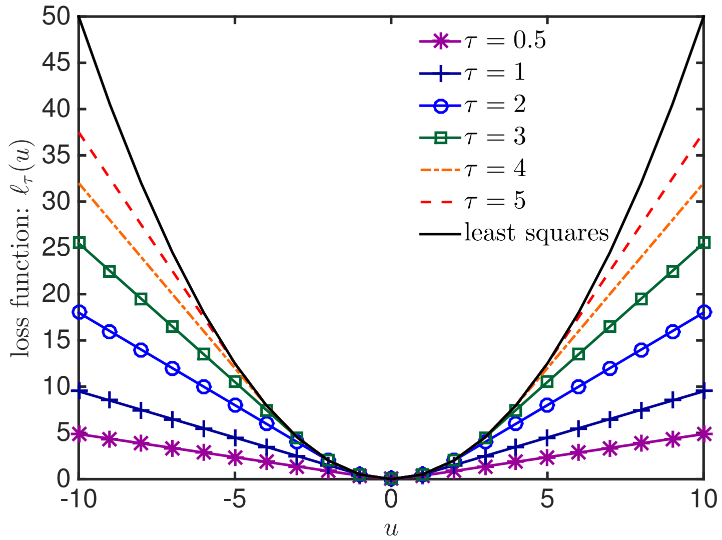

Figure 2: The Huber loss function $\ell_\tau(\cdot)$ with varying robustification parameters. The ℓ_2-loss function is shown for comparison.

1964) is defined as

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

where τ depends on the moment condition, sample size and dimension nonasymptotically. In Sun *et al.* (2017), τ is referred to as the *robustification parameter* that trades bias for robustness.

We refer to the Huber loss in Definition 2.1 above as the adaptive Huber loss to recognize the nonasymptotic adaptivity of the robustification parameter τ. For any $1 \leqslant j \leqslant p$, with a robustification parameter $\tau_j > 0$, we consider the following adaptive and robust M-estimator of μ_j:

$$\widehat{\mu}_j = \underset{\theta \in \mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^{n} \ell_{\tau_j}(X_{ij} - \theta), \tag{2.5}$$

where we suppress the dependence of $\widehat{\mu}_j$ on τ_j for simplicity. As shown in our theoretical results, the robustification parameter τ plays an important role in controlling the bias-robustness tradeoff. To guarantee the asymptotic normality of $\widehat{\mu}_j$ uniformly over $j = 1, \ldots p$, and to achieve optimal bias-robustness tradeoff, we choose $\tau = \tau(n, p)$ of the form

$C\sqrt{n/\log(np)}$, where the constant $C > 0$ can be selected via cross-validation. Specifically, we show that $\sqrt{n}\,(\widehat{\mu}_j - \boldsymbol{b}_j^{\mathrm{T}} \bar{\boldsymbol{f}})$ is asymptotically normal with mean μ_j and variance $\sigma_{\varepsilon,jj}$ (with details given in the supplemental material):

$$\sqrt{n}\,\big(\widehat{\mu}_j - \mu_j - \boldsymbol{b}_j^{\mathrm{T}} \bar{\boldsymbol{f}}\,\big) = \mathcal{N}\big(0, \sigma_{\varepsilon,jj}\big) + o_{\mathbb{P}}(1) \ \text{ uniformly over } j = 1, \ldots, p. \tag{2.6}$$

Here, $\widehat{\mu}_j$'s can be regarded as robust versions of the sample averages $\bar{X}_j = \mu_j + \boldsymbol{b}_j^{\mathrm{T}} \bar{\boldsymbol{f}} + \bar{\varepsilon}_j$, where $\bar{X}_j = n^{-1} \sum_{i=1}^n X_{ij}$ and $\bar{\varepsilon}_j = n^{-1} \sum_{i=1}^n \varepsilon_{ij}$.

Given a prespecified level $\alpha \in (0, 1)$, our testing procedure consists of three steps: (i) robust estimation of the loading vectors and factors; (ii) construction of factor-adjusted marginal test statistics and their P-values; and (iii) computing the critical value or threshold level with the estimated FDP controlled at α. The detailed algorithm is stated as follows.

We expect that the factor-adjusted test statistic T_j given in (2.8) is close in distribution to standard normal for all $j = 1, \ldots, p$. Hence, according to the law of large numbers, the number of false discoveries $V(z) = \sum_{j \in \mathcal{H}_0} 1(|T_j| \geqslant z)$ should be close to $2p_0 \Phi(-z)$ for any $z \geqslant 0$. The number of null hypotheses p_0 is typically unknown. In the high dimensional and sparse regime, where both p and p_0 are large and $p_1 = p - p_0 = o(p)$ is relatively small, $\mathrm{FDP}^{\mathrm{A}}$ in (2.9) serves as a slightly conservative surrogate for the asymptotic approximation $2p_0 \Phi(-z)/R(z)$. If the proportion $\pi_0 = p_0/p$ is bounded away from 1 as $p \to \infty$, $\mathrm{FDP}^{\mathrm{A}}$ tends to overestimate the true FDP. The estimation of π_0 has long been known as an interesting problem. See, for example, Storey (2002), Langaas and Lindqvist (2005), Meinshausen and Rice (2006), Jin and Cai (2007) and Jin (2008), among others. Therefore, a more adaptive method is to combine the above procedure with, for example Storey's approach, to calibrate the rejection region for individual hypotheses. Let $\{P_j = 2\Phi(-|T_j|)\}_{j=1}^p$ be the approximate P-values. For a predetermined $\eta \in [0, 1)$, Storey (2002) suggested to estimate π_0 by

$$\widehat{\pi}_0(\eta) = \frac{1}{(1-\eta)p} \sum_{j=1}^p 1\big(P_j > \eta\big). \tag{2.10}$$

The fundamental observation that underpins Storey's procedure is that most of the large P-values come from the true null hypotheses and thus are uniformly distributed. For a

FARM-Test algorithm.

Input: Observations $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^{\mathrm{T}} \in \mathbb{R}^p$ for $i = 1, \ldots, n$, a prespecified level $\alpha \in (0,1)$ and an integer $K \geq 1$.

Procedure:

Step 1. Denote by $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{\Sigma}}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) \in \mathbb{R}^{p\times p}$ a generic robust covariance matrix estimator. Let $\widehat{\lambda}_1 \geqslant \widehat{\lambda}_2 \geqslant \cdots \geqslant \widehat{\lambda}_K$ be the top K eigenvalues of $\widehat{\boldsymbol{\Sigma}}$, and let $\widehat{\boldsymbol{v}}_1, \widehat{\boldsymbol{v}}_2, \ldots, \widehat{\boldsymbol{v}}_K$ be the corresponding eigenvectors. Define $\widehat{\mathbf{B}} = (\widetilde{\lambda}_1^{1/2}\widehat{\boldsymbol{v}}_1, \ldots, \widetilde{\lambda}_K^{1/2}\widehat{\boldsymbol{v}}_K) \in \mathbb{R}^{p\times K}$, where $\widetilde{\lambda}_k = \max(\widehat{\lambda}_k, 0)$ for $k = 1, \ldots, K$. Let $\widehat{\boldsymbol{b}}_1, \ldots, \widehat{\boldsymbol{b}}_p \in \mathbb{R}^K$ be the p rows of $\widehat{\mathbf{B}}$, and define

$$\widehat{\boldsymbol{f}} = \underset{\boldsymbol{f}\in\mathbb{R}^K}{\mathrm{argmin}} \sum_{j=1}^{p} \ell_\gamma(\bar{X}_j - \widehat{\boldsymbol{b}}_j^{\mathrm{T}}\boldsymbol{f}), \tag{2.7}$$

where $\gamma = \gamma(n,p) > 0$ is a robustification parameter.

Step 2. Construct factor-adjusted test statistics

$$T_j = \left(\frac{n}{\widehat{\sigma}_{\varepsilon,jj}}\right)^{1/2} (\widehat{\mu}_j - \widehat{\boldsymbol{b}}_j^{\mathrm{T}}\widehat{\boldsymbol{f}}), \quad j = 1, \ldots, p, \tag{2.8}$$

where $\widehat{\sigma}_{\varepsilon,jj} = \widehat{\theta}_j - \widehat{\mu}_j^2 - \|\widehat{\boldsymbol{b}}_j\|_2^2$, $\widehat{\theta}_j = \mathrm{argmin}_{\theta>\widehat{\mu}_j^2+\|\widehat{\boldsymbol{b}}_j\|_2^2} \sum_{i=1}^n \ell_{\tau_{jj}}(X_{ij}^2 - \theta)$ and τ_{jj}'s are robustification parameters. Here, we use the fact that $\mathbb{E}(X_j^2) = \mu_j^2 + \|\boldsymbol{b}_j\|_2^2 + \mathrm{var}(\varepsilon_j)$, according to the identifiability condition.

Step 3. Calculate the critical value z_α as

$$z_\alpha = \inf\left\{z \geq 0 : \mathrm{FDP}^{\mathrm{A}}(z) \leqslant \alpha\right\}, \tag{2.9}$$

where $\mathrm{FDP}^{\mathrm{A}}(z) = 2p\Phi(-z)/R(z)$ denotes the approximate FDP and $R(z)$ is as in (2.4). Finally, for $j = 1, \ldots, p$, reject H_{0j} whenever $|T_j| \geqslant z_\alpha$.

---

sufficiently large λ, about $(1-\eta)\pi_0$ of the P-values are expected to lie in $(\eta, 1]$. Thus, the proportion of P-values that exceed η should be close to $(1-\eta)\pi_0$. A value of $\eta = 1/2$ is used in the SAM software (Storey and Tibshirani, 2003); while it was shown in Blanchard and Roquain (2009) that the choice $\eta = \alpha$ may have better properties for dependent P-values.

Incorporating the above estimate of π_0, a modified estimate of FDP takes the form

$$\mathrm{FDP}^{\mathrm{A}}(z;\eta) = 2p\,\widehat{\pi}_0(\eta)\Phi(-z)/R(z),\ z \geqslant 0.$$

Finally, for any prespecified $\alpha \in (0,1)$, we reject H_{0j} whenever $|T_j| \geqslant z_{\alpha,\eta}$, where

$$z_{\alpha,\eta} = \inf\big\{z \geqslant 0 : \mathrm{FDP}^{\mathrm{A}}(z;\eta) \leqslant \alpha\big\}. \tag{2.11}$$

By definition, it is easy to see that $z_{\alpha,0}$ is equal to z_α given in (2.9).

3 Theoretical properties

To fully understand the impact of factor-adjustment as well as robust estimation, we gradually investigate the theoretical properties of the FARM-Test through several steps, starting with an oracle procedure which provides key insights into the problem.

3.1 An oracle procedure

First we consider an oracle procedure that serves as a heuristic device. In this section, we assume the loading matrix $\mathbf{B}$ is known and the factors $\{\boldsymbol{f}_i\}_{i=1}^n$ are observable. In this case, it is natural to use the factor-adjusted data: $\boldsymbol{Y}_i = \boldsymbol{X}_i - \mathbf{B}\boldsymbol{f}_i = \boldsymbol{\mu} + \boldsymbol{\varepsilon}_i$, which has smaller componentwise variances (which are $\{\sigma_{\varepsilon,jj}\}_{j=1}^p$ and assumed known for the moment) than those of $\boldsymbol{X}_i$. Thus, instead of using $\sqrt{n}\,\widehat{\mu}_j$ given in (2.5), it is more efficient to construct robust mean estimates using factor-adjusted data. This is essentially the same as using the test statistic

$$T_j^{\circ} = \left(\frac{n}{\sigma_{\varepsilon,jj}}\right)^{1/2} \big(\widehat{\mu}_j - \boldsymbol{b}_j^{\mathrm{T}} \bar{\boldsymbol{f}}\,\big), \tag{3.1}$$

whose distribution is close to the standard normal distribution under the j-th null hypothesis. Recall that $p_0 = |\mathcal{H}_0|$ is the number of true null hypotheses. Then, for any $z \geqslant 0$,

$$\frac{1}{p_0} V(z) = \frac{1}{p_0} \sum_{j \in \mathcal{H}_0} 1\big(|T_j^{\circ}| \geqslant z\big).$$

Intuitively, using the (conditional) law of large numbers yields $p_0^{-1}V(z) = 2\Phi(-z) + o_{\mathbb{P}}(1)$. Hence, the FDP based on oracle test statistics admits an asymptotic expression

$$\mathrm{AFDP}_{\mathrm{orc}}(z) = 2p_0\Phi(-z)/R(z),\ z \geqslant 0, \tag{3.2}$$

where "AFDP" stands for the asymptotic FDP and a subscript "orc" is added to highlight the use of oracle statistics.

Remark 1. For testing the individual hypothesis $\mathcal{H}_{0j}$, Fan and Han (2017) considered the test statistic $\sqrt{n}\bar{X}_j$, where $\bar{X}_j = (1/n)\sum_{i=1}^n X_{ij}$. The empirical means, without factor adjustments, are inefficient, as elucidated above. In addition, they are sensitive to the tails of the error distributions (Catoni, 2012). In fact, with many collected variables, by chance only, some test statistics $\sqrt{n}\bar{X}_j$ can be so large empirically that they may be mistakenly regarded as discoveries.

We will show that $\mathrm{AFDP}_{\mathrm{orc}}(z)$ provides a valid asymptotic approximation of the (unknown) true FDP using oracle statistics $\{T_j^\circ\}$ in high dimensions. The latter will be denoted as $\mathrm{FDP}_{\mathrm{orc}}(z)$. Let $\mathbf{R}_\varepsilon = (r_{\varepsilon,jk})_{1\leqslant j,k\leqslant p}$ be the correlation matrix of $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_p)^{\mathrm{T}}$, i.e., $\mathbf{R}_\varepsilon = \mathbf{D}_\varepsilon^{-1}\mathbf{\Sigma}_\varepsilon\mathbf{D}_\varepsilon^{-1}$ where $\mathbf{D}_\varepsilon^2 = \mathrm{diag}(\sigma_{\varepsilon,11}, \ldots, \sigma_{\varepsilon,pp})$. Moreover, write

$$\omega_{n,p} = \sqrt{\frac{n}{\log(np)}}. \tag{3.3}$$

We impose the following moment and regularity assumptions.

Assumption 1. (i) $p = p(n) \to \infty$ and $\log(p) = o(\sqrt{n})$ as $n \to \infty$; (ii) $\boldsymbol{X} \in \mathbb{R}^p$ follows the approximate factor model (2.1) with $\boldsymbol{f}$ and $\boldsymbol{\varepsilon}$ being independent; (iii) $\mathbb{E}(\boldsymbol{f}) = \mathbf{0}$, $\mathrm{cov}(\boldsymbol{f}) = \mathbf{I}_K$ and $\|\boldsymbol{f}\|_{\psi_2} \leqslant A_f$ for some constant $A_f > 0$. Here $\|\cdot\|_{\psi_2}$ denotes the vector sub-Gaussian norm (Vershynin, 2012); (iv) There exist constants $C_\varepsilon, c_\varepsilon > 0$ such that $c_\varepsilon \leqslant \min_{1\leqslant j\leqslant p}\sigma_{\varepsilon,jj}^{1/2} \leqslant \max_{1\leqslant j\leqslant p}\upsilon_j \leqslant C_\varepsilon$, where $\upsilon_j := (\mathbb{E}\varepsilon_j^4)^{1/4}$; (v) There exist constants $\kappa_0 \in (0,1)$ and $\kappa_1 > 0$ such that $\max_{1\leqslant j,k\leqslant p}|r_{\varepsilon,jk}| \leqslant \kappa_0$ and $p^{-2}\sum_{1\leqslant j,k\leqslant p}|r_{\varepsilon,jk}| = O(p^{-\kappa_1})$ as $p \to \infty$.

Theorem 3.1. Suppose that Assumption 1 holds and $p_0 \geqslant ap$ for some constant $a \in (0,1)$. Let $\tau_j = a_j\omega_{n,p}$ with $a_j \geqslant \sigma_{jj}^{1/2}$ for $j = 1, \ldots, p$, where $\omega_{n,p}$ is given by (3.3). Then, for any

$z \geqslant 0$, on the event $\{p^{-1}R(z) \geqslant c\}$ for some $c > 0$, we have

$$\big|\mathrm{FDP}_{\mathrm{orc}}(z) - \mathrm{AFDP}_{\mathrm{orc}}(z)\big| = o_{\mathbb{P}}(1) \quad \text{as } n, p \to \infty.$$

3.2 Robust estimation of loading matrix

Next we focus on estimating $\mathbf{B}$ under the identifiability condition (2.2). Write $\mathbf{B} = (\bar{\boldsymbol{b}}_1, \ldots, \bar{\boldsymbol{b}}_K)$ and assume without loss of generality that $\bar{\boldsymbol{b}}_1, \ldots, \bar{\boldsymbol{b}}_K \in \mathbb{R}^p$ are ordered such that $\{\|\bar{\boldsymbol{b}}_\ell\|_2\}_{\ell=1}^K$ is in a non-increasing order. In this notation, we have $\boldsymbol{\Sigma} = \sum_{\ell=1}^K \bar{\boldsymbol{b}}_\ell \bar{\boldsymbol{b}}_\ell^{\mathrm{T}} + \boldsymbol{\Sigma}_\varepsilon$, and $\bar{\boldsymbol{b}}_{\ell_1}^{\mathrm{T}} \bar{\boldsymbol{b}}_{\ell_2} = 0$ for $1 \leqslant \ell_1 \neq \ell_2 \leqslant K$. Let $\lambda_1, \ldots, \lambda_p$ be the eigenvalues of $\boldsymbol{\Sigma}$ in a descending order, with associated eigenvectors denoted by $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p \in \mathbb{R}^p$. By Weyl's theorem,

$$\big|\lambda_j - \|\bar{\boldsymbol{b}}_j\|_2^2\big| \leqslant \|\boldsymbol{\Sigma}_\varepsilon\| \quad \text{for } 1 \leq j \leq K \quad \text{and} \quad |\lambda_j| \leqslant \|\boldsymbol{\Sigma}_\varepsilon\| \quad \text{for } j > K.$$

Moreover, under the pervasiveness condition (see Assumption 2 below), the eigenvectors $\boldsymbol{v}_j$ and $\bar{\boldsymbol{b}}_j / \|\bar{\boldsymbol{b}}_j\|_2$ of $\boldsymbol{\Sigma}$ and $\mathbf{B}\mathbf{B}^{\mathrm{T}}$, respectively, are also close to each other for $1 \leqslant j \leqslant K$. The estimation of $\mathbf{B}$ thus depends heavily on estimating $\boldsymbol{\Sigma}$ along with its eigenstructure.

In Section 3.2.1 and 3.2.2, we propose two different robust covariance matrix estimators that are also of independent interest. Then, the construction of $\widehat{\mathbf{B}}$ follows from Step 1 of the FARM-Test algorithm described in Section 2.2.

3.2.1 U-type covariance estimation

First, we propose a U-type covariance matrix estimator, which leads to estimates of the unobserved factors under condition (2.2). Let $\psi_\tau(\cdot)$ be the derivative of $\ell_\tau(\cdot)$ given by

$$\psi_\tau(u) = \big(|u| \wedge \tau\big)\,\mathrm{sgn}(u), \quad u \in \mathbb{R},$$

which is a simple truncation operation. Given n scalar random variables $X_1, \ldots, X_n$ from X with mean μ, a fast and robust estimator of μ is given by

$$\widehat{\mu}_\tau = \frac{1}{n} \sum_{i=1}^n \psi_\tau(X_i).$$

Recently, Minsker (2016) extended this univariate estimation scheme to matrix settings based on the following definition on matrix functionals.

Definition 3.1. Given a real-valued function f defined on $\mathbb{R}$ and a symmetric $\mathbf{A} \in \mathbb{R}^{K\times K}$ with eigenvalue decomposition $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathrm{T}}$ such that $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_K)$, $f(\mathbf{A})$ is defined as $f(\mathbf{A}) = \mathbf{U}f(\mathbf{\Lambda})\mathbf{U}^{\mathrm{T}}$, where $f(\mathbf{\Lambda}) = \mathrm{diag}(f(\lambda_1), \ldots, f(\lambda_K))$.

Suppose we observe n random samples $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ from $\boldsymbol{X}$ with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma} = \mathbb{E}\{(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^{\mathrm{T}}\}$. If $\boldsymbol{\mu}$ were known, a robust estimator of $\mathbf{\Sigma}$ can be simply constructed by $(1/n)\sum_{i=1}^n \psi_\tau\{(\boldsymbol{X}_i-\boldsymbol{\mu})(\boldsymbol{X}_i-\boldsymbol{\mu})^{\mathrm{T}}\}$. In practice, however, the initial assumption of a known $\boldsymbol{\mu}$ is often unrealistic. Instead, we suggest to estimate $\mathbf{\Sigma}$ using the following U-type statistic:

$$\widehat{\mathbf{\Sigma}}_U(\tau) = \frac{1}{\binom{n}{2}} \sum_{1\leqslant i<j\leqslant n} \psi_\tau\bigg\{\frac{1}{2}(\boldsymbol{X}_i-\boldsymbol{X}_j)(\boldsymbol{X}_i-\boldsymbol{X}_j)^{\mathrm{T}}\bigg\}.$$

Observe that $(\boldsymbol{X}_i-\boldsymbol{X}_j)(\boldsymbol{X}_i-\boldsymbol{X}_j)^{\mathrm{T}}$ is a rank one matrix with eigenvalue $\|\boldsymbol{X}_i-\boldsymbol{X}_j\|_2^2$ and its associated eigenvector $(\boldsymbol{X}_i-\boldsymbol{X}_j)/\|\boldsymbol{X}_i-\boldsymbol{X}_j\|_2$. Therefore, by Definition 3.1, $\widehat{\mathbf{\Sigma}}_U(\tau)$ can be equivalently written as

$$\frac{1}{\binom{n}{2}} \sum_{1\leqslant i<j\leqslant n} \psi_\tau\bigg(\frac{1}{2}\|\boldsymbol{X}_i-\boldsymbol{X}_j\|_2^2\bigg)\frac{(\boldsymbol{X}_i-\boldsymbol{X}_j)(\boldsymbol{X}_i-\boldsymbol{X}_j)^{\mathrm{T}}}{\|\boldsymbol{X}_i-\boldsymbol{X}_j\|_2^2}.$$

This alternative expression makes it much easier to compute. The following theorem provides an exponential-type deviation inequality for $\widehat{\mathbf{\Sigma}}_U(\tau)$, which represents a useful complement to the results in Minsker (2016). See, for example, Remark 8 therein.

Theorem 3.2. Let

$$v^2 := \frac{1}{2}\Big\|\mathbb{E}\big\{(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^{\mathrm{T}}\big\}^2 + \mathrm{tr}(\mathbf{\Sigma})\mathbf{\Sigma} + 2\mathbf{\Sigma}^2\Big\|. \tag{3.4}$$

For any $t > 0$, the estimator $\widehat{\mathbf{\Sigma}}_U = \widehat{\mathbf{\Sigma}}_U(\tau)$ with $\tau \geqslant (v/2)(n/t)^{1/2}$ satisfies

$$\mathbb{P}\big\{\|\widehat{\mathbf{\Sigma}}_U - \mathbf{\Sigma}\| \geqslant 4v(t/n)^{1/2}\big\} \leqslant 2p\exp(-t).$$

Given $\widehat{\mathbf{\Sigma}}_U$, we can construct an estimator of $\mathbf{B}$ following Step 1 of the FARM-Test algorithm. Recall that $\widehat{\boldsymbol{b}}_1, \ldots, \widehat{\boldsymbol{b}}_p$ are the p rows of $\widehat{\mathbf{B}}$. To investigate the consistency of $\widehat{\boldsymbol{b}}_j$'s, let $\overline{\lambda}_1, \ldots, \overline{\lambda}_K$ be the top K (nonzero) eigenvalues of $\mathbf{B}\mathbf{B}^{\mathrm{T}}$ in a descending order and $\overline{\boldsymbol{v}}_1, \ldots, \overline{\boldsymbol{v}}_K$ be their corresponding eigenvectors. Under the identifiability condition (2.2), we have $\overline{\lambda}_\ell = \|\overline{\boldsymbol{b}}_\ell\|_2^2$ and $\overline{\boldsymbol{v}}_\ell = \overline{\boldsymbol{b}}_\ell / \|\overline{\boldsymbol{b}}_\ell\|_2$ for $\ell = 1, \ldots, K$.

Assumption 2 (Pervasiveness)**.** There exist positive constants c_1, c_2 and c_3 such that $c_1 p \leqslant \overline{\lambda}_\ell - \overline{\lambda}_{\ell+1} \leqslant c_2 p$ for $\ell = 1, \ldots, K$ with $\overline{\lambda}_{K+1} := 0$, and $\|\boldsymbol{\Sigma}_\varepsilon\| \leqslant c_3 < \overline{\lambda}_K$.

Remark 2. The pervasiveness condition is required for high dimensional spiked covariance model with the first several eigenvalues well separated and significantly larger than the rest. In particular, Assumption 2 requires the top K eigenvalues grow linearly with the dimension p. Therefore, the corresponding eigenvectors can be consistently estimated so long as sample size diverges (Fan *et al.*, 2013). This condition is widely assumed in existing literature (Stock and Watson, 2002; Bai and Ng, 2002). The following proposition provides convergence rate of the robust estimators $\{\widehat{\lambda}_\ell, \widehat{\boldsymbol{v}}_\ell\}_{\ell=1}^K$ under Assumption 2. The proof is based on Weyl's inequality and a useful variant of the Davis-Kahan theorem (Yu *et al.*, 2015), and is given in the supplemental material. We notice that some preceding works (Onatski, 2012; Shen *et al.*, 2016; Wang and Fan, 2017) have provided similar results under a weaker pervasiveness assumption which allows $p/n \to \infty$ in any manner and the spiked eigenvalues $\{\overline{\lambda}_\ell\}_{\ell=1}^K$ are allowed to grow slower than p so long as $c_\ell = p/(n\overline{\lambda}_\ell)$ is bounded.

Proposition 3.1. Under Assumption 2, we have

$$\max_{1 \leqslant \ell \leqslant K} |\widehat{\lambda}_\ell - \overline{\lambda}_\ell| \leqslant \|\widehat{\boldsymbol{\Sigma}}_U - \boldsymbol{\Sigma}\| + \|\boldsymbol{\Sigma}_\varepsilon\| \quad \text{and} \tag{3.5}$$

$$\max_{1 \leqslant \ell \leqslant K} \|\widehat{\boldsymbol{v}}_\ell - \overline{\boldsymbol{v}}_\ell\|_2 \leqslant C p^{-1} \big(\|\widehat{\boldsymbol{\Sigma}}_U - \boldsymbol{\Sigma}\| + \|\boldsymbol{\Sigma}_\varepsilon\| \big), \tag{3.6}$$

where $C > 0$ is a constant independent of (n, p).

We now show the properties of estimated loading vectors and the residual variance defined in (2.8).

Theorem 3.3. Suppose Assumption 1(iv) and Assumption 2 hold. Let $\tau = v_0 \omega_{n,p}$ with

$v_0 \geqslant v/2$ for v given in (3.4). Then, with probability at least $1 - 2n^{-1}$,

$$\max_{1 \leqslant j \leqslant p} \|\widehat{\boldsymbol{b}}_j - \boldsymbol{b}_j\|_2 \leqslant C_1 \big\{ v \sqrt{\log(np)}\, (np)^{-1/2} + p^{-1/2} \big\} \tag{3.7}$$

as long as $n \geqslant v^2 p^{-1} \log(np)$. In addition, if $n \geqslant C_2 \log(np)$, $\tau_j = a_j \omega_{n,p}, \tau_{jj} = a_{jj} \omega_{n,p}$ with $a_j \geqslant \sigma_{jj}^{1/2}, a_{jj} \geqslant \mathrm{var}(X_j^2)^{1/2}$, we have

$$\max_{1 \leqslant j \leqslant p} |\widehat{\sigma}_{\varepsilon,jj} - \sigma_{\varepsilon,jj}| \leqslant C_3 \big(v p^{-1/2} w_{n,p}^{-1} + p^{-1/2} \big) \tag{3.8}$$

with probability greater than $1 - C_4 n^{-1}$. Here, C_1–C_4 are positive constants that are independent of (n, p).

Remark 3. According to Theorem 3.3, the robustification parameters can be set as $\tau_j = a_j \omega_{n,p}$ and $\tau_{jj} = a_{jj} \omega_{n,p}$, where $w_{n,p}$ is given in (3.3). In practice, the constants a_j and a_{jj} can be chosen by cross-validation.

3.2.2 Adaptive Huber covariance estimation

In this section, we adopt an estimator that was first considered in Fan *et al.* (2017). For every $1 \leqslant j \neq k \leqslant p$, we define the robust estimate $\widehat{\sigma}_{jk}$ of $\sigma_{jk} = \mathbb{E}(X_j X_k) - \mu_j \mu_k$ to be

$$\widehat{\sigma}_{jk} = \widehat{\theta}_{jk} - \widehat{\mu}_j \widehat{\mu}_k \quad \text{with} \quad \widehat{\theta}_{jk} = \operatorname*{argmin}_{\theta \in \mathbb{R}} \sum_{i=1}^{n} \ell_{\tau_{jk}}(X_{ij} X_{ik} - \theta), \tag{3.9}$$

where $\tau_{jk} > 0$ is a robustification parameter and $\widehat{\mu}_j$ is defined in (2.5). This yields the adaptive Huber covariance estimator $\widehat{\boldsymbol{\Sigma}}_{\mathrm{H}} = (\widehat{\sigma}_{jk})_{1 \leqslant j,k \leqslant p}$. The dependence of $\widehat{\boldsymbol{\Sigma}}_{\mathrm{H}}$ on $\{\tau_{jk} : 1 \leqslant j \leqslant k \leqslant p\}$ and $\{\tau_j\}_{j=1}^{p}$ is assumed without displaying.

Theorem 3.4. Suppose Assumption 1 (iv) and Assumption 2 hold. Let $\tau_j = a_j \omega_{n,p}, \tau_{jk} = a_{jk} \omega_{n,p^2}$ with $a_j \geqslant \sigma_{jj}^{1/2}, a_{jk} \geqslant \mathrm{var}(X_j^2)^{1/2}$ for $1 \leqslant j, k \leqslant p$. Then, there exist positive constants C_1–C_3 independent of (n, p) such that as long as $n \geqslant C_1 \log(np)$,

$$\max_{1 \leqslant j \leqslant p} \|\widehat{\boldsymbol{b}}_j - \boldsymbol{b}_j\|_2 \leqslant C_2 \big(\omega_{n,p}^{-1} + p^{-1/2} \big)$$

$$\text{and} \quad \max_{1 \leqslant j \leqslant p} |\widehat{\sigma}_{\varepsilon,jj} - \sigma_{\varepsilon,jj}| \leqslant C_3 \big(\omega_{n,p}^{-1} + p^{-1/2} \big)$$

with probability greater than $1-4n^{-1}$, where $w_{n,p}$ is given in (3.3).

3.3 Estimating realized factors

To make the oracle statistics T_j°'s given in (3.1) usable, we still need to estimate $\bar{\boldsymbol{f}}$. Since the loadings can be estimated in two different ways, let us first assume $\mathbf{B}$ is given and treat it as an input variable.

Averaging the approximate factor model (2.1), we have $\bar{\boldsymbol{X}} = \boldsymbol{\mu} + \mathbf{B}\bar{\boldsymbol{f}} + \bar{\varepsilon}$, where $\bar{\boldsymbol{X}} = (\bar{X}_1, \ldots, \bar{X}_p)^{\mathrm{T}} = n^{-1}\sum_{i=1}^n \boldsymbol{X}_i$ and $\bar{\boldsymbol{\varepsilon}} := (\bar{\varepsilon}_1, \ldots, \bar{\varepsilon}_p) = n^{-1}\sum_{i=1}^n \boldsymbol{\varepsilon}_i$. This leads to

$$\bar{X}_j = \boldsymbol{b}_j^{\mathrm{T}}\bar{\boldsymbol{f}} + \mu_j + \bar{\varepsilon}_j, \; j = 1, \ldots, p. \tag{3.10}$$

Among all $\{\mu_j + \bar{\varepsilon}_j\}_{j=1}^p$, we may consider $\mu_j + \bar{\varepsilon}_j$ with $\mu_j \neq 0$ as outliers. Therefore, to achieve robustness, we estimate $\bar{\boldsymbol{f}}$ by

$$\widehat{\boldsymbol{f}}(\mathbf{B}) = \operatorname*{argmin}_{\boldsymbol{f}\in\mathbb{R}^K} \sum_{j=1}^p \ell_\gamma(\bar{X}_j - \boldsymbol{b}_j^{\mathrm{T}}\boldsymbol{f}), \tag{3.11}$$

where $\gamma = \gamma(n,p) > 0$ is a robustification parameter. Next, we define robust variance estimators $\widehat{\sigma}_{\varepsilon,jj}$'s by

$$\widehat{\sigma}_{\varepsilon,jj}(\mathbf{B}) = \widehat{\theta}_j - \widehat{\mu}_j^2 - \|\boldsymbol{b}_j\|_2^2 \;\text{ with }\; \widehat{\theta}_j = \operatorname*{argmin}_{\theta > \widehat{\mu}_j^2 + \|\boldsymbol{b}_j\|_2^2} \sum_{i=1}^n \ell_{\tau_{jj}}(X_{ij}^2 - \theta), \tag{3.12}$$

where τ_{jj}'s are robustification parameters and $\widehat{\mu}_j$'s are as in (2.5). Plugging $\{\widehat{\sigma}_{\varepsilon,jj}\}_{j=1}^p$ and $\widehat{\boldsymbol{f}}$ into (3.1), we obtain the following factor-adjusted test statistics

$$T_j(\mathbf{B}) = \left\{\frac{n}{\widehat{\sigma}_{\varepsilon,jj}(\mathbf{B})}\right\}^{1/2} \{\widehat{\mu}_j - \boldsymbol{b}_j^{\mathrm{T}}\widehat{\boldsymbol{f}}(\mathbf{B})\}, \quad j = 1, \ldots, p. \tag{3.13}$$

For a given threshold $z \geqslant 0$, the corresponding FDP is defined as

$$\mathrm{FDP}(z;\mathbf{B}) = V(z;\mathbf{B})/R(z;\mathbf{B}), \tag{3.14}$$

where $V(z;\mathbf{B}) = \sum_{j\in\mathcal{H}_0} 1\{|T_j(\mathbf{B})| \geqslant z\}$ and $R(z;\mathbf{B}) = \sum_{1\leqslant j\leqslant p} 1\{|T_j(\mathbf{B})| \geqslant z\}$. Similarly to (3.2), we approximate $\mathrm{FDP}(z;\mathbf{B})$ by

$$\mathrm{AFDP}(z;\mathbf{B}) = 2p_0\Phi(-z)/R(z;\mathbf{B}),\ z \geqslant 0. \tag{3.15}$$

Regarding the accuracy of $\mathrm{AFDP}(z;\mathbf{B})$ as an asymptotic approximation of $\mathrm{FDP}(z;\mathbf{B})$, we need to account for the statistical errors of $\{\widehat{\sigma}_{\varepsilon,jj}(\mathbf{B})\}_{j=1}^p$ and $\widehat{\boldsymbol{f}}(\mathbf{B})$. To this end, we make the following structural assumptions on $\boldsymbol{\mu}$ and $\mathbf{B}$.

Assumption 3. $\varepsilon_1,\ldots,\varepsilon_p$ are independent, and there exist constants $c_l, c_u > 0$ such that $\lambda_{\min}(p^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{B}) \geqslant c_l$ and $\|\mathbf{B}\|_{\max} \leqslant c_u$.

Assumption 4 (Sparsity)**.** There exist constants $C_\mu > 0$ and $c_\mu \in (0, 1/2)$ such that $\|\boldsymbol{\mu}\|_\infty = \max_{1\leqslant j\leqslant p}|\mu_j| \leqslant C_\mu$ and $\|\boldsymbol{\mu}\|_0 = \sum_{j=1}^p 1(\mu_j \neq 0) \leqslant p^{1/2-c_\mu}$. Moreover, (n,p) satisfies that $n\log(n) = o(p)$ as $n,p\to\infty$.

The following proposition, which is of independent interest, reveals an exponential-type deviation inequality for $\widehat{\boldsymbol{f}}(\mathbf{B})$ with a properly chosen $\gamma > 0$.

Proposition 3.2. Suppose that Assumption 3 holds. For any $t > 0$, the estimator $\widehat{\boldsymbol{f}}(\mathbf{B})$ given in (3.11) with $\gamma = \gamma_0(p/t)^{1/2}$ for $\gamma_0 \geqslant \overline{\sigma}_\varepsilon := (p^{-1}\sum_{j=1}^p \sigma_{\varepsilon,jj})^{1/2}$ satisfies that with probability greater than $1-(2eK+1)e^{-t}$,

$$\|\widehat{\boldsymbol{f}}(\mathbf{B}) - \bar{\boldsymbol{f}}\|_2 \leqslant C_1\gamma_0(Kt)^{1/2}p^{-1/2} \tag{3.16}$$

as long as $p \geqslant \max\{\|\boldsymbol{\mu}\|_2^2/\overline{\sigma}_\varepsilon^2, (\|\boldsymbol{\mu}\|_1/\overline{\sigma}_\varepsilon)^2 t, C_2K^2t\}$, where $C_1, C_2 > 0$ are constants depending only on c_l, c_u in Assumption 3.

The convergence in probability of $\mathrm{FDP}(z;\mathbf{B})$ to $\mathrm{AFDP}(z;\mathbf{B})$ for any $z \geqslant 0$ is investigated in the following theorem.

Theorem 3.5. Suppose that Assumptions 1(i)–(iv), Assumptions 3 and 4 hold. Let $\tau_j = a_j\omega_{n,p}$, $\tau_{jj} = a_{jj}\omega_{n,p}$ with $a_j \geqslant \sigma_{jj}^{1/2}$, $a_{jj} \geqslant \mathrm{var}(X_j^2)^{1/2}$ for $j = 1,\ldots,p$, and $\gamma = \gamma_0\{p/\log(n)\}^{1/2}$ with $\gamma_0 \geqslant \overline{\sigma}_\varepsilon$. Then, for any $z \geqslant 0$, on the event $\{p^{-1}R(z;\mathbf{B}) \geqslant c\}$ for some $c > 0$, we have

$$\big|\mathrm{AFDP}(z; \mathbf{B}) - \mathrm{FDP}(z; \mathbf{B})\big| = o_{\mathbb{P}}(1) \quad \text{as } n, p \to \infty. \tag{3.17}$$

3.4 Sample splitting

Note that, our procedure described in Sections 3.2 and 3.3 consists of two parts, the calibration of a factor model (i.e., estimating $\mathbf{B}$ in (2.1)) and multiple inference. The construction of the test statistics, or equivalently, the P-values, relies on a "fine" estimate of $\bar{\boldsymbol{f}}$ based on the linear model in (3.10). In practice, $\boldsymbol{b}_j$'s are replaced by the fitted loadings $\widehat{\boldsymbol{b}}_j$'s using the methods in Section 3.2.

To avoid mathematical challenges caused by the reuse of the sample, we resort to the simple idea of sample splitting (Hartigan, 1969; Cox, 1975): half the data are used for calibrating a factor model and the other half are used for multiple inference. We refer to Rinaldo et al. (2016) for a modern look at inference based on sample splitting. Specifically, the steps are summarized below.

(1) Split the data $\mathcal{X} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ into two halves $\mathcal{X}_1$ and $\mathcal{X}_2$. For simplicity, we assume that the data are divided into two groups of equal size $m = n/2$.

(2) We use $\mathcal{X}_1$ to estimate $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_p$ using either the U-type method (Section 3.2.1) or the adaptive Huber method (Section 3.2.2). For simplicity, we focus on the latter and denote the estimators by $\widehat{\boldsymbol{b}}_1(\mathcal{X}_1), \ldots, \widehat{\boldsymbol{b}}_p(\mathcal{X}_1)$.

(3) Proceed with the remain steps in the FARM-Test algorithm using the data in $\mathcal{X}_2$. Denote the resulting test statistics by $T_1, \ldots, T_p$. For a given threshold $z \geqslant 0$, the corresponding FDP and its asymptotic expression are defined as

$$\mathrm{FDP}_{\mathrm{sp}}(z) = V(z)/R(z) \quad \text{and} \quad \mathrm{AFDP}_{\mathrm{sp}}(z) = 2p\Phi(-z)/R(z),$$

respectively, where $V(z) = \sum_{j \in \mathcal{H}_0} 1(|T_j| \geqslant z)$, $R(z) = \sum_{1 \leqslant j \leqslant p} 1(|T_j| \geqslant z)$ and the subscript "sp" stands for sample splitting.

The purpose of sample splitting employed in the above procedure is to facilitate the theoretical analysis. The following result shows that the asymptotic FDP $\mathrm{AFDP}_{\mathrm{sp}}(z)$, constructed via sample splitting, provides a consistent estimate of $\mathrm{FDP}(z)$.

Theorem 3.6. Suppose that Assumptions 1(i)–(iv), Assumptions 2–4 hold. Let $\tau_j = a_j\omega_{n,p}$, $\tau_{jj} = a_{jj}\omega_{n,p}$ with $a_j \geqslant \sigma_{jj}^{1/2}$, $a_{jj} \geqslant \mathrm{var}(X_j^2)^{1/2}$ for $j = 1,\ldots,p$, and let $\gamma = \gamma_0\{p/\log(np)\}^{1/2}$ with $\gamma_0 \geqslant \overline{\sigma}_\varepsilon$. Then, for any $z \geqslant 0$, on the event $\{p^{-1}R(z) \geqslant c\}$ for some $c > 0$, we have

$$\left|\mathrm{AFDP}_{\mathrm{sp}}(z) - \mathrm{FDP}_{\mathrm{sp}}(z)\right| = o_{\mathbb{P}}(1) \quad \text{as } n, p \to \infty.$$

4 Simulation studies

4.1 Settings

In the simulation studies, we take $(p_1, p) = (25, 500)$ so that $\pi_1 = p_1/p = 0.05$, $n \in \{100, 150, 200\}$ and use $t = 0.01$ as the threshold value for P-values. Moreover, we set the mean vector $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_p)^{\mathrm{T}}$ to be $\mu_j = 0.5$ for $1 \leqslant j \leqslant 25$ and $\mu_j = 0$ otherwise. We repeat 1000 replications in each of the scenarios below. The robustificatons parameters are selected by five-fold cross-validation under the guidance of their theoretically optimal orders. The data-generating processes are as follows.

Model 1: Normal factor model. Consider a three-factor model $\boldsymbol{X}_i = \boldsymbol{\mu} + \mathbf{B}\boldsymbol{f}_i + \boldsymbol{\varepsilon}_i$, $i = 1,\ldots,n$, where $\boldsymbol{f}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_3)$, $\mathbf{B} = (b_{j\ell})_{1\leqslant j\leqslant p, 1\leqslant \ell\leqslant 3}$ has i.i.d. entries $b_{j\ell}$'s generated from the uniform distribution $\mathcal{U}(-2, 2)$.

Model 2: Synthetic factor model. Consider a similar three-factor model as in Model 1, except now $\{\boldsymbol{f}_i\}_{i=1}^n$ and $\{b_j\}_{j=1}^p$ are generated independently from $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_f)$ and $\mathcal{N}(\boldsymbol{\mu}_B, \boldsymbol{\Sigma}_B)$, respectively. The parameters $\boldsymbol{\Sigma}_f$, $\boldsymbol{\mu}_B$ and $\boldsymbol{\Sigma}_B$ are calibrated from the daily returns of S&P 500's top 100 constituents (ranked by the market cap) between July 1st, 2008 and June 29th, 2012.

Model 3: Serial dependent factor model. Consider a similar three-factor model as in Model 1, except the factors $\boldsymbol{f}_i$'s are now generated from a stationary VAR(1) model $\boldsymbol{f}_i = \boldsymbol{\Pi}\boldsymbol{f}_{i-1} + \boldsymbol{\xi}_i$ for $i = 1,\ldots,n$, with $\boldsymbol{f}_0 = \mathbf{0}$ and $\boldsymbol{\xi}_i$'s i.i.d. drawn from $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_3)$. The

(j,k)-th entry of $\mathbf{\Pi}$ is set to be 0.5 when $j = k$ and $0.4^{|j-k|}$ otherwise.

Error distributions. In all the three models above, the idiosyncratic errors are generated from one of the following four distributions. Let $\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}$ be a sparse matrix whose diagonal entries are 3 and off diagonal entries are drawn from i.i.d. $0.3 \times Bernoulli(0.05)$;

(1) Multivariate Normal distribution $\mathcal{N}(0, \mathbf{\Sigma}_{\boldsymbol{\varepsilon}})$;

(2) Multivariate t-distribution $t_3(0, \mathbf{\Sigma}_{\boldsymbol{\varepsilon}})$ with 3 degrees of freedom;

(3) i.i.d. Gamma distribution with shape parameter 3 and scale parameter 1;

(4) i.i.d. Re-scaled Log-Normal distribution $a\{\exp(1+1.2Z)-b\}$, where $Z \sim \mathcal{N}(0,1)$ and $a, b > 0$ are chosen such that it has mean zero and variance 3.

4.2 Comparing methods of estimating FDP

In our robust proposal, the covariance matrix is either estimated by the entry-wise adaptive Huber method or by the U-type robust covariance estimator. The corresponding test procedures are denoted by **FARM-H** and **FARM-U**, respectively.

In this subsection, we compare **FARM-H** and **FARM-U** with three existing non-robust test procedures. The first non-robust alternative is a factor-adjusted procedure using the sample mean and sample covariance matrix, denoted by **FAM**. The second one is the **PFA** method, short for principal factor approximation, proposed by Fan and Han (2017). In contrast to **FAM**, **PFA** method only accounts for the effect of the latent factors, but not removing the effect of latent factors. The third non-robust procedure is the **Naive** method, which ignores completely the factor dependence.

The performance of FDP estimation is assessed by the empirical mean absolute error between the estimated FDP, using either the entry-wise adaptive Huber estimator or the U-type robust covariance estimator, and the oracle FDP defined in (3.2). The results are presented in Table 1. We see that, although the **PFA** and **FAM** methods achieve the smallest estimation errors in the normal case, both **FARM-H** and **FARM-U** perform comparably well. The **Naive** method performs worst as it ignores the impact of the latent factors. In heavy-tailed cases, both **FARM-H** and **FARM-U** outperform the non-robust competitors by a wide margin, still with the **Naive** method being the least favorable. In summary, the proposed methods achieve high degree of robustness against heavy-tailed errors, while almost lose no efficiency under normality.

Table 1: Empirical mean absolute error between estimated and oracle FDP

			$p = 500$				
	ε_i	n	**FARM-H**	**FARM-U**	**FAM**	**PFA**	**Naive**
Model 1	Normal	100	0.0654	0.0662	0.0628	0.0635	0.1399
		150	0.0628	0.0637	0.0593	0.0615	0.1216
		200	0.0604	0.0609	0.0580	0.0585	0.1162
	t_3	100	0.0820	0.0852	0.1463	0.1524	0.2166
		150	0.0777	0.0715	0.1231	0.1315	0.1903
		200	0.0751	0.0623	0.1046	0.1176	0.1682
	Gamma	100	0.0848	0.0864	0.1629	0.1821	0.3468
		150	0.0802	0.0809	0.1553	0.1637	0.2970
		200	0.0750	0.0757	0.1457	0.1522	0.2169
	LN	100	0.0987	0.0993	0.1973	0.2152	0.3987
		150	0.0831	0.0861	0.1814	0.1988	0.3391
		200	0.0797	0.0810	0.1620	0.1669	0.2811
Model 2	Normal	100	0.0679	0.0688	0.0635	0.0639	0.1528
		150	0.0664	0.0664	0.0609	0.0611	0.1430
		200	0.0640	0.0648	0.0588	0.0594	0.1371
	t_3	100	0.0858	0.0880	0.1640	0.1719	0.2570
		150	0.0820	0.0827	0.1451	0.1520	0.2182
		200	0.0789	0.0798	0.1232	0.1283	0.1755
	Gamma	100	0.0847	0.0864	0.1952	0.2041	0.3622
		150	0.0820	0.0809	0.1598	0.1697	0.3046
		200	0.0765	0.0772	0.1307	0.1376	0.2378
	LN	100	0.1042	0.1077	0.2468	0.2528	0.4475
		150	0.0870	0.0882	0.2159	0.2284	0.3632
		200	0.0804	0.0812	0.1961	0.2154	0.2997
Model 3	Normal	100	0.0681	0.0695	0.0651	0.0673	0.1623
		150	0.0637	0.0650	0.0612	0.0634	0.1470
		200	0.0613	0.0615	0.0599	0.0608	0.1392
	t_3	100	0.0904	0.0910	0.1584	0.1761	0.2383
		150	0.0815	0.0816	0.1396	0.1428	0.2098
		200	0.0768	0.0788	0.1252	0.1245	0.1957
	Gamma	100	0.0904	0.0908	0.1935	0.2173	0.4193
		150	0.0822	0.0829	0.1662	0.1850	0.3193
		200	0.0787	0.0790	0.1574	0.1585	0.2460
	LN	100	0.0998	0.1022	0.2133	0.2286	0.4439
		150	0.0895	0.0918	0.1976	0.2051	0.3714
		200	0.0820	0.0842	0.1806	0.1827	0.3274

4.3 Empirical power

In this section, we compare the powers of five methods under consideration. The empirical power is defined as the average ratio between the number of correct rejections and p_1. The results are displayed in Table 2. In the normal case, **FAM** has a higher power than **PFA**. This is because **FAM** adjusts the effect of latent factors for each individual hypothesis so that the signal-to-noise ratio is higher.

Once again, both the **FARM-H** and **FARM-U** methods pay only a negligible price in power under normality. In heavy-tailed cases, however, these two robust methods have much higher empirical powers than their non-robust counterparts. We further investigate the relationship between the empirical power and signal strength. In Model 1 with $(n,p) = (200,500)$ and t_3-distributed errors, we allow μ_j to vary from 0.1 to 0.8 and draw the empirical power of each method with respect to signal strength; see Figure 3.

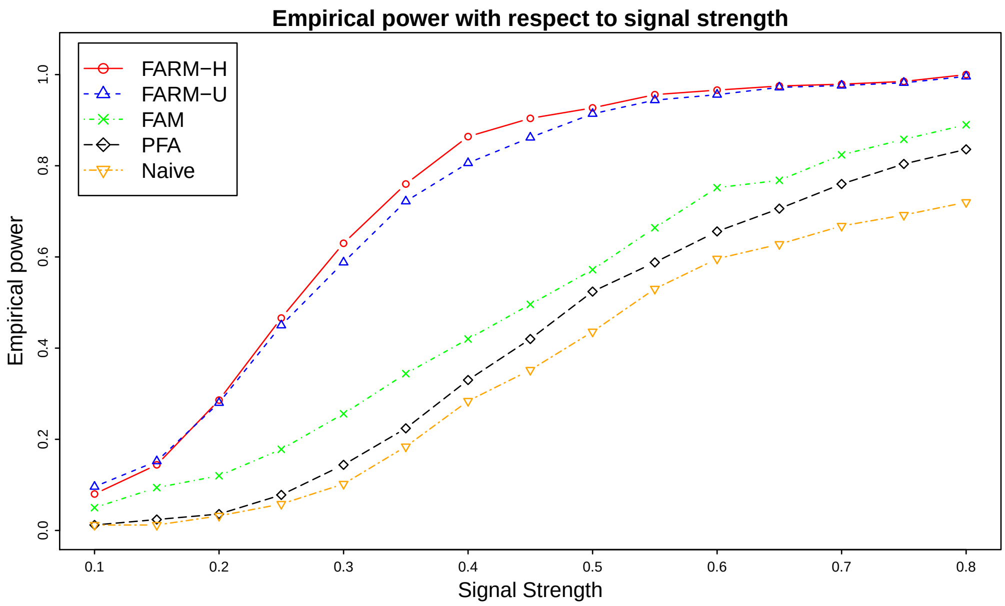

Figure 3: Empirical power versus signal strength, which is non-zero μ_j that ranges from 0.1 to 0.8. The data are generated from Model 1 with $(n,p) = (200,500)$ and t_3-distributed noise.

5 Real data analysis

Oberthuer et al. (2006) analyzed the German Neuroblastoma Trials NB90-NB2004 (diag-

Table 2: Empirical powers

	ε_i	n	$p = 500$ **FARM-H**	**FARM-U**	**FAM**	**PFA**	**Naive**
Model 1	Normal	100	0.853	0.849	0.872	0.863	0.585
		150	0.877	0.870	0.890	0.882	0.624
		200	0.909	0.907	0.924	0.915	0.671
	t_3	100	0.816	0.815	0.630	0.610	0.442
		150	0.828	0.826	0.668	0.657	0.464
		200	0.894	0.870	0.702	0.691	0.502
	Gamma	100	0.816	0.813	0.658	0.639	0.281
		150	0.830	0.825	0.684	0.663	0.391
		200	0.889	0.873	0.712	0.707	0.433
	LN	100	0.798	0.786	0.566	0.534	0.242
		150	0.817	0.805	0.587	0.673	0.292
		200	0.844	0.835	0.613	0.605	0.369
Model 2	Normal	100	0.801	0.799	0.864	0.855	0.584
		150	0.856	0.846	0.880	0.870	0.621
		200	0.904	0.900	0.911	0.904	0.659
	t_3	100	0.810	0.802	0.612	0.601	0.402
		150	0.825	0.814	0.638	0.632	0.457
		200	0.873	0.859	0.695	0.683	0.484
	Gamma	100	0.804	0.798	0.527	0.509	0.216
		150	0.821	0.819	0.594	0.557	0.289
		200	0.885	0.875	0.638	0.606	0.379
	LN	100	0.763	0.757	0.463	0.434	0.206
		150	0.799	0.795	0.495	0.479	0.228
		200	0.826	0.819	0.529	0.511	0.312
Model 3	Normal	100	0.837	0.832	0.848	0.833	0.535
		150	0.856	0.848	0.864	0.857	0.594
		200	0.875	0.871	0.902	0.896	0.628
	t_3	100	0.801	0.796	0.606	0.591	0.403
		150	0.818	0.816	0.640	0.612	0.426
		200	0.881	0.872	0.675	0.643	0.501
	Gamma	100	0.792	0.785	0.385	0.329	0.205
		150	0.818	0.809	0.472	0.435	0.281
		200	0.874	0.867	0.581	0.565	0.367
	LN	100	0.783	0.776	0.355	0.336	0.187
		150	0.804	0.795	0.442	0.406	0.231
		200	0.859	0.849	0.514	0.487	0.326

nosed between 1989 and 2004) and developed a gene expression based classifier. For 246 neuroblastoma patients, gene expressions over 10,707 probe sites were measured. The binary response variable is the 3-year event-free survival information of the patients (56 positive and 190 negative). We refer to Oberthuer et al. (2006) for a detailed description of the dataset.

In this study, we divide the data into two groups, one with positive responses and the other with negative responses, and test the equality of gene expression levels at all the 10,707 probe sites simultaneously. To that end, we generalize the proposed FARM-Test to the two-sample case by defining the following two-sample t-type statistic

$$T_j = \frac{(\widehat{\mu}_{1j} - \widehat{\boldsymbol{b}}_{1j}^{\mathrm{T}} \widehat{\boldsymbol{f}}_1) - (\widehat{\mu}_{2j} - \widehat{\boldsymbol{b}}_{2j}^{\mathrm{T}} \widehat{\boldsymbol{f}}_2)}{(\widehat{\sigma}_{1\varepsilon,jj}/56 + \widehat{\sigma}_{2\varepsilon,jj}/190)^{1/2}}, \quad j = 1, \ldots, 10707,$$

where the subscripts 1 and 2 correspond to the positive and negative groups, respectively. Specifically, $\widehat{\mu}_{1j}$ and $\widehat{\mu}_{2j}$ are the robust mean estimators obtained from minimizing the empirical Huber risk (2.5), and $\widehat{\boldsymbol{b}}_{1j}$, $\widehat{\boldsymbol{b}}_{2j}$, $\widehat{\boldsymbol{f}}_1$ and $\widehat{\boldsymbol{f}}_2$ are robust estimators of the factors and loadings based on the U-type covariance estimator. In addition, $\widehat{\sigma}_{1\varepsilon,jj}$ and $\widehat{\sigma}_{2\varepsilon,jj}$ are the variance estimators defined in (3.13). Once again, the robustificatons parameters are selected via five-fold cross-validation with their theoretically optimal orders taking into account.

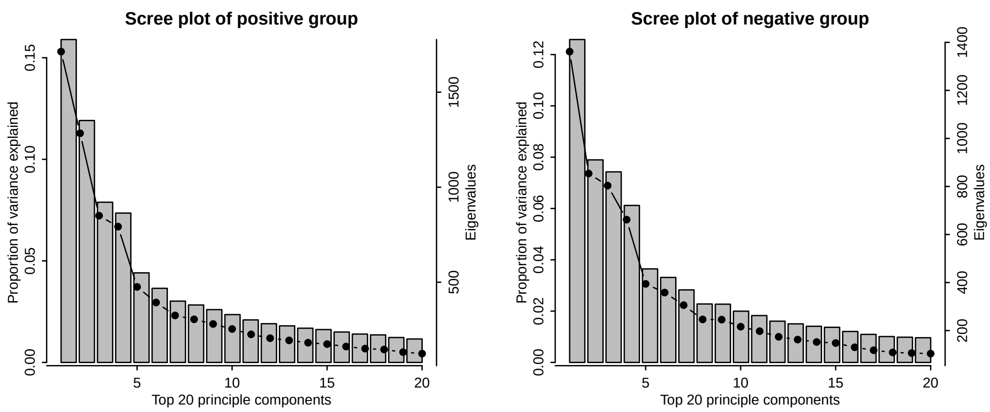

Figure 4: **Scree plots for positive and negative groups**. The bars represent the proportion of variance explained by the top 20 principal components. The dots represent the corresponding eigenvalues in descending order.

We use the eigenvalue ratio method (Lam and Yao, 2012; Ahn and Horenstein, 2013) to determine the number of factors. Let $\lambda_k(\widehat{\boldsymbol{\Sigma}})$ be the k-th largest eigenvalue of $\widehat{\boldsymbol{\Sigma}}$ and $K_{\max}$ a prespecified upper bound. The number of factors can then be estimated by

$$\widehat{K} = \underset{1 \leqslant k \leqslant K_{\max}}{\operatorname{argmax}} \; \lambda_k(\widehat{\boldsymbol{\Sigma}}) / \lambda_{k+1}(\widehat{\boldsymbol{\Sigma}}).$$

The eigenvalue ratio method suggests $K = 4$ for both positive and negative groups. Figure

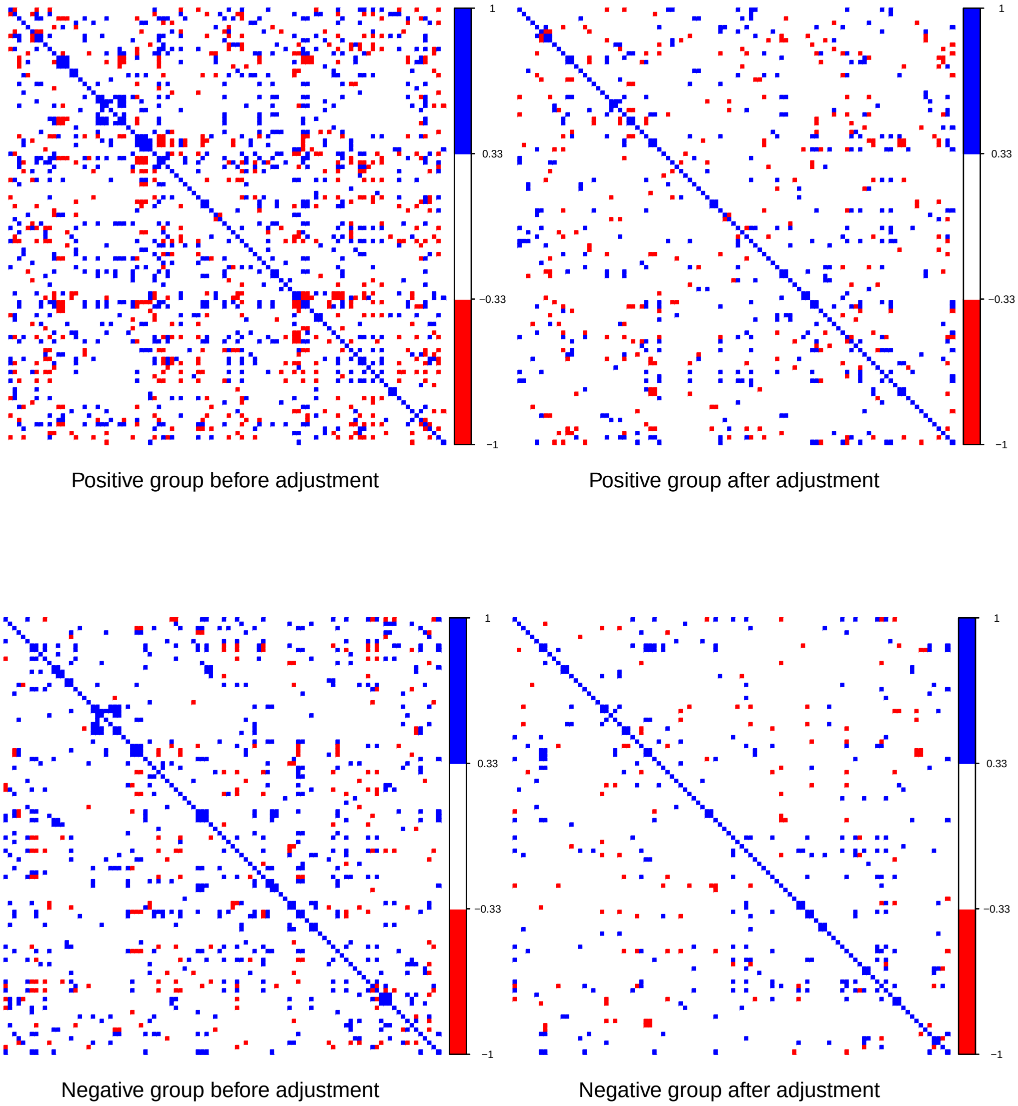

Figure 5: **Correlations among the first 100 genes before and after factor-adjustment**. The pixel plots are the correlation matrices of the first 100 gene expressions. In the plots, the blue pixels represent the entries with correlation greater than 1/3 and the red pixels represent the entries with correlation smaller than -1/3.

4 depicts scree plots of the top 20 eigenvalues for each group. The gene expressions in both groups are highly correlated. As an evidence, the top 4 principal components (PCs) explain 42.6% and 33.3% of the total variance for the positive and negative groups, respectively.

To justify the importance of the factor-adjustment procedure, for each group, we plot the correlation matrices of the first 100 gene expressions before and after adjusting the

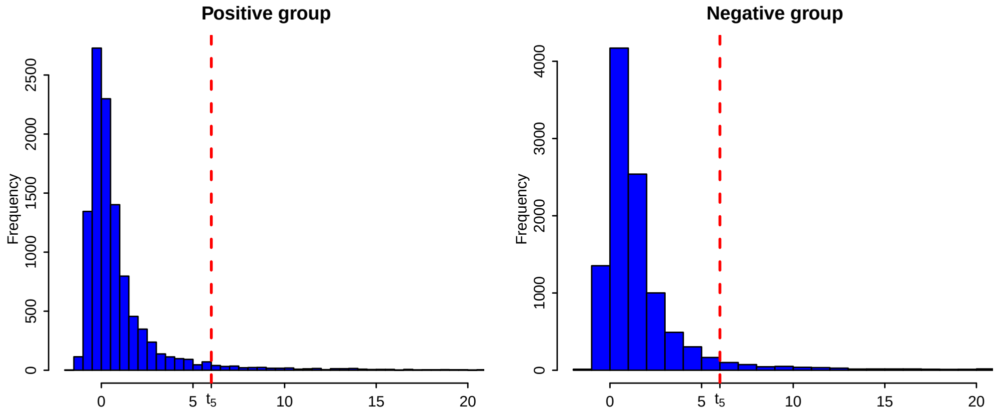

Figure 6: **Histogram of excess kurtosises for the gene expressions in positive and negative groups**. The dashed line at 6 is the excess kurtosis of t-distribution with d.f. 5.

top 4 PCs; see Figure 5. The blue and red pixels in Figure 5 represent the pairs of gene expressions whose absolute correlations are greater than 1/3. Therefore, after adjusting the top 4 PCs, the number of off-diagonal entries with strong correlations is significantly reduced in both groups. To be more specific, the number drops from 1452 to 666 for the positive group and from 848 to 414 for the negative group.

Another stylized feature of the data is that distributions of many gene expressions are heavy-tailed. To see this, we plot histograms of the excess kurtosis of the gene expressions in Figure 6. The left panel of the Figure 6 shows that 6518 gene expressions have positive excess kurtosis with 420 of them greater than 6. In other words, more than 60% of the gene expressions in the positive group have tails heavier than the normal distribution and about 4% are severely heavy tailed as their tails are fatter than the t-distribution with 5 degrees of freedom. Similarly, in the negative group, 9341 gene expressions exhibit positive excess kurtosis with 671 of them greater than 6. Such a heavy-tailed feature indicates the necessity of using robust methods to estimate the mean and covariance of the data.

We apply all the three methods, the two-sample FARM-Test (**FARM-U**), the **FAM**-Test and the naive method, to this dataset. Given threshold value fixed at $t = 0.05$, the **FARM-U** method identifies 2128 probes with different gene expressions, while the **FAM** and naive methods discover 1767 and 1131 probes, respectively. For this dataset, accounting for latent factor dependence indeed leads to different statistical conclusions. In particular, this visible discrepancy between **FARM-U** and **FAM** highlights the importance of robustness and reflects the difference of power in detecting differently expressed probes. The importance of factor adjustments is also highlighted in the discovery of significance genes.

6 Discussion and extensions

In this paper, we have developed a factor-adjusted multiple testing procedure (FARM-Test) for large-scale simultaneous inference with dependent and heavy-tailed data, the key of which lies in a robust estimate of the false discovery proportion. The procedure has two attractive features: First, it incorporates dependence information to construct marginal test statistics. Intuitively, subtracting common factors out leads to higher signal-to-noise ratios, and therefore makes the resulting FDP control procedure more efficient and powerful. Second, to achieve robustness against heavy-tailed errors that may also be asymmetric, we used the adaptive Huber regression method (Fan *et al.*, 2017; Sun *et al.*, 2017) to estimate the realized factors, factor loadings and variances. We believe that these two properties will have further applications to higher criticism for detecting sparse signals with dependent and non-Gaussian data; see Delaigle *et al.* (2011) for the independent case.

In other situations, it may be more instructive to consider the mixed effects regression modeling of the data (Friguet *et al.*, 2009; Wang *et al.*, 2015), that is, $X_j = \mu_j + \boldsymbol{\beta}_j^{\mathrm{T}} \boldsymbol{Z} + \boldsymbol{b}_j^{\mathrm{T}} \boldsymbol{f} + \varepsilon_j$ for $j = 1, \ldots, p$, where $\boldsymbol{Z} \in \mathbb{R}^q$ is a vector of explanatory variables (e.g., treatment-control, phenotype, health trait), $\boldsymbol{\beta}_j$'s are $q \times 1$ vectors of unknown slope coefficients, and $\boldsymbol{f}$, $\boldsymbol{b}_j$'s and ε_j's have the same meanings as in (2.1). Suppose we observe independent samples $(\boldsymbol{X}_1, \boldsymbol{Z}_1), \ldots, (\boldsymbol{X}_n, \boldsymbol{Z}_n)$ from $(\boldsymbol{X}, \boldsymbol{Z})$ satisfying

$$\boldsymbol{X}_i = \boldsymbol{\mu} + \boldsymbol{\Theta} \boldsymbol{Z}_i + \mathbf{B} \boldsymbol{f}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, n,$$

where $\boldsymbol{\Theta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p)^{\mathrm{T}} \in \mathbb{R}^{p \times q}$. In this case, we have $\mathbb{E}(\boldsymbol{X}_i | \boldsymbol{Z}_i) = \boldsymbol{\mu} + \boldsymbol{\Theta} \boldsymbol{Z}_i$ and $\mathrm{cov}(\boldsymbol{X}_i | \boldsymbol{Z}_i) = \mathbf{B} \boldsymbol{\Sigma}_f \mathbf{B}^{\mathrm{T}} + \boldsymbol{\Sigma}_\varepsilon$. The main issue in extending our methodology to such a mixed effects model is the estimation of $\boldsymbol{\Theta}$. For this, we construct robust estimators $(\widehat{\mu}_j, \widehat{\boldsymbol{\beta}}_j)$ of

$(\mu_j, \boldsymbol{\beta}_j)$, defined as

$$(\widehat{\mu}_j, \widehat{\boldsymbol{\beta}}_j) \in \underset{\mu \in \mathbb{R},\, \boldsymbol{\beta}_j \in \mathbb{R}^q}{\operatorname{argmin}} \sum_{i=1}^{n} \ell_{\tau_j}(X_{ij} - \mu - \boldsymbol{\beta}_j^{\mathrm{T}} \boldsymbol{Z}_i), \quad 1 \leqslant j \leqslant p,$$

where τ_j's are robustification parameters. Taking $\widehat{\boldsymbol{\Theta}} = (\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_p)^{\mathrm{T}}$, the FARM-Test procedure in Section 2.2 can be directly applied with $\{\boldsymbol{X}_i\}_{i=1}^n$ replaced by $\{\boldsymbol{X}_i - \widehat{\boldsymbol{\Theta}} \boldsymbol{Z}_i\}_{i=1}^n$. We leave the theoretical justification of this modified method to future work.

References

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

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, **71**, 135–171.

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.

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, **57**, 289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, **29**, 1165–1188.

Blanchard, G. and Roquain, E. (2009). Adaptive false discovery rate control under independence and dependence. *Journal of Machine Learning Research*, **10**, 2837–2871.

Catoni, O. (2012). Challenging the empirical mean and empirical variance: A deviation study. *Annales de I'Institut Henri Poincaré - Probabilités et Statistiques*, **48**, 1148–1185.

Clarke, S. and Hall, P. (2009). Robustness of multiple testing procedures against dependence. *The Annals of Statistics*, **37**, 332–358.

Cont, R. (2001). Empirical properties of asset returns: Stylized facts and statistical issues.

Quantitative Finance, **1**, 223–236.

Cox, D. R. (1975). A note on data-splitting for the evaluation of significance levels. *Biometrika*, **62**, 441–444.

Delaigle, A., Hall, P. and Jin, J. (2011). Robustness and accuracy of methods for high dimensional data analysis based on Student's t-statistic. *Journal of the Royal Statistical Society,* Series B, **73**, 283–301.

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.

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.

Eklund, A., Nichols, T. and Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*, **113**, 7900–7905.

Fama, E. F. (1963). Mandelbrot and the stable paretian hypothesis. *Journal of Business* **36** 420–429.

Fan, J. and Han, X. (2017). Estimation of the false discovery proportion with unknown dependence. *Journal of the Royal Statistical Society,* Series B, **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., Li, Q. and Wang, Y. (2017). Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *Journal of the Royal Statistical Society,* Series B, **79**, 247–265.

Fan, J., Liao, Y. and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society,* Series B, **75**, 603–680.

Ferreira, J. A. and Zwinderman, A. H. (2006). On the Benjamini-Hochberg method. *The Annals of Statistics*, **34**, 1827–1849.

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.

Genovese, C. and Wasserman, L. (2004). A stochastic process approach to false discovery control. *The Annals of Statistics*, **32**, 1035–1061.

Hartigan, J. A. (1969). Using subsample values as typical values. *Journal of the American Statistical Association*, **64**, 1303–1317.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, **35**, 73–101.

Jin, J. (2008). Proportion of non-zero normal means: universal oracle equivalences and uniformly consistent estimators. *Journal of the Royal Statistical Society,* Series B, **70** 461–493.

Jin, J. and Cai, T. T. (2007). Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons. *Journal of the American Statistical Association*, **102**, 495–506.

Kosorok, M. R. and Ma, S. (2007). Marginal asymptotics for the "large p, small n" paradigm: With applications to microarray data. *The Annals of Statistics*, **35**, 1456–1486.

Kustra, R., Shioda, R. and Zhu, M. (2006). A factor analysis model for functional genomics. *BMC Bioinformatics*, **7**: 216.

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

Langaas, M. and Lindqvist, B. (2005). Estimating the proportion of true null hypotheses, with application to DNA microarray data. *Journal of the Royal Statistical Society,* Series B, **67**, 555–572.

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.

Lehmann, E. L. and Romano, J. P. (2005). Generalizations of the familywise error rate. *The Annals of Statistics*, **33**, 1138–1154.

Liu, L., Hawkins, D. M., Ghosh, S. and Young, S. S. (2003). Robust singular value decomposition analysis of microarray data. *Proceedings of the National Academy of Sciences*, **100**, 13167–13172.

Liu, W. and Shao, Q.-M. (2014). Phase transition and regularized bootstrap in large-scale t-tests with false discovery rate control. *The Annals of Statistics*, **42**, 2003–2025.

Mandelbrot, B. (1963). The variation of certain speculative prices. *Journal of Business* **36** 394–419.

Medland, S., Jahanshad, N., Neale, B. and Thompson, P. (2014). The variation of certain speculative prices. *Nature Neuroscience* **17** 791–800.

Meinshausen, N. and Rice, J (2006). Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *The Annals of Statistics*, **34**, 373–393.

Minsker, S. (2016). Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries. Available at *arXiv:1605.07129*.

Oberthuer, A., Berthold, F., Warnat, P., Hero, B., Kahlert, Y., Spitz, R., Ernestus, K., König, R., Haas, S., Eils, R., Schwab, M., Brors, B., Westermann, F. and Fischer, M. (2006). Customized oligonucleotide microarray gene expression based classification of neuroblastoma patients outperforms current clinical risk stratification. *Journal of Clinical Oncology*, **24**, 5070–5078.

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

Owen, A. B. (2005). Variance of the number of false discoveries. *Journal of the Royal Statistical Society,* Series B, **67**, 411–426.

Posekany, A., Felsenstein, K. and Sykacek, P. (2011). Biological assessment of robust noise models in microarray data analysis. *Bioinformatics*, **27**, 807–814.

Pournara, I. and Wernish, L. (2007). Factor analysis for gene regulatory networks and transcription factor activity profiles. *BMC Bioinformatics*, **8**: 61.

Purdom, E. and Holmes, S. P. (2005). Error distribution for gene expression data. *Statistical Applications in Genetics and Molecular Biology*, **4**: 16.

Rinaldo A., Wasserman, L., G'Snell, M., Lei, J. and Tibshirani, R. (2016). Bootstrapping and sample splitting for high-dimensional, assumption-free inference. Available at *arXiv:1611.05401*.

Schwartzman, A. and Lin, X. (2011). The effect of correlation in false discovery rate estimation. *Biometrika*, **98**, 199–214.

Shen, D., Shen, H., Zhu, H. and Marron, J. S. (2016). The statistics and mathematics of high dimension low sample size asymptotics. *Statistica Sinica*, **26**, 1747–1770.

Sun, W. and Cai, T. T. (2009). Large-scale multiple testing under dependence. *Journal of the Royal Statistical Society,* Series B, **71**, 393–424.

Sun, Q., Zhou, W.-X. and Fan, J. (2017). Adaptive Huber regression: Nonasymptotic optimality and phase transition. Available at *arXiv:1706.06991*.

Stock, J. and Watson, M. (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, **64**, 479–498.

Storey, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *Journal of the Royal Statistical Society,* Series B, **64**, 479–498.

Storey, J. D., Taylor, J. E. and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rate: A unified approach. *Journal of the Royal Statistical Society,* Series B, **66**, 187–205.

Storey, J. D. and Tibshirani, R. (2003). SAM thresholding and false discovery rates for detecting differential gene expression in DNA microarrays. In *The Analysis of Gene Expression Data: Methods and Software* (eds G. Parmigiani, E. S. Garrett, R. A. Irizarry and S. L. Zeger). New York: Springer.

Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing* (eds Y. C. Eldar and G. Kutyniok), pp. 210–268. Cambridge: Cambridge University Press.

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. *The Annals of Statistics*, **45**, 1342–1374.

Wu, W. B. (2008). On false discovery control under dependence. *The Annals of Statistics*, **36**, 364–380.

Yu, Y., Wang, T. and Samworth, R. J. (2015). A useful variant of the Davis-Kahan theorem for statisticians. *Biometrika*, **102**, 315–323.