Submitted to the Annals of Statistics

A NEW PERSPECTIVE ON ROBUST M-ESTIMATION: FINITE SAMPLE THEORY AND APPLICATIONS TO DEPENDENCE-ADJUSTED MULTIPLE TESTING[*]

BY WEN-XIN ZHOU[†], KOUSHIKI BOSE[†], JIANQING FAN[†,‡] AND HAN LIU[†]

Princeton University[†] *and Fudan University*[‡]

Heavy-tailed errors impair the accuracy of the least square estimate, which can be spoiled by a single grossly outlying observation. As argued in the seminar work of Peter Huber in 1973 [*Ann. Statist.* **1** (1973) 799–821], robust alternatives to the method of least squares are sorely needed. To achieve robustness against heavy tailedness, we revisit the Huber estimator from a new perspective by letting the regularization parameter involved diverge with the sample size. In this paper we develop nonasymptotic concentration results for such an adaptive Huber estimator, namely, the Huber estimator with the regularization parameter adapted to sample size, dimension and the variance of the noise. Specifically, we obtain a sub-Gaussian-type deviation inequality and a nonasymptotic Bahadur representation in the presence of heavy-tailed errors. The nonasymptotic results further yield two important normal approximation results, the Berry-Esseen inequality and Cramér-type moderate deviation.

As an important application to large-scale simultaneous inference, we apply these robust normal approximation results to analyze a dependence-adjusted multiple testing procedure for moderately heavy-tailed data. It is shown that the robust, dependence-adjusted procedure asymptotically controls the overall false discovery proportion at the nominal level under mild moment conditions. Thorough numerical results on both simulated and real datasets are also provided to back up our theory.

1. Introduction. High dimensional data are often automatically collected with low quality. For each feature, the samples drawn from a moderate-tailed distribution may comprise one or two very large outliers in the measurements. When dealing with thousands or tens of thousands of features simultaneously, the chance of including a fair amount of outliers is high. Therefore, the development of robust procedures is arguably even more im-

[*]J. Fan and W.-X. Zhou were partially supported by NIH Grant R01–GM072611 and NSF Grant DMS–1206464. K. Bose was supported in part by NSF Grant DGE–1148900.

MSC 2010 subject classifications: Primary 62F03, 62F35; secondary 62J05, 62E17

Keywords and phrases: Approximate factor model, Bahadur representation, false discovery proportion, heavy-tailed data, Huber loss, large-scale multiple testing, M-estimator

portant for high dimensional problems. In this paper, we develop a finite sample theory for robust M-estimation from a new perspective. Such a finite sample theory is motivated by contemporary problems of simultaneously testing many hypotheses. In these problems, our goal is either to control the false discovery rate (FDR)/false discovery proportion (FDP), or to control the familywise error rate (FWER).

The main contributions of this paper are described and summarized in the following two subsections.

1.1. *A finite sample theory of robust M-estimation.* Consider a linear regression model $Y = \mu^* + \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^* + \varepsilon$, where $\mu^* \in \mathbb{R}$, $\boldsymbol{\beta}^* \in \mathbb{R}^d$, $\boldsymbol{X} \in \mathbb{R}^d$ is the covariate and ε is the random noise variable with mean zero and finite variance. Under the assumption that ε has exponentially light tails, statistical properties of the ordinary least squares (OLS) estimator of μ^* and $\boldsymbol{\beta}^*$ have been well studied. When ε follows a heavy-tailed distribution, more robust methods, such as the Huber estimator [Huber (1964)], are needed to replace the OLS estimator. However, unlike the OLS estimator, all the existing theoretical results for the Huber estimator are asymptotic, including asymptotic normality [Huber (1973), Yohai and Maronna (1979), Portnoy (1985), Mammen (1989)] and the Bahadur representation [He and Shao (1996, 2000)]. The main reason for the lack of nonasymptotic results is that the Huber estimator does not bare an explicit closed-form solution. This is in sharp contrast to the OLS estimator which has a closed-form representation. Most existing nonasymptotic analysis of the OLS estimator depends on such a closed-form representation.

The first contribution of this paper is to develop a new finite sample theory for the Huber estimator. Specifically, we denote the Huber loss

$$\ell_\tau(u) = \begin{cases} \frac{1}{2}u^2 & \text{if } |u| \le \tau \\ \tau|u| - \frac{1}{2}\tau^2 & \text{if } |u| > \tau \end{cases}, \tag{1.1}$$

where $\tau > 0$ is a regularization parameter that balances robustness and efficiency. In line with this notation, we use ℓ_∞ to denote the quadratic loss $\ell_\infty(u) = u^2/2$, $u \in \mathbb{R}$. Let $\{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n$ be independent random samples from the model $Y = \mu^* + \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^* + \varepsilon$. We define the robust M-estimator of $\boldsymbol{\theta}^* := (\mu^*, \boldsymbol{\beta}^{*\mathrm{T}})^{\mathrm{T}}$ by

$$\widehat{\boldsymbol{\theta}} := (\widehat{\mu}, \widehat{\boldsymbol{\beta}}^{\mathrm{T}})^{\mathrm{T}} \in \underset{\mu \in \mathbb{R},\, \boldsymbol{\beta} \in \mathbb{R}^d}{\operatorname{argmin}} \sum_{i=1}^n \ell_\tau(Y_i - \mu - \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}). \tag{1.2}$$

It is worth noticing that our robustness concern is fairly different from the conventional one. In Huber's robust mean estimation [Huber (1964)], it is assumed that the error distribution is in the neighbourhood of a normal distribution, which gives the possibility to replace the mean by some location parameter. Unless the shape of the distribution is constrained (e.g., symmetry), in general this location parameter will not be equal to the mean. Our interest, however, is pinned on the heavy-tailed case where the error distribution is allowed to be asymmetric and exhibit heavy tails. Therefore, unlike the classical Huber estimator [Huber (1973)] which requires τ to be fixed, we allow τ to diverge with the sample size n such that $\ell_\tau(\cdot)$ can be viewed as an approximate quadratic (RA) loss function. As in Fan, Li and Wang (2017), this is needed to reduce the bias for estimating the (conditional) mean function when the (conditional) distribution of ε is asymmetric and has heavy tails. In particular, by taking $\tau = \infty$, $\widehat{\boldsymbol{\theta}}$ coincides with the OLS estimator of $\boldsymbol{\theta}^*$ and by taking $\tau = 0$, the resulting estimator is the least absolute deviation (LAD) estimator.

For every $\tau > 0$, by definition $\widehat{\boldsymbol{\theta}}$ is an M-estimator of

$$\boldsymbol{\theta}^*_\tau := (\mu_\tau, \boldsymbol{\beta}_\tau^{\mathrm{T}})^{\mathrm{T}} = \operatorname*{argmin}_{\mu\in\mathbb{R},\, \boldsymbol{\beta}\in\mathbb{R}^d} \mathbb{E}\{\ell_\tau(Y - \mu - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})\}, \tag{1.3}$$

which can differ from the target parameter $\boldsymbol{\theta}^* = \operatorname{argmin}_{\mu,\boldsymbol{\beta}} \mathbb{E}\ell_\infty(Y - \mu - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})$. Note that the total error $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|$ can be divided into two parts:

$$\underbrace{\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|}_{\text{Total error}} \le \underbrace{\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*_\tau\|}_{\text{Estimation error}} + \underbrace{\|\boldsymbol{\theta}^*_\tau - \boldsymbol{\theta}^*\|}_{\text{Approximation error}},$$

where $\|\cdot\|$ denotes the Euclidean norm. We define $\mathrm{Bias}(\tau) := \|\boldsymbol{\theta}^*_\tau - \boldsymbol{\theta}^*\|$ to be the approximation error introduced by the Huber loss. Proposition 1.1 in the supplemental material [Zhou et al. (2017)] shows that $\mathrm{Bias}(\tau)$ scales at the rate τ^{-1}, which decays as τ grows. A large τ reduces the approximation bias but jeopardizes the degree of robustness. Hence, the tuning parameter τ controls this bias and robustness trade-off of the estimator. Our main theoretical result (Theorem 2.1) is to prove a nonasymptotic Bahadur representation for $\widehat{\boldsymbol{\theta}}$. Such a Bahadur representation provides a finite sample approximation of $\widehat{\boldsymbol{\theta}}$ by a sum of independent variables with a higher-order remainder. More specifically, let $\boldsymbol{\Sigma} = \mathbb{E}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})$ and $\ell'_\tau(\cdot)$ be the derivative function of $\ell_\tau(\cdot)$, we have with properly chosen $\tau = \tau_n$ that

$$\left\| \sqrt{n} \begin{bmatrix} \widehat{\mu} - \mu^* \\ \boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \end{bmatrix} - \frac{1}{\sqrt{n}} \sum_{i=1}^n \ell'_\tau(\varepsilon_i) \begin{bmatrix} 1 \\ \boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}_i \end{bmatrix} \right\| \le R_n(\tau) \simeq \frac{d}{\sqrt{n}}, \tag{1.4}$$

where $R_n(\tau)$ is a finite sample error bound which characterizes the accuracy of such linear approximation, and d is the number of covariates that possibly grows with n. We refer to Theorem 2.1 for a rigorous description of the result (1.4), where we obtain an exponential-type probability inequality for the above Bahadur representation.

Many asymptotic Bahadur-type representations for robust M-estimators have been obtained in the literature. See, for example, Portnoy (1985), Mammen (1989), and He and Shao (1996, 2000), among others. Our result in (1.4), however, is nonasymptotic and provides an explicit tail bound for the remainder term $R_n(\tau)$. To obtain such a result, we first derive a sub-Gaussian-type deviation bound for $\widehat{\boldsymbol{\theta}}$; and then conduct a careful analysis on the higher-order remainder term using this bound and techniques from empirical process theory. The result in (1.4) further yields a number of important normal approximation results, including the Berry-Esseen inequality and a Cramér-type moderate deviation result. These results have important applications to large-scale inference as we shall describe below. In particular, we consider the problems of simultaneously testing many hypotheses (e.g., controlling FDP) or globally inferring a high dimensional parameter (e.g., controlling familywise error). For multiple testing, the obtained Berry-Esseen bound and Cramér-type moderate deviation result can be used to investigate the robustness and accuracy of the P-values and critical values. For globally testing a high dimensional parameter, the result in (1.4) combined with parametric bootstrap can be used to construct a valid test. In this paper, we only focus on the large-scale multiple testing problem, which is described in the next sub-section.

1.2. *FDP control for robust dependent tests.* We apply the representation in (1.4) to control FDP for dependent robust test statistics. Conventional tasks of large-scale testing, including controlling FDR/FDP or FWER, have been extensively explored and are now well understood when all the test statistics are independent [Benjamini and Hochberg (1995), Storey (2002), Genovese and Wasserman (2004), Lehmann and Romano (2005)]. It is becoming increasingly important to understand the dependence among multiple test statistics. Under the positive regression dependence condition, the FDR control can be conducted in the same manner as that for the independent case [Benjamini and Yekutieli (2001)], which provides a conservative upper bound. For more general dependence, directly applying standard FDR control procedures developed for independent P-values can lead to inaccurate false discovery rate control and spurious outcomes [Efron (2004), Efron (2007), Sun and Cai (2009), Clarke and Hall (2009), Schwartzman and Lin (2011), Fan, Han and Gu (2012)]. In this more challenging situation, various multi-factor models have been proposed to investigate the dependence structure in high dimensional data. See, for example, Leek and Storey (2008), Friguet, Kloareg and Causeur (2009), Desai and Storey (2012) and Fan, Han and Gu (2012).

The multi-factor model relies on the identification of a linear space of random vectors capturing the dependence structure of the data. Friguet, Kloareg and Causeur (2009) and Desai and Storey (2012) assume that the data are drawn from a strict factor model with independent idiosyncratic errors. They use the expectation-maximization algorithm to estimate the factor loadings and realized factors in the model, and then obtain an estimator for the FDP by subtracting out realized common factors. These methods, however, depend on stringent modeling assumptions, including the independence of idiosyncratic errors and the joint normality of the factor and noise. In contrast, Fan, Han and Gu (2012) and Fan and Han (2017) propose a more general approximate factor model which allows dependent noise.

More specifically, 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 $\boldsymbol{\Sigma} = (\sigma_{jk})_{1 \le j,k \le p}$. We aim to simultaneously test the hypotheses

$$H_{0j} : \mu_j = 0 \quad \text{versus} \quad H_{1j} : \mu_j \neq 0, \quad \text{for } j = 1, \ldots, p. \tag{1.5}$$

We are interested in the case where there is strong dependence across the components in $\boldsymbol{X}$. The approximate factor model assumes the dependence of a high dimensional random vector $\boldsymbol{X}$ can be captured by a few factors, i.e., it assumes that $\boldsymbol{X}$ satisfies

$$\boldsymbol{X} = \boldsymbol{\mu} + \mathbf{B}\boldsymbol{f} + \boldsymbol{\Sigma}(\boldsymbol{f})\boldsymbol{u}, \tag{1.6}$$

from which we observe independent random samples $(\boldsymbol{X}_1, \boldsymbol{f}_1), \ldots, (\boldsymbol{X}_n, \boldsymbol{f}_n)$ satisfying

$$\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^{\mathrm{T}} = \boldsymbol{\mu} + \mathbf{B}\boldsymbol{f}_i + \boldsymbol{\Sigma}(\boldsymbol{f}_i)\boldsymbol{u}_i, \quad i = 1, \ldots, n.$$

Here, $\mathbf{B} = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_p)^{\mathrm{T}} \in \mathbb{R}^{p \times K}$ represents the factor loading matrix, $\boldsymbol{f}_i$ is the K-dimensional common factor to the ith observation and is independent of the idiosyncratic noise $\boldsymbol{u}_i \in \mathbb{R}^p$, and $\boldsymbol{\Sigma}(\boldsymbol{f}) = \mathrm{diag}(\sigma_1(\boldsymbol{f}), \ldots, \sigma_p(\boldsymbol{f}))$ with $\sigma_j(\cdot) : \mathbb{R}^K \mapsto (0, \infty)$ as unknown variance heteroscedasticity functions. We allow the components of $\boldsymbol{u}_i$ to be dependent. To fully understand the influence of heavy tailedness, in this paper we restrict our attention to such an observable factor model.

For testing the hypotheses in (1.5) under model (1.6), when the factors $\boldsymbol{f}$ are latent, a popular and natural approach is based on the (marginal) sample averages of $\{\boldsymbol{X}_i\}$ and attentions focus on the valid control of FDR [Benjamini and Yekutieli (2001), Genovese and Wasserman (2004), Efron (2007), Sun and Cai (2009), Schwartzman and Lin (2011), Fan and Han (2017)]. As pointed out in Fan, Han and Gu (2012), the power of such an approach is dominated by the factor-adjusted approach that produces alternative rankings of statistical significance from those of the marginal statistics. They focus on a Gaussian model where both $\boldsymbol{f}$ and $\boldsymbol{u}$ follow multivariate normal distributions, while its statistical properties haven't been studied. The normality assumption, however, is an idealization which provides insights into the key issues underlying the problems. Data subject to heavy-tailed and asymmetric errors are repeatedly observed in many fields of research [Finkenstadt and Rootzén (2003)]. For example, it is known that financial returns typically exhibit heavy tails. The important papers by Mandelbrot (1963) and Fama (1963) provide evidence of power-law behavior in asset prices in the early 1960s. Since then, the non-Gaussian character of the distribution of price changes has been widely observed in various market data. Cont (2001) provides further numerical evidence to show that a Student's t-distribution with four degrees of freedom displays a tail behavior similar to many asset returns.

For multiple testing with heavy-tailed data, test statistics based on OLS estimators are sensitive to outliers, and thus lack robustness against heavy tailedness. This issue is amplified further by high dimensionality: When dimension is large, even moderate tails may lead to significant false discoveries. This motivates us to develop new test statistics that are robust to the tails of error distributions. Also, since the multiple testing problem is more complicated when the variables are dependent, theoretical guarantee of the FDP control for the existing dependence-adjusted methods remains unclear.

To illustrate the impact of heavy tailedness, we generate independent and identically distributed (i.i.d.) random variables $\{X_{ij},\, i=1,\ldots,30,\, j=1,\ldots,10000\}$ from a normalized t-distribution with 2.5 degrees of freedom. In Figure 1, we compare the histogram of the empirical means $\overline{X}_j$ with that of the robust mean estimates constructed using (1.3) without covariates, after rescaling both estimates by $\sqrt{30}$. For a standard normal distribution, we expect 99.73% data points to lie with three standard deviations of the mean or inside $[-3,3]$. Hence for this experiment, if the sampling distribution of the mean estimates indeed approached a Gaussian, we would expect about 27 out of 10000 estimates to lie outside $[-3,3]$. From Figure 1 we see that the robust procedure gives 28 points that fall outside this interval, whereas the sample average gives a much larger number, 79, many of which would surely be regarded as discoveries. We see that in the presence of heavy tails and high dimensions, using the robust estimator leads to a more accurate Gaussian limiting approximation than using a non-robust one. In fact, for

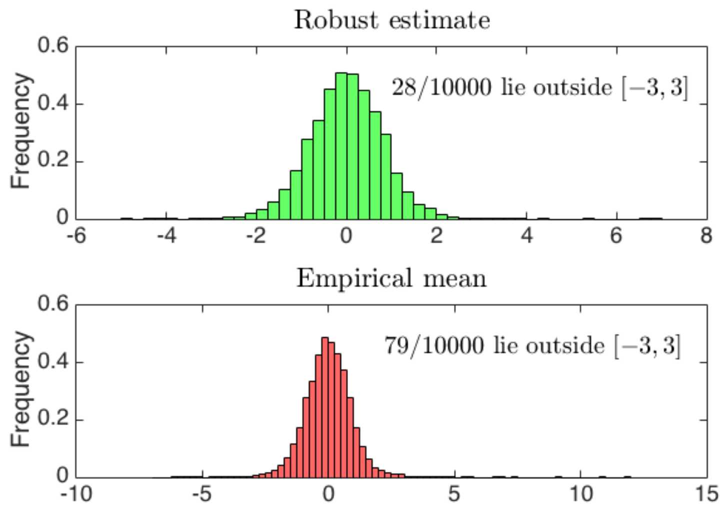

FIG 1. *Histograms of 10000 robust mean estimates and empirical means based on 30 i.i.d. samples drawn from a normalized* t*-distribution with* 2.5 *degrees of freedom. Both the mean estimates are rescaled by* $\sqrt{30}$.

the empirical means there are many outliers that are outside $[-6, 6]$ based on a sample of size 30. This inaccuracy in the tail approximation by the non-robust estimator gives rise to false discoveries. In summary, outliers from the test statistics $\overline{X}_j$ can be so large that they are mistakenly regarded as discoveries, whereas the robust approach results in fewer outliers.

In Section 3, we develop robust, dependence-adjusted multiple testing procedures with solid theoretical guarantees. We use the approximate factor model (1.6) with an observable factor and relatively heavy-tailed errors to characterize the dependence structure in high dimensional data. Assuming such a model, we construct robust test statistics based on the Huber estimator with a diverging parameter, denoted by $T_1, \ldots, T_p$, for testing the individual hypotheses. At a prespecified level $0 < \alpha < 1$, we apply a family of FDP controlling procedures to the dependence-adjusted P-values $\{P_j = 2\Phi(-|T_j|)\}_{j=1}^p$ to decide which null hypotheses are rejected, where Φ is the standard normal distribution function. To establish rigorous theoretical guarantee of the resulting procedure on FDP control, a delicate analysis of the impact of dependence-adjustment on the distribution of the P-values is required. We show that, under mild moment and regularity conditions, the robust multiple testing procedure controls the FDP at any prespecified level asymptotically. Specifically, applying Storey's procedure [Storey (2002)] to the above P-values gives a data-driven rejection threshold $\widehat{z}_{\mathrm{N}}$ such that H_{0j} is rejected whenever $|T_j| \geq \widehat{z}_{\mathrm{N}}$. Let $\mathrm{FDP}(z) = V(z)/\max\{1, R(z)\}$ be the FDP at threshold $z \geq 0$, where $V(z) = \sum_{j=1}^{p} 1(|T_j| \geq z, \mu_j = 0)$ and $R(z) = \sum_{j=1}^{p} 1(|T_j| \geq z)$ are the number of false discoveries and the number of total discoveries, respectively. In the ultra-high dimensional setting that p can be of order e^{n^c} for some $0 < c < 1$, we prove that

$$\frac{p}{p_0}\,\mathrm{FDP}(\widehat{z}_{\mathrm{N}}) \to \alpha \quad \text{in probability} \tag{1.7}$$

as $(n, p) \to \infty$, where $p_0 = \sum_{j=1}^{p} 1(\mu_j = 0)$ is the number of true null hypotheses. We also illustrate the usefulness of the robust techniques by contrasting the behaviors of inference procedures using robust and OLS estimates in synthetic numerical experiments.

Key technical tools in proving (1.7) are the Berry-Esseen bound and Cramér-type deviation result for the marginal statistic T_j. These results are built upon the nonasymptotic Bahadur representation (1.4), and may be of independent interest for other statistical applications. For example, Delaigle, Hall and Jin (2011) explore moderate and large deviations of the t-statistic in a variety of high dimensional settings.

1.3. *Organization of the paper.* The lay-out of the paper is as follows. In Section 2, we develop a general finite sample theory for Huber's robust M-estimator from a new perspective where a diverging parameter is involved. In Section 3, we formally describe the proposed robust, dependence-adjusted multiple testing procedure with rigorous theoretical guarantees. Section 4 consists of Monte Carlo evaluations and real data analysis. The simulation study provides empirical evidence that the proposed robust inference procedure improves performance in the presence of asymmetric and heavy-tailed errors, and maintains efficiency under light-tail situations. A discussion is given in Section 5. Proofs of the theoretical results in Sections 2 and 3 are provided in the supplemental material [Zhou et al. (2017)].

NOTATION. For a vector $\boldsymbol{u} = (u_1, \ldots, u_p)^{\mathrm{T}} \in \mathbb{R}^p$ $(p \geq 2)$, we use $\|\boldsymbol{u}\| = \big(\sum_{j=1}^{p} u_j^2\big)^{1/2}$ to denote its ℓ_2-norm. Let $\mathbb{S}^{p-1} = \{\boldsymbol{u} \in \mathbb{R}^p : \|\boldsymbol{u}\| = 1\}$ represent the unit sphere in $\mathbb{R}^p$. For a matrix $\mathbf{A} \in \mathbb{R}^{p\times p}$, $\|\mathbf{A}\| = \sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}} \|\mathbf{A}\boldsymbol{u}\|$ denotes the spectral norm of $\mathbf{A}$. For any two sequences $\{a_n\}$ and $\{b_n\}$ of positive numbers, denote by $a_n \asymp b_n$ when $cb_n \leq a_n \leq Cb_n$ for some absolute constants $C \geq c > 0$, denote by $a_n \sim b_n$ if $a_n/b_n \to 1$ as $n \to \infty$. Moreover, we write $a_n = O(b_n)$ if $a_n \leq Cb_n$ for some absolute constant $C > 0$, write $a_n = o(b_n)$ if $a_n/b_n \to 0$ as $n \to \infty$, and write $a_n = o_{\mathbb{P}}(b_n)$ if $a_n/b_n \to 0$ in probability as $n \to \infty$. For a set S, we use S^{c} to denote its complement and $\mathrm{Card}(S)$ for its cardinality. For $x \in \mathbb{R}$, denote by $\lfloor x \rfloor$ the largest integer not greater than x and $\lceil x \rceil$ the smallest integer not less than x. For any two real numbers a and b, we write $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$.

2. Robust M-estimation: A finite sample theory. Consider a heteroscedastic linear regression model $Y = \mu^* + \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^* + \sigma(\boldsymbol{X})\varepsilon$, from which we collect independent and identically distributed observations

$$Y_i = \mu^* + \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^* + \sigma(\boldsymbol{X}_i)\varepsilon_i, \quad i = 1, \ldots, n, \tag{2.1}$$

where μ^* is the intercept, $\boldsymbol{X}$ is a d-dimensional covariate vector, $\boldsymbol{\beta}^*$ is a d-dimensional regression coefficient vector, ε denotes the random error that is independent of $\boldsymbol{X}$, and $\sigma(\cdot) : \mathbb{R}^d \mapsto (0, \infty)$ is an unknown variance function. We assume that both the distributions of $\boldsymbol{X}$ and ε have mean 0. Under this assumption, μ^* and $\boldsymbol{\beta}^*$ together are related to the conditional mean effects of Y given $\boldsymbol{X}$, and μ^* is the unconditional mean effect of Y that is of independent interest in many applications. For simplicity, we introduce the following notations:

$$\boldsymbol{\theta}^* = (\mu^*, \boldsymbol{\beta}^{*\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{d+1}, \quad \boldsymbol{Z} = (1, \boldsymbol{X}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{d+1}, \quad \nu = \sigma(\boldsymbol{X})\varepsilon,$$
$$\text{and} \quad \boldsymbol{Z}_i = (1, \boldsymbol{X}_i^{\mathrm{T}})^{\mathrm{T}}, \quad \nu_i = \sigma(\boldsymbol{X}_i)\varepsilon_i, \quad i = 1, \ldots, n.$$

In this section, we study the robust estimator of $\boldsymbol{\theta}^*$ defined in (1.2). In particular, we show that it admits exponential-type deviation bounds even for heavy-tailed error distributions. Note that, under the heteroscedastic model (2.1), $\boldsymbol{\theta}^*$ differs from the median effect of Y conditioning on $\boldsymbol{X}$. In this setting, the LAD-based methods are not applicable to estimate $\boldsymbol{\theta}^*$. Instead, we focus on Huber's robust estimator $\widehat{\boldsymbol{\theta}}$ given in (1.2) with a diverging regularization parameter $\tau = \tau_n$ that balances the approximation error and robustness of the estimator.

First, we impose the following conditions on the linear model (2.1).

(M1). The random vector $\boldsymbol{X}$ satisfies $\mathbb{E}(\boldsymbol{X}) = \mathbf{0}$ and $\mathbb{E}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}) = \boldsymbol{\Sigma}$ for some positive definite matrix $\boldsymbol{\Sigma}$. Moreover, $\boldsymbol{X}$ is sub-Gaussian with $K_0 = \|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}\|_{\psi_2} < \infty$, where $\|\cdot\|_{\psi_2}$ denotes the vector sub-Gaussian norm [Vershynin (2012)]. The error variable ε is independent of $\boldsymbol{X}$ satisfying $\mathbb{E}(\varepsilon) = 0$ and $\mathbb{E}(\varepsilon^2) = 1$. In addition, we assume $\sigma(\cdot) : \mathbb{R}^d \mapsto (0, \infty)$ to be a positive function and assume that $\sigma^2 = \mathbb{E}\{\sigma^2(\boldsymbol{X})\}$ is finite.

Condition (M1) allows a family of conditional heteroscedastic models with heavy-tailed error ε. Specifically, it only requires the second moment of $\nu = \sigma(\boldsymbol{X})\varepsilon$ to be finite.

Our first result, Theorem 2.1, provides an exponential bound as well as a nonasymptotic version of Bahadur representation for the robust estimator $\widehat{\boldsymbol{\theta}} = (\widehat{\mu}, \widehat{\boldsymbol{\beta}}^{\mathrm{T}})^{\mathrm{T}}$ defined in (1.2).

THEOREM 2.1. *Under the linear model* (2.1) *with Condition (M1) satisfied, we have for any* $w > 0$, *the robust estimator* $\widehat{\boldsymbol{\theta}}$ *in* (1.2) *with* $\tau = \tau_n = \tau_0 \sqrt{n}(d+1+w)^{-1/2}$ *and* $\tau_0 \ge \sigma$ *satisfies*

$$\mathbb{P}\big\{\|\mathbf{S}^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\| > a_1(d+w)^{1/2} n^{-1/2}\big\} \le 7e^{-w} \quad \textit{and} \tag{2.2}$$

$$\mathbb{P}\left\{\left\|\sqrt{n}\,\mathbf{S}^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \mathbf{S}^{-1/2}\frac{1}{\sqrt{n}}\sum_{i=1}^n \ell_\tau'(\nu_i)\boldsymbol{Z}_i\right\| > a_2 \frac{d+w}{\sqrt{n}}\right\} \le 8e^{-w} \tag{2.3}$$

as long as $n \ge a_3(d+w)^{3/2}$, *where* $\mathbf{S} = \mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}})$ *and* a_1*–*a_3 *are positive constants depending only on* τ_0, K_0 *and* $\|\mathbf{S}^{-1/2}\overline{\mathbf{S}}\mathbf{S}^{-1/2}\|$ *with* $\overline{\mathbf{S}} = \mathbb{E}\{\sigma^2(\boldsymbol{X})\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}\}$.

An important message of Theorem 2.1 is that even for heavy-tailed error with only finite second moment, the robust estimator $\widehat{\boldsymbol{\theta}}$ with properly chosen τ has sub-Gaussian tails. See inequality (2.2). To some extent, the tuning parameter τ plays a similar role as the bandwidth in constructing nonparametric estimators. Secondly, we show in (2.3) that the remainder of the Bahadur representation for $\widehat{\boldsymbol{\theta}}$ exhibits sub-exponential tails. To the best of our knowledge, no nonasymptotic results of this type exist in the literature, and the existing asymptotic results can only be used to derive polynomial-type tail probability bounds.

A direct consequence of (2.3) is that $\sqrt{n}(\widehat{\mu} - \mu^*)$ is close to $W_n := n^{-1/2}\sum_{i=1}^n \ell_\tau'(\nu_i)$ with probability exponentially fast approaching one. The next result reveals that, under higher moment condition on $\nu = \sigma(\boldsymbol{X})\varepsilon$, W_n has an asymptotic normal distribution with mean 0 and limiting variance $\sigma^2 = \mathbb{E}(\nu^2)$.

THEOREM 2.2. *Assume that Condition (M1) holds and that* $v_\kappa := \mathbb{E}(|\nu|^\kappa)$ *is finite for some* $\kappa \ge 3$. *Then, there exists an absolute constant* $C > 0$ *such that for any* $\tau > 0$,

$$\begin{aligned}&\sup_{x\in\mathbb{R}} |\mathbb{P}(\sigma^{-1}W_n \le x) - \Phi(x)| \\ &\quad \le C\big(\sigma^{-3}v_3\, n^{-1/2} + \sigma^{-2}v_\kappa\, \tau^{2-\kappa} + \sigma^2\tau^{-2} + \sigma^{-1}v_\kappa\, \tau^{1-\kappa}\sqrt{n}\big).\end{aligned}$$

In particular, we have

$$\sup_{x\in\mathbb{R}} |\mathbb{P}(\sigma^{-1}W_n \le x) - \Phi(x)| \le C\big(\sigma^{-3}v_3\, n^{-1/2} + \sigma^{-2}v_4\, \tau^{-2} + \sigma^{-1}v_4\, \tau^{-3}\sqrt{n}\big).$$

Together, Theorems 2.1 and 2.2 lead to a Berry-Esseen type bound for $T := \sqrt{n}(\widehat{\mu} - \mu^*)/\sigma$ for properly chosen τ. The following theorem gives a Cramér-type deviation result for T, which quantifies the relative error of the normal approximation.

THEOREM 2.3. *Assume that Condition (M1) holds and that* $\mathbb{E}(|\nu|^3) < \infty$. *Let* $\{w_n\}_{n \geq 1}$ *be an arbitrary sequence of numbers satisfying* $w_n \to \infty$ *and* $w_n = o(\sqrt{n})$. *Then, the statistic* T *with* $\tau = \tau_0 \sqrt{n}(d + w_n)^{-1/2}$ *for some constant* $\tau_0 \geq \sigma$ *satisfies that*

$$\mathbb{P}\big(|T| \geq z\big) = \big(1 + C_{n,z}\big)\mathbb{P}\big(|G| \geq z\big) \tag{2.4}$$

uniformly for $0 \leq z = o\{\min(\sqrt{w_n}, \sqrt{n} w_n^{-1})\}$ *as* $n \to \infty$, *where* $G \sim N(0,1)$,

$$|C_{n,z}| \leq C\big\{\big(\sqrt{\log n} + z\big)^3 n^{-1/2} + (1+z)\big(n^{-3/10} + n^{-1/2} w_n\big) + e^{-w_n}\big\}$$

and $C > 0$ *is a constant independent of* n. *In particular, we have*

$$\sup_{0 \leq z \leq o\{\min(\sqrt{w_n}, \sqrt{n} w_n^{-1})\}} \left| \frac{\mathbb{P}(|T| \geq z)}{2 - 2\Phi(z)} - 1 \right| \to 0. \tag{2.5}$$

REMARK 2.1. From Theorem 2.3 we see that the ratio $\mathbb{P}(|T| \geq z)/\{2 - 2\Phi(z)\}$ is close to 1 for a wide range of nonnegative z-values, whose length depends on both the sample size n and w_n. In particular, taking $w_n \asymp n^{1/3}$ gives the widest possible range $[0, o(n^{1/6}))$, which is also optimal for Cramér-type moderate deviation results [Linnik (1961)]. In this case, the regularization parameter $\tau = \tau_n \asymp n^{1/3}$.

REMARK 2.2. Motivated by the applications to large-scale simultaneous inference considered in Section 3, Theorems 2.2 and 2.3 only focus on the intercept μ^*. In fact, similar results can be obtained for $\boldsymbol{\theta}^*$ or a specific coordinate of $\boldsymbol{\theta}^*$ based on the joint Bahadur representation result (2.3).

3. Large-scale multiple testing for heavy-tailed dependent data. In this section, we present a detailed description of the dependence-adjusted multiple testing procedure for the means $\mu_1, \ldots, \mu_p$ in model (1.6), based on random samples from the population vector $\boldsymbol{X}$ that exhibits strong dependence and heavy tails.

3.1. *Robust test statistics.* Suppose we are given independent random samples $\{(\boldsymbol{X}_i, \boldsymbol{f}_i)\}_{i=1}^n$ from model (1.6). We are interested in the simultaneous testing of mean effects (1.5). A naive approach is to directly use the information $X_{ij} \sim N(\mu_j, \sigma_{jj})$ for the dependent case as done in the literature, where σ_{jj} is the jth diagonal element of $\mathbf{\Sigma} = \mathrm{cov}(\boldsymbol{X})$. Such an approach is very natural and popular when factors are unobservable and focus is on the valid control of FDR, but is inefficient as noted in Fan, Han and Gu (2012). Indeed, if the loading matrix $\mathbf{B}$ is known and the factors are observable (otherwise, replaced by their estimates), for each j, we can construct the marginal test statistic using dependence-adjusted observations $\{X_{ij} - \boldsymbol{b}_j^{\mathrm{T}} \boldsymbol{f}_i\}_{i=1}^n$ from $\mu_j + \sigma_j(\boldsymbol{f}) u_j$ for testing the jth hypothesis $H_{0j} : \mu_j = 0$.

We consider the approximate factor model (1.6) and write

$$\boldsymbol{u} = (u_1, \ldots, u_p)^{\mathrm{T}}, \quad \boldsymbol{\nu} = (\nu_1, \ldots, \nu_p)^{\mathrm{T}} = \mathbf{\Sigma}(\boldsymbol{f})\boldsymbol{u},$$
$$\boldsymbol{\nu}_i = (\nu_{i1}, \ldots, \nu_{ip})^{\mathrm{T}} = \mathbf{\Sigma}(\boldsymbol{f}_i)\boldsymbol{u}_i,\ i = 1, \ldots, n.$$

Let $\mathbf{\Sigma}_f$ and $\mathbf{\Sigma}_\nu = (\sigma_{\nu,jk})_{1 \le j,k \le p}$ denote the covariance matrices of $\boldsymbol{f}$ and $\boldsymbol{\nu}$, respectively. Under certain sparsity condition on $\mathbf{\Sigma}_\nu$ (see Section 3.4 for an elaboration), $\nu_1, \ldots, \nu_p$ are weakly dependent random variables with higher signal-to-noise ratios since $\mathrm{var}(\nu_j) = \sigma_{jj} - \|\mathbf{\Sigma}_f^{1/2} \boldsymbol{b}_j\|^2 < \sigma_{jj}$. Therefore, we expect subtracting common factors out would make the resulting FDP control procedure more efficient and powerful. It provides an alternative ranking of the significance of hypothesis from the tests based on marginal statistics.

For each $j = 1, \ldots, p$, we have a linear regression model

$$X_{ij} = \mu_j + \boldsymbol{b}_j^{\mathrm{T}} \boldsymbol{f}_i + \nu_{ij}, \quad i = 1, \ldots, n. \tag{3.1}$$

A natural approach is to estimate μ_j and $\boldsymbol{b}_j$ by the method of least squares. However, the least squares method is sensitive to the tails of the error distributions. Also, as demonstrated by Fan, Li and Wang (2017), the LAD-based methods are not applicable in the presence of asymmetric and heteroscedastic errors. Hence, we consider the following robust procedure, which simultaneously estimates μ_j and $\boldsymbol{b}_j$ by solving

$$(\widehat{\mu}_j, \widehat{\boldsymbol{b}}_j^{\mathrm{T}})^{\mathrm{T}} \in \underset{\mu \in \mathbb{R},\, \boldsymbol{b} \in \mathbb{R}^d}{\mathrm{argmin}} \sum_{i=1}^n \ell_\tau(X_{ij} - \mu - \boldsymbol{b}^{\mathrm{T}} \boldsymbol{f}_i), \tag{3.2}$$

where ℓ_τ is given in (1.1). By Theorems 2.1 and 2.2, the adaptive Huber estimator $\widehat{\mu}_j$ follows a normal distribution asymptotically as $n \to \infty$:

$$\sqrt{n}(\widehat{\mu}_j - \mu_j) \xrightarrow{\mathscr{D}} N(0, \sigma_{\nu,jj}) \ \text{ with } \ \sigma_{\nu,jj} = \mathrm{var}(\nu_j). \tag{3.3}$$

To construct a test statistic for the individual hypothesis $H_{0j} : \mu_j = 0$ with pivotal limiting distribution, we need to estimate $\sigma_{\nu,jj} = \sigma_{jj} - \|\mathbf{\Sigma}_f^{1/2} \boldsymbol{b}_j\|^2$, which is essentially the asymptotic variance of $\sqrt{n}(\widehat{\mu}_j - \mu_j)$. Since $\mathbf{\Sigma}_f$ is the covariance matrix of the low-dimensional factor $\boldsymbol{f}$, it is natural to estimate it by $\widehat{\mathbf{\Sigma}}_f = n^{-1}\sum_{i=1}^n \boldsymbol{f}_i\boldsymbol{f}_i^{\mathrm{T}}$ and use $\widehat{\boldsymbol{b}}_j^{\mathrm{T}}\widehat{\mathbf{\Sigma}}_f\widehat{\boldsymbol{b}}_j$ to estimate $\mathrm{var}(\boldsymbol{b}_j^{\mathrm{T}}\boldsymbol{f})$. Let $\widehat{\sigma}_{jj}$ and $\widehat{\sigma}_{\nu,jj}$ be generic estimators of σ_{jj} and $\sigma_{\nu,jj}$. To simultaneously infer all the hypotheses of interest, we require the uniform convergence results

$$\max_{1\le j\le p}\left|\frac{\widehat{\sigma}_{jj}}{\sigma_{jj}} - 1\right| = o_{\mathbb{P}}(1) \quad \text{and} \quad \max_{1\le j\le p}\left|\frac{\widehat{\sigma}_{\nu,jj}}{\sigma_{\nu,jj}} - 1\right| = o_{\mathbb{P}}(1).$$

For $\sigma_{jj} = \mathrm{var}(X_j)$, it is known that the sample variance $n^{-1}\sum_{i=1}^n (X_{ij} - \overline{X}_j)^2$ performs poorly when X_j has heavy tails. Based on the recent developments on robust mean estimation for heavy-tailed data [Catoni (2012), Joly and Lugosi (2016), Fan, Li and Wang (2017)], we consider the following two types of robust variance estimators.

1 (Adaptive Huber variance estimator). Write $\theta_j = \mathbb{E}(X_j^2)$ so that $\sigma_{jj} = \theta_j - \mu_j^2$. Construct the adaptive Huber estimator of θ_j using the squared data, i.e., $\widehat{\theta}_j = \mathrm{argmin}_{\theta>0}\sum_{i=1}^n \ell_\gamma(X_{ij}^2 - \theta)$, where $\gamma = \gamma_n$ is a tuning parameter. Then, we compute the adaptive Huber estimator $(\widehat{\mu}_j, \widehat{\boldsymbol{b}}_j^{\mathrm{T}})^{\mathrm{T}}$ given in (3.2). The final variance estimator is then defined by

$$\widehat{\sigma}_{\nu,jj} = \begin{cases} \widehat{\theta}_j - \widehat{\mu}_j^2 - \widehat{\boldsymbol{b}}_j^{\mathrm{T}}\widehat{\mathbf{\Sigma}}_f\widehat{\boldsymbol{b}}_j & \text{if } \widehat{\theta}_j > \widehat{\mu}_j^2 + \widehat{\boldsymbol{b}}_j^{\mathrm{T}}\widehat{\mathbf{\Sigma}}_f\widehat{\boldsymbol{b}}_j, \\ \widehat{\theta}_j & \text{otherwise.} \end{cases} \tag{3.4}$$

2 (Median-of-means variance estimator). The median-of-means technique, which dates back to Nemirovsky and Yudin (1983), robustifies the empirical mean by first dividing the given observations into several blocks, computing the sample mean within each block and then taking the median of these sample means as the final estimator. Although the sample variance cannot be represented in a simple average form, it is a U-statistic with a symmetric kernel $h : \mathbb{R}^2 \to \mathbb{R}$ given by $h(x, y) = (x - y)^2/2$. The recent work of Joly and Lugosi (2016) extends the median-of-means technique to construct U-statistics based sub-Gaussian estimators for heavy-tailed data.
Back to the current problem, we aim to estimate σ_{jj} based on independent observations $X_{1j}, \ldots, X_{nj}$. Let $V = V_n < n$ be an integer and decompose n as $n = Vm + r$ for some integer $0 \le r < V$. Let $B_1, \ldots, B_V$ be a partition of $\{1, \ldots, n\}$ defined by

$$B_k = \begin{cases} \{(k-1)m+1, (k-1)m+2, \ldots, km\}, & \text{if } 1 \le k \le V-1, \\ \{(V-1)m+1, (V-1)m+2, \ldots, n\}, & \text{if } k = V. \end{cases} \tag{3.5}$$

For each pair (k,ℓ) satisfying $1 \le k < \ell \le V$, define decoupled U-statistic $U_{j,k\ell} = (2|B_k||B_\ell|)^{-1} \sum_{i_1 \in B_k} \sum_{i_2 \in B_\ell} (X_{i_1 j} - X_{i_2 j})^2$. Then, we estimate σ_{jj} by the median of $\{U_{j,k\ell} : 1 \le k < \ell \le V\}$, i.e., $\widetilde{\sigma}_{jj}(V) \in \operatorname{argmin}_{u \in \mathbb{R}} \sum_{1 \le k < \ell \le V} |U_{j,k\ell} - u|$. As before, we compute the adaptive Huber estimator $(\widehat{\mu}_j, \widehat{\boldsymbol{b}}_j^{\mathrm{T}})^{\mathrm{T}}$. Finally, our robust variance estimators are

$$\widetilde{\sigma}_{\nu,jj} = \begin{cases} \widetilde{\sigma}_{jj}(V) - \widehat{\boldsymbol{b}}_j^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}_f \widehat{\boldsymbol{b}}_j & \text{if } \widetilde{\sigma}_{jj}(V) > \widehat{\boldsymbol{b}}_j^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}_f \widehat{\boldsymbol{b}}_j, \\ \widetilde{\sigma}_{jj}(V) & \text{otherwise,} \end{cases} \tag{3.6}$$

Given robust mean and variance estimators of each type, we construct dependence-adjusted test statistics

$$T_j = \sqrt{n}\, \widehat{\sigma}_{\nu,jj}^{-1/2} \widehat{\mu}_j \quad \text{and} \quad S_j = \sqrt{n}\, \widetilde{\sigma}_{\nu,jj}^{-1/2} \widehat{\mu}_j \quad \text{for } j = 1, \ldots, p. \tag{3.7}$$

In fact, as long as the fourth moment $\mathbb{E}(X_j^4)$ is finite, the estimators $\widehat{\sigma}_{\nu,jj}$ and $\widetilde{\sigma}_{\nu,jj}$ given in (3.4) and (3.6), respectively, are concentrated around $\sigma_{\nu,jj}$ with high probability. Hence, in view of (3.3), under the null hypothesis $H_{0j} : \mu_j = 0$, the adjusted test statistics T_j and S_j satisfy that $T_j \xrightarrow{\mathscr{D}} N(0,1)$ and $S_j \xrightarrow{\mathscr{D}} N(0,1)$ as $n \to \infty$.

3.2. *Dependence-adjusted FDP control procedure.* To carry out multiple testing (1.5) using the test statistics T_j's, let $z > 0$ be the critical value such that H_{0j} is rejected whenever $|T_j| \ge z$. One of the main objects of interest in the present paper is the false discovery proportion

$$\mathrm{FDP}(z) = \frac{V(z)}{\max\{R(z), 1\}}, \quad z \ge 0, \tag{3.8}$$

where $V(z) = \sum_{j \in \mathcal{H}_0} 1(|T_j| \ge z)$ is the number of false discoveries, $R(z) = \sum_{j=1}^{p} 1(|T_j| \ge z)$ and $\mathcal{H}_0 = \{j : 1 \le j \le p, \mu_j = 0\}$ represents the set of true null hypotheses.

The statistical behavior of $\mathrm{FDP}(z)$ is of significant interest in multiple testing. However, the realization of $V(z)$ for a given experiment is unknown and thus needs to be estimated. When the sample size is large, it is natural to approximate $V(z)$ by its expectation $2p_0\Phi(-z)$, where $p_0 = \mathrm{Card}(\mathcal{H}_0)$. In the high dimensional sparse setting, both p and p_0 are large and $p_1 = p - p_0 = o(p)$ is relatively small. Therefore, we can use p as a slightly conservative surrogate for p_0 so that $\mathrm{FDP}(z)$ can be approximated by

$$\mathrm{FDP}_{\mathrm{N}}(z) = \frac{2p\Phi(-z)}{\max\{R(z), 1\}}, \quad z \ge 0. \tag{3.9}$$

We shall prove in Theorem 3.2 that under mild conditions, $\mathrm{FDP}_{\mathrm{N}}(z)$ provides a consistent estimate of $\mathrm{FDP}(z)$ uniformly in $0 \le z \le \Phi^{-1}(1 - m_p/(2p))$ for any sequence of positive number $m_p \le 2p$ satisfying $m_p \to \infty$.

In the non-sparse case where $\pi_0 = p_0/p$ is bounded away from 0 and 1 as $p \to \infty$, $\mathrm{FDP}_{\mathrm{N}}$ given in (3.9) tends to overestimate the true FDP. Therefore, we need to estimate the proportion π_0, which has been studied by Efron et al. (2001), Storey (2002), Genovese and Wasserman (2004), Langaas and Lindqvist (2005) and Meinshausen and Rice (2006), among others. For simplicity, we focus on Storey's approach. Let $\{P_j = 2\Phi(-|T_j|)\}_{j=1}^p$ be the approximate P-values. For a predetermined $\lambda \in [0, 1)$, Storey (2002) suggests the following conservative estimate of π_0:

$$\widehat{\pi}_0(\lambda) = \frac{1}{(1-\lambda)p} \sum_{j=1}^{p} 1(P_j > \lambda). \tag{3.10}$$

The intuition of such an estimator is as follows. Since most of the large P-values correspond to the null and thus are uniformly distributed, for a sufficiently large λ, we expect about $(1-\lambda)\pi_0$ of the P-values to lie in $(\lambda, 1]$. Hence, the proportion of P-values that exceed λ, $p^{-1}\sum_{j=1}^{p} 1(P_j > \lambda)$, should be close to $(1-\lambda)\pi_0$. This gives rise to Storey's procedure.

Incorporating such an estimate of π_0, we obtain a modified estimate of $\mathrm{FDP}(z)$ by

$$\mathrm{FDP}_{\mathrm{N},\lambda}(z) = \frac{2p\,\widehat{\pi}_0(\lambda)\Phi(-z)}{\max\{R(z), 1\}}, \quad z \ge 0. \tag{3.11}$$

In view of (3.9)–(3.11) and the fact $\widehat{\pi}_0(0) = 1$, we have $\mathrm{FDP}_{\mathrm{N},0}(z) = \mathrm{FDP}_{\mathrm{N}}(z)$.

There is also substantial interest in controlling the FDP at a prespecified level $0 < \alpha < 1$ for which the ideal rejection threshold is $z_{\mathrm{oracle}} = \inf\{z \ge 0 : \mathrm{FDP}(z) \le \alpha\}$. By replacing the unknown quantity $\mathrm{FDP}(z)$ by $\mathrm{FDP}_{\mathrm{N},\lambda}(z)$ given in (3.11) for some $\lambda \in [0, 1)$, we reject H_{0j} whenever $|T_j| \ge \widehat{z}_{\mathrm{N},\lambda}$, where

$$\widehat{z}_{\mathrm{N},\lambda} = \inf\bigl\{z \ge 0 : \mathrm{FDP}_{\mathrm{N},\lambda}(z) \le \alpha\bigr\}. \tag{3.12}$$

By Lemmas 1 and 2 in Storey, Taylor and Siegmund (2004), this procedure is equivalent to a variant of the seminal Benjamini-Hochberg (B-H) procedure [Benjamini and Hochberg (1995)] for selecting $\mathcal{S} = \{j : 1 \le j \le p, P_j \le P_{(k_p(\lambda))}\}$ based on the P-values $P_j = 2\Phi(-|T_j|)$, where $k_p(\lambda) := \max\{j : 1 \le j \le p, P_{(j)} \le \frac{\alpha j}{\widehat{\pi}_0(\lambda)p}\}$ and $P_{(1)} \le \cdots \le P_{(p)}$ are the ordered P-values. In Theorem 3.3, we show that under weak moment conditions, the FDP of this dependence-adjusted procedure with $\lambda = 0$ converges to α in ultra-high dimensions.

Note that $\mathrm{FDP}_{\mathrm{N},0}(z)$ is the most conservatively biased estimate of $\mathrm{FDP}(z)$ among all $\lambda \in [0,1)$ using normal calibration. The statistical power of the corresponding procedure can be compromised if π_0 is much smaller than 1. In general, the procedure requires the choice of a tuning parameter λ in the estimate $\widehat{\pi}(\lambda)$, which leads to an inherent bias-variance trade-off. We refer to Section 9 in Storey (2002) and Section 6 in Storey, Taylor and Siegmund (2004) for two data-driven methods for automatically choosing λ.

3.3. *Bootstrap calibration.* When the sample size is large, it is suitable to use the normal distribution for calibration. Here we consider bootstrap calibration, which has been widely used due to its good numerical performance when the sample size is relatively small. In particular, we focus on the weighted bootstrap procedure [Barbe and Bertail (1995)].

The weighted bootstrap perturbs the objective function of an M-estimator with i.i.d. weights. Let W be a random variable with unit mean and variance, i.e., $\mathbb{E}(W) = 1$ and $\mathrm{var}(W) = 1$. Let $\{W_{ij,b}, 1 \le i \le n, 1 \le j \le p, 1 \le b \le B\}$ be i.i.d. random samples from W that are independent of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, where B is the number of bootstrap replications. For each j, the bootstrap counterparts of $(\widehat{\mu}_j, \widehat{\boldsymbol{b}}_j^{\mathrm{T}})^{\mathrm{T}}$ given in (3.2) are defined by

$$(\widehat{\mu}_{j,b}^*, (\widehat{\boldsymbol{b}}_{j,b}^*)^{\mathrm{T}})^{\mathrm{T}} \in \operatorname*{argmin}_{\mu \in \mathbb{R}, \boldsymbol{b} \in \mathbb{R}^d} \sum_{i=1}^n W_{ij,b}\, \ell_\tau(X_{ij} - \mu - \boldsymbol{b}^{\mathrm{T}} \boldsymbol{f}_i), \quad b = 1, \ldots, B.$$

For $j = 1, \ldots, p$, define empirical tail distributions

$$G_{j,B}^*(z) = \frac{1}{B+1} \sum_{b=1}^B 1\big(|\widehat{\mu}_{j,b}^* - \widehat{\mu}_j| \ge z\big), \quad z \ge 0.$$

The bootstrap P-values are thus given by $\{P_j^* = G_{j,B}^*(|\widehat{\mu}_j|)\}_{j=1}^p$, to which either the B-H procedure or Storey's procedure can be applied. For the former, we reject H_{0j} whenever $P_j^* \le P_{(k_p^*)}^*$, where $k_p^* = \max\{j : 1 \le j \le p, P_{(j)}^* \le j\alpha/p\}$ for a predetermined $0 < \alpha < 1$ and $P_{(1)}^* \le \cdots \le P_{(p)}^*$ are the ordered bootstrap P-values. For the distribution of the bootstrap weights, it is common to choose $W \sim 2\mathrm{Bernoulli}(0.5)$ or $W \sim N(1,1)$ in practice. The nonnegative random variables have the advantage that the objective function is convex.

REMARK 3.1. Weighted bootstrap procedure serves as an alternative method to normal calibration in multiple testing. We refer to Spokoiny and

Zhilova (2015) and Zhilova (2016) for the most advanced recent results of weighted bootstrap as well as a comprehensive literature review, and leave the theoretical guarantee of this procedure for future research.

3.4. *Theoretical properties.* First, we impose some conditions on the distribution of $\boldsymbol{X}$ and the tuning parameters τ and γ that are used in the robust regression and robust estimation of the second moment.

(C1). (i) $\boldsymbol{X} \in \mathbb{R}^p$ follows the model (1.6) with $\boldsymbol{f}$ and $\boldsymbol{u}$ being independent; (ii) $\mathbb{E}(u_j) = 0$, $\mathbb{E}(u_j^2) = 1$ for $j = 1, \ldots, p$, and $c_v \le \min_{1\le j\le p} \sigma_{\nu,jj} \le \max_{1\le j\le p} \mathbb{E}(\nu_j^4) \le C_v$ for some $C_\nu > c_\nu > 0$; (iii) $\mathbb{E}(\boldsymbol{f}) = \mathbf{0}$, $\boldsymbol{\Sigma}_f = \mathrm{cov}(\boldsymbol{f})$ is positive definite and $\|\boldsymbol{\Sigma}_f^{-1/2}\boldsymbol{f}\|_{\psi_2} \le C_f$ for some $C_f > 0$.

(C2). $(\tau, \gamma) = (\tau_n, \gamma_n)$ satisfies $\tau = \tau_0\sqrt{n}w_n^{-1/2}$ and $\gamma = \gamma_0\sqrt{n}w_n^{-1/2}$ for some constants $\tau_0 \ge \max_{1\le j\le p} \sigma_{\nu,jj}^{1/2}$ and $\gamma_0 \ge \max_{1\le j\le p} \mathrm{var}^{1/2}(X_j^2)$, where the sequence w_n is such that $w_n \to \infty$ and $w_n = o(\sqrt{n})$.

In addition, we need the following assumptions on the covariance structure of $\boldsymbol{\nu} = \boldsymbol{\Sigma}(\boldsymbol{f})\boldsymbol{u}$, and the number and magnitudes of the signals (nonzero coordinates of $\boldsymbol{\mu}$). Let $\mathbf{R}_\nu = (\rho_{\nu,jk})_{1\le j,k\le p}$ be the correlation matrix of $\boldsymbol{\nu}$, where by the independence of $\boldsymbol{f}$ and $\boldsymbol{u}$, $\rho_{\nu,jk} = \frac{\mathbb{E}\{\sigma_j(\boldsymbol{f})\sigma_k(\boldsymbol{f})\}}{\sqrt{\mathbb{E}\sigma_j^2(\boldsymbol{f})\mathbb{E}\sigma_k^2(\boldsymbol{f})}} \times \mathrm{corr}(u_j, u_k)$.

(C3). $\max_{1\le j<k\le p} |\rho_{\nu,jk}| \le \rho$ and

$$s_p := \max_{1\le j\le p} \sum_{k=1}^{p} 1\big\{|\rho_{\nu,jk}| > (\log p)^{-2-\kappa}\big\} = O(p^r)$$

for some $0 < \rho < 1$, $\kappa > 0$ and $0 < r < (1-\rho)/(1+\rho)$. As $n, p \to \infty$, $p_0/p \to \pi_0 \in (0, 1]$, $\log p = o(n^{1/5})$ and $w_n \asymp n^{1/5}$, where w_n is as in Condition (C2).

(C4). $\mathrm{Card}\big\{j : 1 \le j \le p, \sigma_{\nu,jj}^{-1/2}|\mu_j| \ge \lambda\sqrt{(\log p)/n}\big\} \to \infty$ as $n, p \to \infty$ for some $\lambda > 2\sqrt{2}$.

Condition (C3) allows weak dependence between $\nu_1, \ldots, \nu_p$ in the sense that each variable is moderately correlated with s_p other variables and weakly correlated with the remaining ones. The technical assumption (C4) imposes a constraint on the number of significant true alternatives, which is slightly stronger than $p_1 \to \infty$. According to Proposition 2.1 in Liu and Shao (2014), this condition is nearly optimal for the results on FDP control in the sense that if p_1 is fixed, the B-H method fails to control the FDP at any level $0 < \beta < 1$ with overwhelming probability even if the true P-values were known.

For robust test statistics T_j's (3.7), define the null distribution $F_{j,n}(x) = \mathbb{P}(T_j \le x|H_{0j})$ and the corresponding P-value $P_j^{\mathrm{true}} = F_{j,n}(-|T_j|) + 1 - F_{j,n}(|T_j|)$. In practice, we use $P_j = 2\Phi(-|T_j|)$ to estimate the true (unknown) P-values. A natural question is on how fast p can diverge with n so as to maintain valid simultaneous inference. This problem has been studied in Fan, Hall and Yao (2007), Kosorok and Ma (2007) and Liu and Shao (2010). There it is shown that the simple consistency $\max_{1\le j\le p}|P_j - P_j^{\mathrm{true}}| = o(1)$ is not enough, and the level of accuracy required must increase with n. More precisely, to secure a valid inference, we require

$$\max_{1\le j\le p}\left|\frac{P_j}{P_j^{\mathrm{true}}} - 1\right| 1\{\mathcal{S}_j\} = o(1) \quad \text{as } n\to\infty, \tag{3.13}$$

where $\mathcal{S}_j = \{P_j^{\mathrm{true}} > \alpha/p\}$, $j = 1,\ldots,p$.

THEOREM 3.1. *Assume that Conditions (C1) and (C2) hold and* $\log p = o\{\min(w_n, nw_n^{-2})\}$. *Then* (3.13) *holds.*

Theorem 3.1 shows that, to ensure the accuracy of the normal distribution calibration, the number of simultaneous tests can be as large as $\exp\{o(n^{1/3})\}$, when taking $w_n \asymp n^{1/3}$. We are also interested in estimating FDP in the high dimensional sparse setting, that is, p is large, but the number of $\mu_j \ne 0$ is relatively small. The following result indicates that $\mathrm{FDP}_{\mathrm{N}}(z)$ given in (3.9) provides a consistent estimator of the realized FDP in a uniform sense.

THEOREM 3.2. *Assume that Conditions (C1)–(C3) hold. Then, for any sequence of positive numbers* $m_p \le 2p$ *satisfying* $m_p \to \infty$, *we have as* $(n,p)\to\infty$,

$$\max_{0\le z\le \Phi^{-1}(1-m_p/(2p))}\left|\frac{\mathrm{FDP}(z)}{\mathrm{FDP}_{\mathrm{N}}(z)} - 1\right| \to 0 \quad \textit{in probability.} \tag{3.14}$$

Further, Theorem 3.3 shows that the proposed robust, dependence-adjusted inference procedure controls the FDP at a given level α asymptotically with P-values estimated from the standard normal distribution.

THEOREM 3.3. *Assume that Conditions (C1)–(C4) hold. Then, for any prespecified* $0 < \alpha < 1$,

$$(p_0/p)^{-1}\mathrm{FDP}(\widehat{z}_{\mathrm{N},0}) \to \alpha \quad \textit{in probability} \tag{3.15}$$

as $(n,p)\to\infty$, *where* $\widehat{z}_{\mathrm{N},0}$ *is defined in* (3.12).

Constraints on the tuning parameters (τ, γ) and the dimension p in Theorems 3.2 and 3.3 can be relaxed in a strict factor model with independent idiosyncratic errors.

(C5). $\nu_1, \ldots, \nu_p$ in model (1.6) are independent. As $n, p \to \infty$, $p_0/p \to \pi_0 \in (0, 1]$, $\log p = o(w_n)$ and $w_n = O(n^{1/3})$, where w_n is as in Condition (C2).

THEOREM 3.4. *Assume that Conditions (C1), (C2), (C4) and (C5) hold. Then, for any prespecified* $0 < \alpha < 1$, $(p_0/p)^{-1}\mathrm{FDP}(\widehat{z}_{\mathrm{N},0}) \to \alpha$ *in probability as* $(n, p) \to \infty$, *where* $\widehat{z}_{\mathrm{N},0}$ *is defined in* (3.12).

Theorems 3.2–3.4 provide theoretical guarantees on the FDP control for the B-H procedure with dependence-adjusted P-values $P_j = 2\Phi(-|T_j|)$, $j = 1, \ldots, p$. A similar approach can be defined by using the median-of-means approach, namely, replacing T_j's with S_j's in the definition of $\mathrm{FDP}_{\mathrm{N}}(z)$ in (3.9), which is equivalent to the B-H procedure with P-values $Q_j = 2\Phi(-|S_j|)$, $j = 1, \ldots, p$. Under similar conditions, the theoretical results on the FDP control remain valid.

THEOREM 3.5. *Let* $\mathrm{FDP}(z)$ *and* $\widehat{z}_{\mathrm{N},0}$ *be defined in* (3.8) *and* (3.12) *with* T_j*'s replaced by* S_j*'s, and let* $V = V_n$ *in* (3.5) *satisfy* $V \asymp w_n$ *for* w_n *as in Condition (C2). Moreover, let* $\tau = \tau_n$ *be as in Condition (C2).*

(i). Under Conditions (C1), (C3) and (C4), (3.15) *holds for any prespecified* $0 < \alpha < 1$.

(ii). Under Conditions (C1), (C4) and (C5), (3.15) *holds for any prespecified* $0 < \alpha < 1$.

4. Numerical study.

4.1. *Implementation.* To implement the proposed procedure, we solve the convex program (3.2) by using our own implementation in Matlab of the traditional method of scoring, which is an iterative method starting at an initial estimate $\widehat{\boldsymbol{\theta}}^0 \in \mathbb{R}^{K+1}$. Here, we take $\widehat{\boldsymbol{\theta}}^0 = \mathbf{0}$; using the current estimate $\widehat{\boldsymbol{\theta}}^t$ at iteration $t = 0, 1, 2, \ldots$, we update the estimate by the Newton-Raphson step:

$$\widehat{\boldsymbol{\theta}}^{t+1} = \widehat{\boldsymbol{\theta}}^t + \left\{ \frac{1}{n} \sum_{i=1}^n \ell''_\tau(\boldsymbol{z}_i^t) \right\}^{-1} (\mathbb{G}^{\mathrm{T}}\mathbb{G})^{-1}\mathbb{G}^{\mathrm{T}}(\ell'_\tau(\boldsymbol{z}_1^t), \ldots, \ell'_\tau(\boldsymbol{z}_n^t))^{\mathrm{T}},$$

where $\boldsymbol{z}^t = (X_{1j}, \ldots, X_{nj})^{\mathrm{T}} - \mathbb{G}\widehat{\boldsymbol{\theta}}^t$ and $\mathbb{G} = (\boldsymbol{g}_1, \ldots, \boldsymbol{g}_n)^{\mathrm{T}} \in \mathbb{R}^{n \times (K+1)}$ with $\boldsymbol{g}_i = (1, \boldsymbol{f}_i^{\mathrm{T}})^{\mathrm{T}}$.

For each $1 \leq j \leq p$, we apply the above algorithm with $\tau = \tau_j := c\,\widehat{\sigma}_j\sqrt{n/\log(np)}$ to obtain $(\widehat{\mu}_j, \widehat{\boldsymbol{b}}_j^{\mathrm{T}})^{\mathrm{T}}$, where $\widehat{\sigma}_j^2$ denotes the sample variance of the fitted residuals using OLS and $c > 0$ is a control parameter. We take $c = 2$ in all the simulations reported below. In practice, we can use a cross-validation procedure to pick c from only a few candidates, say $\{0.5, 1, 2\}$.

4.2. *Simulations via a synthetic factor model.* In this section, we perform Monte Carlo simulations to illustrate the performance of the robust test statistic under approximate factor models with general errors. Consider the Fama-French three factor model:

$$X_{ij} = \mu_j + \boldsymbol{b}_j^{\mathrm{T}}\boldsymbol{f}_i + u_{ij}, \quad i = 1, \ldots, n, \tag{4.1}$$

where $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{ip})^{\mathrm{T}}$ are i.i.d. copies of $\boldsymbol{u} = (u_1, \ldots, u_p)^{\mathrm{T}}$. We simulate $\{\boldsymbol{b}_j\}_{j=1}^p$ and $\{\boldsymbol{f}_i\}_{i=1}^n$ independently from $N_3(\boldsymbol{\mu}_B, \boldsymbol{\Sigma}_B)$ and $N_3(\mathbf{0}, \boldsymbol{\Sigma}_f)$, respectively. To make the model more realistic, parameters are calibrated from the daily returns of S&P 500's top 100 constituents (chosen by market cap), for the period July 1st, 2008 to June 29th, 2012.

To generate dependent errors, we set $\boldsymbol{\Sigma}_u = \mathrm{cov}(\boldsymbol{u})$ to be a block diagonal matrix where each block is four-by-four correlation matrix with equal off-diagonal entries generated from Uniform$[0, 0.5]$. The hypothesis testing is carried out under the alternative: $\mu_j = \mu$ for $1 \leq j \leq \pi_1 p$ and $\mu_j = 0$ otherwise. In the simulations reported here, the ambient dimension $p = 2000$, the proportion of true alternatives $\pi_1 = 0.25$ and the sample size n takes values in $\{80, 120\}$. For simplicity, we set $\lambda = 0.5$ in our procedure and use the Matlab package `mafdr` to compute the estimate $\widehat{\pi}_0(\lambda)$ of $\pi_0 = 1 - \pi_1$. For each test, the empirical false discovery rate (FDR) is calculated based on 500 replications with FDR level α taking values in $\{5\%, 10\%, 20\%\}$. The errors $\{\boldsymbol{u}_i\}_{i=1}^n$ are generated independently from the following distributions:

- **Model 1**. $\boldsymbol{u} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_u)$: Centered normal random errors with covariance matrix $\boldsymbol{\Sigma}_u$;
- **Model 2**. $\boldsymbol{u} \sim (1/\sqrt{5})\, t_{2.5}(\mathbf{0}, \boldsymbol{\Sigma}_u)$: Symmetric and heavy-tailed errors following a multivariate t-distribution with degrees of freedom 2.5 and covariance matrix $\boldsymbol{\Sigma}_u$;
- **Model 3**. $\boldsymbol{u} = 0.5\boldsymbol{u}_{\mathrm{N}} + 0.5(\boldsymbol{u}_{\mathrm{LN}} - \mathbb{E}\boldsymbol{u}_{\mathrm{LN}})$, where $\boldsymbol{u}_{\mathrm{N}} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_u)$ and $\boldsymbol{u}_{\mathrm{LN}} \sim \exp\{N(\mathbf{0}, \boldsymbol{\Sigma}_u)\}$ is independent of $\boldsymbol{u}_{\mathrm{N}}$. This model admits asymmetric and heavy-tailed errors;
- **Model 4**. $\boldsymbol{u} = 0.25\boldsymbol{u}_t + 0.75(\boldsymbol{u}_{\mathrm{W}} - \mathbb{E}\boldsymbol{u}_{\mathrm{W}})$, where $\boldsymbol{u}_t \sim t_4(\mathbf{0}, \boldsymbol{\Sigma}_u)$ and the p coordinates of $\boldsymbol{u}_{\mathrm{W}}$ are i.i.d. random variables following the Weibull distribution with shape parameter 0.75 and scale parameter 0.75.

The proposed Robust Dependence-Adjusted (RD-A) testing procedure is compared with the Ordinary Dependence-Adjusted (OD-A) procedure that uses OLS to estimate unknown parameters in the factor model, and also with the naive procedure where we directly perform multiple marginal t-tests ignoring the common factors. We use RD-$\mathrm{A_N}$ and RD-$\mathrm{A_B}$ to denote the RD-A procedure with normal and bootstrap calibration. The number of bootstrap replications is set to be $B = 2000$. The signal strength μ is taken to be $\sqrt{2(\log p)/n}$ for Models 1, 2 and 4, and $\sqrt{3(\log p)/n}$ for Model 3. Define the false negative rate FNR $= \mathbb{E}\{T/(p - R)\}$, where T is the number of falsely accepted null hypotheses and R is the number of discoveries. The true positive rate (TPR) is defined as the average ratio between the number of correct rejections and $p_1 = \pi_1 p$. Empirical FDR, FNR and TPR for the RD-$\mathrm{A_B}$, RD-$\mathrm{A_N}$, OD-A and naive procedures under different scenarios are presented in Tables 1 and 2. To save space, we leave the numerical comparison between the RD-$\mathrm{A_N}$ and OD-A procedures for some additional models in the supplementary material [Zhou et al. (2017)], along with comparisons across a range of sample sizes and signal strengths.

For weakly dependent errors following the normal distribution and t-distribution, Table 1 shows that the RD-A procedure consistently outperforms the OD-A method, in the sense that the RD-A method provides a much better control of the FDR at the expense of slight compromises of the FNR and TPR. Should the FDR being controlled at the same level, the robust method will be more powerful. In Table 2, when the errors are both asymmetric and heavy-tailed, the RD-A procedure has the biggest advantage in that it significantly outperforms the OD-A on controlling the FDR at all levels while maintaining low FNR and high TPR. Together, these results show that the RD-A procedure is indeed robust to outliers and does not lose efficiency when the errors are symmetric and light-tailed. In terms of controlling FDR, Models 3 and 4 present more challenges than Models 1 and 2 due to being both heavy-tailed and asymmetric. In Table 2 we see that although both the RD-A and OD-A methods achieve near-perfect power, the empirical FDR is higher than the desired level across all settings and much more higher for OD-A. Hence we compare the FDR of the RD-A and OD-A methods for various sample sizes in Figure 2. We see that the empirical FDR decreases with increase in sample size, while consistently outperforming the OD-A procedure. The difference between the two methods is greater for lower sample sizes, reinforcing the usefulness of our method for high dimensional heavy-tailed data with moderate sample sizes.

The naive procedure suffers from a significant loss in FNR and TPR. The reasons are twofold: (a) the naive procedure ignores the actual dependency

		Normal					
		$n = 80$			$n = 120$		
		α =5%	10%	20%	5%	10%	20%
FDR	RD-A$_{\mathrm{B}}$	3.66%	7.79%	16.64%	4.10%	8.45%	17.67%
	RD-A$_{\mathrm{N}}$	6.22%	11.61%	21.86%	5.69%	10.93%	21.05%
	OD-A	7.51%	13.70%	24.92%	6.49%	12.24%	23.01%
	Naive	8.35%	13.35%	19.40%	7.61%	10.97%	17.67%
FNR	RD-A$_{\mathrm{B}}$	3.87%	2.17%	0.99%	3.12%	1.72%	0.79%
	RD-A$_{\mathrm{N}}$	2.65%	1.49%	0.68%	2.44%	1.36%	0.62%
	OD-A	2.16%	1.21%	0.55%	2.13%	1.19%	0.53%
	Naive	20.10%	19.29%	18.61%	20.25%	19.17%	18.26%
TPR	RD-A$_{\mathrm{B}}$	88.05%	93.50%	97.18%	90.45%	94.89%	97.79%
	RD-A$_{\mathrm{N}}$	91.99%	95.64%	98.13%	92.64%	96.03%	98.30%
	OD-A	93.53%	96.50%	98.53%	93.59%	96.55%	98.57%
	Naive	22.71%	28.80%	37.63%	21.96%	28.20%	36.79%
		Student's t					
		$n = 80$			$n = 120$		
		α =5%	10%	20%	5%	10%	20%
FDR	RD-A$_{\mathrm{B}}$	2.98%	6.76%	15.36%	3.69%	7.89%	17.01%
	RD-A$_{\mathrm{N}}$	4.22%	8.72%	17.99%	4.26%	8.83%	18.44%
	OD-A	6.43%	12.72%	25.17%	5.77%	11.64%	23.63%
	Naive	10.05%	13.36%	19.45%	7.52%	11.07%	17.50%
FNR	RD-A$_{\mathrm{B}}$	1.97%	1.31%	0.80%	1.58%	1.07%	0.68%
	RD-A$_{\mathrm{N}}$	2.17%	1.51%	0.97%	2.00%	1.40%	0.91%
	OD-A	1.93%	1.39%	1.00%	1.86%	1.34%	0.93%
	Naive	19.75%	19.12%	18.71%	19.93%	18.99%	18.46%
TPR	RD-A$_{\mathrm{B}}$	93.79%	95.95%	97.60%	95.03%	96.72%	97.98%
	RD-A$_{\mathrm{N}}$	93.04%	95.23%	97.03%	93.55%	95.54%	97.18%
	OD-A	93.91%	95.74%	97.47%	94.09%	95.86%	97.43%
	Naive	25.45%	31.95%	41.39%	23.63%	29.70%	38.81%

TABLE 1
Empirical FDR, FNR and TPR based on a factor model with dependent errors following a normal distribution (Model 1) and a t-distribution (Model 2).

structure among the variables; (b) the signal-to-noise ratio of $H_{0j} : \mu_j = 0$ for the naive procedure is $\sigma_{jj}^{-1/2}|\mu_j|$, which can be much smaller than $\sigma_{\nu,jj}^{-1/2}|\mu_j|$ for the dependence-adjusted procedure.

4.3. *Stock market data.* In this section, we apply our proposed robust, dependence-adjusted multiple testing procedure to monthly stock market data. Consider Carhart's four-factor model [Carhart (1997)] on S&P 500

		Mixture Normal/Lognormal					
		$n = 80$			$n = 120$		
		α =5%	10%	20%	5%	10%	20%
FDR	RD-A_B	8.40%	13.33%	22.04%	8.02%	12.98%	21.99%
	RD-A_N	10.29%	16.02%	25.87%	9.16%	14.65%	24.61%
	OD-A	12.18%	18.46%	29.24%	10.28%	16.17%	26.79%
	Naive	7.79%	12.18%	18.49%	7.99%	11.77%	17.93%
FNR	RD-A_B	0.32%	0.09%	0.02%	0.29%	0.11%	0.03%
	RD-A_N	0.57%	0.25%	0.07%	0.56%	0.23%	0.06%
	OD-A	0.26%	0.08%	0.02%	0.29%	0.10%	0.03%
	Naive	18.66%	16.99%	15.26%	18.51%	16.67%	14.78%
TPR	RD-A_B	99.07%	99.73%	99.95%	99.15%	99.70%	99.92%
	RD-A_N	98.36%	99.31%	99.81%	98.38%	99.36%	99.82%
	OD-A	99.26%	99.76%	99.96%	99.16%	99.72%	99.93%
	Naive	28.39%	37.37%	50.14%	29.72%	39.34%	51.95%
		Mixture Student's t/Weibull					
		$n = 80$			$n = 120$		
		α =5%	10%	20%	5%	10%	20%
FDR	RD-A_B	7.51%	12.05%	20.43%	7.03%	11.66%	20.49%
	RD-A_N	8.91%	14.34%	24.11%	7.85%	13.17%	23.11%
	OD-A	11.00%	17.11%	27.67%	9.09%	14.80%	25.37%
	Naive	9.33%	13.48%	20.30%	8.14%	12.52%	18.85%
FNR	RD-A_B	0.62%	0.23%	0.06%	0.62%	0.25%	0.08%
	RD-A_N	0.62%	0.24%	0.06%	0.60%	0.24%	0.07%
	OD-A	0.30%	0.11%	0.03%	0.42%	0.16%	0.05%
	Naive	19.91%	18.78%	18.08%	20.71%	20.02%	19.36%
TPR	RD-A_B	98.16%	99.34%	99.83%	98.16%	99.29%	99.78%
	RD-A_N	98.18%	99.31%	99.82%	98.22%	99.30%	99.80%
	OD-A	99.14%	99.69%	99.92%	98.78%	99.54%	99.86%
	Naive	23.41%	30.24%	39.65%	20.49%	25.90%	34.07%

TABLE 2

Empirical FDR, FNR and TPR based on a factor model with dependent errors following a mixture normal/lognormal distribution (Model 3) and a mixture Student's t/Weibull distribution (Model 4).

index, where the excess return of a stock has the following representation:

$$r_{jt} = \mu_j + \beta_{j,\mathrm{MKT}}(\mathrm{MKT}_t - r_{ft}) + \beta_{j,\mathrm{SMB}}\mathrm{SMB}_t + \beta_{j,\mathrm{HML}}\mathrm{HML}_t + \beta_{j,\mathrm{UMD}}\mathrm{UMD}_t + u_{jt}, \tag{4.2}$$

for $j = 1, \dots, p$ and $t = 1, \dots, T$. Here r_{jt} is the excess returns of stock j at month t, r_{ft} is the risk free interest rate at month t, and MKT, HML,

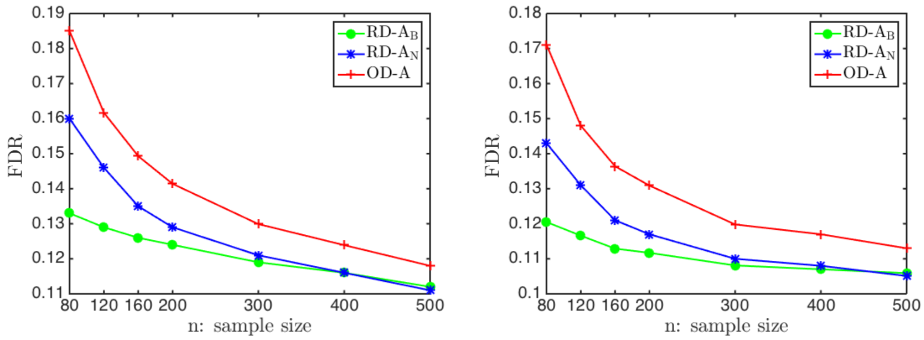

FIG 2. *Empirical FDR of the testing problem at the* 10% *significance level, when the data follows a mixture normal/lognormal distribution (Model 3) on the left panel and a mixture Student's t/Weibull distribution (Model 4) on the right panel, and the sample size varies.*

SMB and UMD represent the market, size, value and momentum factors respectively. We are interested in the value μ_j, which represents the *alpha* of stock j. A stock can be said to have excess returns if its alpha is positive, or in other words, the stock exhibit returns higher than those that can be accounted for by the four factors. If the alpha is negative, the stock is consistently underperforming, given the level of risk it undertakes. Detecting nonzero alpha is important since it is directly related to the efficient equity market hypothesis. When the market is inefficient, we can conduct multiple hypothesis testing to identify those stocks in the market that have statistically significant alphas. When the returns of mutual fund data are used, the test is related to test whether the fund manager has skills or not [Barras, Scaillet and Wermers (2010)]. All the data in this section was obtained from Kenneth French's website and the COMPUSTAT and CRSP databases.

We obtain monthly data for 393 S&P 500 constituents over the time period from January 2005 to December 2013, after removing those stocks that have missing values or have discontinuous inclusion in the index. The stock returns exhibit severely heavy-tails, as illustrated by the histogram of the excess kurtosis of the data in Figure 3. Among the 393 series, 112 have distributions whose tails are fatter than the t-distribution with 5 degrees of freedom.

The regression in (4.2) is carried out over rolling windows: for each month, we evaluate the model using data from the preceding three years. For each rolling window, we simultaneously test the hypotheses $H_{0j} : \mu_j = 0$ versus $H_{1j} : \mu_j \neq 0$ for $j = 1, \ldots, p$, using the proposed robust dependence-adjusted procedure. We see that out of a portfolio of size 393, only a few stocks ex-

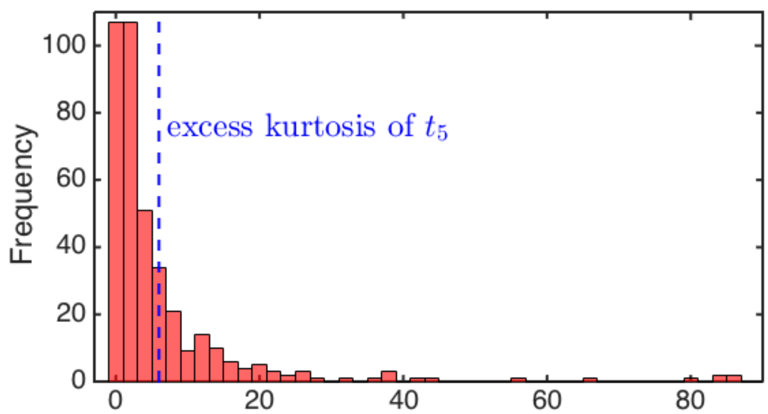

Fig 3. *Histogram of excess kurtosis of monthly returns of* 393 *S&P 500 constituents from January 2005 to December 2013. The excess kurtosis of the t_5 distribution is shown for reference.*

hibit statistically significant nonzero alphas at the FDR threshold of 5%, 10% and 20%. Table 3 summarizes the results for the number of selected stocks and estimated alphas with the FDR controlled at 5%. For the robust dependence-adjusted procedure, we see that no stock is selected on average, with maximum 2 stocks selected over the entire time period. In particular, our method does not select any stocks from the third quarter of 2008 to the third quarter of 2010, coinciding with the financial crisis during which the market volatility is much higher. Moreover, the estimated alphas for the selected stocks are much higher than those not selected by the robust multiple testing procedure. This is represented as $|\widehat{\mu}_j|$ in Table 3. The naive method, which directly performs multiple t-tests ignoring the common factors, appears to be unstable with the number of stocks selected being extremely variable. Additionally, a tremendously large number of stocks are selected in a few time periods, pointing towards false discoveries. In summary, the robust, dependence-adjusted multiple testing procedure is particularly suited for the problem of finding a few stocks with nonzero alphas, which is explained by the focus on a balanced panel of highly traded stocks with large capitalizations, namely, the constituents of the S&P 500.

4.4. *Gene expression data.* In this section, we apply the proposed procedure to the analysis of a neuroblastoma data set reported in Oberthuer et al. (2006) to identify differentially expressed genes between the group of patients who had 3-year event-free survival after the diagnosis of neurob-

Variable	Method	Mean	Std. Dev.	Median	Min	Max
Number of selected stocks	R-DA$_\mathrm{N}$	0.18	0.42	0	0	2
	OD-A	0.94	1.59	0	0	7
	Naive	4.39	22.46	0	0	178
$\lvert\widehat{\mu}_j\rvert$ for selected stocks	R-DA$_\mathrm{N}$	4.19%	1.41%	4.31%	2.43%	6.58%
	OD-A	3.37%	1.00%	3.24%	1.70%	6.34%
	Naive	3.76%	1.20%	3.47%	2.77%	8.01%
$\lvert\widehat{\mu}_j\rvert$ for non-selected stocks	R-DA$_\mathrm{N}$	1.01%	0.09%	1.00%	0.90%	1.25%
	OD-A	1.09%	0.11%	1.08%	0.91%	1.32%
	Naive	1.75%	0.35%	1.63%	1.35%	2.45%

TABLE 3

Summary of the three testing procedures based on 393 *stocks in S&P 500 between January 2005 and December 2013. A rolling window of 3 years is used for estimation and selecting stocks with significant nonzero alpha. The stocks are selected at FDR level* 5%.

lastoma and the group of patients who did not. This data set consists of 251 patients of the German Neuroblastoma Trials NB90-NB2004, diagnosed between 1989 and 2004. The complete data set, obtained via the MicroArray Quality Control phase-II (MAQC-II) project [Shi et al. (2012)], includes gene expression over 10,707 probe sites. There are 246 subjects with 3-year event-free survival information available (56 positive and 190 negative). See Oberthuer et al. (2006) for more details about the data sets.

In the first stage, we use standard principal component analysis on the two samples to obtain the factors, based on which we construct dependence-adjusted P-values to conduct multiple testing in the second step. Note that the test statistic given in (3.7) can be directly generalized to the two-sample case: Given two groups of p-dimensional ($p = 10,707$) observations with sizes $n_1 = 56$ and $n_2 = 190$, we compute robust mean and variance estimators $(\widehat{\mu}_{1j}, \widehat{\mu}_{2j})$ and $(\widehat{\sigma}_{1\nu,jj}, \widehat{\sigma}_{2\nu,jj})$ for $j = 1, \ldots, p$. Define two-sample test statistics $T_j = (\widehat{\mu}_{1j} - \widehat{\mu}_{2j})/(\widehat{\sigma}_{1\nu,jj}/n_1 + \widehat{\sigma}_{2\nu,jj}/n_2)^{1/2}$ so that the corresponding P-values are $\{2\Phi(-|T_j|)\}_{j=1}^p$.

The number of factors is estimated by the eigenvalue ratio estimator proposed in Ahn and Horenstein (2013), which was also used in the context of factor-adjusted multiple testing in Fan and Han (2017). The estimator is defined as $\widehat{K} = \mathrm{argmax}_{1<k<k_{\max}}(\widehat{\lambda}_k/\widehat{\lambda}_{k+1})$, where $\widehat{\lambda}_j$ is the jth eigenvalue of the sample covariance matrix and $k_{\max}$ is the maximum possible number of factors. Following this procedure, we use $K = 2$ to model the latent structure in the data.

Next, we conduct multiple testing using the proposed robust, dependence-adjusted procedure and the naive procedure based on two-sample t-tests.

At FDR level 1%, we detect 3779 genes and the naive procedure detects 3236 genes; while at FDR level 5%, we discover 5223 genes and the naive procedure discovers 4685 genes. In general, taking the latent structure into account causes a visible increase in the number of genes that are declared statistically significant regardless of the prechosen FDR level, reflecting the improved power of our method. This phenomenon is in accord with that in Desai and Storey (2012). These results may serve as an exploratory step for more refined analyses regarding those significant genes.

5. Summary and discussion. This paper consists of two main parts with each one being of independent interest. In the first part, we study the conventional robust M-estimation [Huber (1973)] from a new perspective by allowing the robustification parameter τ to diverge with the sample size to balance the bias and robustness of the estimator. Our main theoretical contribution (Theorem 2.1) is a nonasymptotic Bahadur representation of the proposed robust estimator along with a sub-Gaussian-type deviation bound if the error variable has a finite second moment. As by-products, we prove the Berry-Esseen inequality and a Cramér-type moderate deviation theorem for the estimator. These probabilistic results are particularly useful in investigating robustness and accuracy of the P-values in multiple testing, among other high dimensional statistical inference problems [Fan, Hall and Yao (2007), Delaigle, Hall and Jin (2011)].

In the second part, we focus on large-scale multiple testing for dependent and heavy-tailed data. To characterize the dependence, we employ a multi-factor model similar to that used in Desai and Storey (2012), Fan, Han and Gu (2012) and Fan and Han (2017) but with an observable factor. To achieve robustness, we propose a Huber loss based approach to construct test statistics for testing the individual hypotheses. Under mild conditions, our procedure asymptotically controls the overall false discovery proportion at the nominal level. Thorough numerical results on both simulated and real world datasets are also provided to back up our theory. It is shown that the newly proposed robust, dependence-adjusted method performs well numerically in terms of both the size and power. It significantly outperforms the multiple t-tests under strong dependence, and is applicable even when the true error distribution deviates wildly from the normal distribution. A more interesting and challenging problem is when the dependence structure is characterized by latent factors. In this case, robust estimators of the unobservable factors along with the loadings are required. Large-scale simultaneous inference for latent factor models with heavy-tailed errors is our ongoing work. We leave the details of the results elsewhere in the future.

Acknowledgements. We would like to thank the Editor, Dr. Tailen Hsing, the Associate Editor and three reviewers for their constructive comments which helped us improve the paper, and Dr. Qiang Sun for helpful discussions.

SUPPLEMENTARY MATERIAL

Supplement to "A New Perspective on Robust M-Estimation: Finite Sample Theory and Applications to Dependence-Adjusted Multiple Testing"
(http://www.e-publications.org/ims/support/dowload/imsart-ims.zip). This supplemental material contains the proofs for the theoretical results in the main text and additional simulation results.

References.

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

Barbe, P. and Bertail, P. (1995). *The Weighted Bootstrap.* Springer-Verlag, New York.

Barras, L., Scaillet, O. and Wermers, R. (2010). False discoveries in mutual fund performance: Measuring luck in estimated alphas. *J. Finance* **65**, 179–216.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **57** 289–300.

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

Carhart, M. M. (1997). On persistence in mutual fund performance. *J. Finance* **52** 57–82.

Catoni, O. (2012). Challenging the empirical mean and empirical variance: A deviation study. *Ann. Inst. Henri Poincaré Probab. Stat.* **48** 1148–1185.

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

Cont, R. (2001). Empirical properties of asset returns: Stylized facts and statistical issues. *Quant. Finance* **1** 223–236.

Delaigle, A., Hall, P. and Jin, J. (2011). Robustness and accuracy of methods for high dimensional data analysis based on Student's t-statistic. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 283–301.

Desai, K. H. and Storey, J. D. (2012). Cross-dimensional inference of dependent high-dimensional data. *J. Amer. Statist. Assoc.* **107** 135–151.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *J. Amer. Statist. Assoc.* **99** 96–104.

Efron, B. (2007). Correlation and large-scale simultaneous significance testing. *J. Amer. Statist. Assoc.* **102** 93–103.

Efron, B., Tibshirani, R., Storey, J. D. and Tusher, V. G. (2001). Empirical bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.* **96** 1151–1160.

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

FAN, J., HALL, P. and YAO, Q. (2007). To how many simultaneous hypothesis tests can normal, Student's t or bootstrap calibration be applied? *J. Amer. Statist. Assoc.* **102** 1282–1288.

FAN, J. and HAN, X. (2017). Estimation of the false discovery proportion with unknown dependence. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* To appear. DOI: 10.1111/rssb.12204.

FAN, J., HAN, X. and GU, W. (2012). Estimating false discovery proportion under arbitrary covariance dependence. *J. Amer. Statist. Assoc.* **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. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 247–265.

FINKENSTADT, B. and ROOTZEÉN, H. (2003). *Extreme Values in Finance, Telecommunications and the Environment.* Chapman & Hall, New York.

FRIGUET, C., KLOAREG, M. and CAUSEUR, D. (2009). A factor model approach to multiple testing under dependence. *J. Amer. Statist. Assoc.* **104** 1406–1415.

GENOVESE, C. and WASSERMAN, L. (2004). A stochastic process approach to false discovery control. *Ann. Statist.* **32** 1035–1061.

HE, X. and SHAO, Q.-M. (1996). A general Bahadur representation of M-estimators and its application to linear regression with nonstochastic designs. *Ann. Statist.* **24** 2608–2630.

HE, X. and SHAO, Q.-M. (2000). On parameters of increasing dimensions. *J. Multivariate Anal.* **73** 120–135.

HUBER, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35** 73–101.

HUBER, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *Ann. Statist.* **1** 799–821.

JOLY, E. and LUGOSI, G. (2016). Robust estimation of U-statistics. *Stochastic Process. Appl.* **126** 3760–3773.

KOSOROK, M. R. and MA, S. (2007). Marginal asymptotics for the large p, small n paradigm: With applications to microarray data. *Ann. Statist.* **35** 1456–1486.

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

LEEK, J. T. and STOREY, J. D. (2008). A general framework for multiple testing dependence. *Proc. Natl. Acad. Sci. USA* **105** 18718–18723.

LEHMANN, E. L. and ROMANO, J. P. (2005). Generalizations of the familywise error rate. *Ann. Statist.* **33** 1138–1154.

LINNIK, J. V. (1961). On the probability of large deviations for the sums of independent variables. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. II* 289–306. Univ. California Press, Berkeley, CA.

LIU, W. and SHAO, Q.-M. (2010). Cramér-type moderate deviation for the maximum of the priodogram with application to simultaneous tests in gene expression time series. *Ann. Statist.* **38** 1913–1935.

LIU, W. and SHAO, Q.-M. (2014). Phase transition and regularized bootstrap in large-scale t-tests with false discovery rate control. *Ann. Statist.* **42** 2003–2025.

MAMMEN, E. (1989). Asymptotics with increasing dimension for robust regression with applications to the bootstrap. *Ann. Statist.* **17** 382–400.

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

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

NEMIROVSKY, A. S. and YUDIN, D. B. (1983). *Problem Complexity and Method Efficiency in Optimization.* Wiley, New York.

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. *J. Clin. Oncol.* **24** 5070–5078.

PORTNOY, S. (1985). Asymptotic behavior of M estimators of p regression parameters when p^2/n is large; II. Normal approximation. *Ann. Statist.* **13** 1403–1417.

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

SHI, L. ET AL. (MAQC Consortium) (2010). The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nature Biotechnology* **28** 827–841.

SPOKOINY, V. and ZHILOVA, M. (2015). Bootstrap confidence sets under model misspecification. *Ann. Statist.* **43** 2653–2675.

STOREY, J. D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **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. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** 187–205.

SUN, W. and CAI, T. T. (2009). Large-scale multiple testing under dependence. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 393–424.

VERSHYNIN, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing* 210–268. Cambridge Univ. Press, Cambridge.

YOHAI, V. J. and MARONNA, R. A. (1979). Asymptotic behavior of M-estimators for the linear model. *Ann. Statist.* **7** 258–268.

ZHILOVA, M. (2016). Non-classical Berry-Esseen inequality and accuracy of the weighted bootstrap. Available at arXiv:1611.02686.

ZHOU, W.-X., BOSE, K., FAN, J. and LIU, H. (2017). Supplement to "A new perspective on robust M-estimation: Finite sample theory and applications to dependence-adjusted multiple testing."

DEPARTMENT OF OPERATIONS RESEARCH
AND FINANCIAL ENGINEERING
PRINCETON UNIVERSITY
PRINCETON, NEW JERSEY 08544
USA
E-MAIL: wenxinz@princeton.edu
bose@princeton.edu
hanliu@princeton.edu

DEPARTMENT OF OPERATIONS RESEARCH
AND FINANCIAL ENGINEERING
PRINCETON UNIVERSITY
PRINCETON, NEW JERSEY 08544
USA
AND
SCHOOL OF DATA SCIENCE
FUDAN UNIVERSITY
SHANGHAI 200433
CHINA
E-MAIL: jqfan@princeton.edu