Statistica Sinica **33** (2023), 1831-1852
doi:https://doi.org/10.5705/ss.202021.0003

LARGE-SCALE INFERENCE OF MULTIVARIATE REGRESSION FOR HEAVY-TAILED AND ASYMMETRIC DATA

Youngseok Song, Wen Zhou and Wen-Xin Zhou

Ecole Polytechnique Fédérale de Lausanne, Colorado State University and University of California San Diego

Abstract: Large-scale multivariate regression is a fundamental statistical tool with a wide range of applications. This study considers the problem of simultaneously testing a large number of general linear hypotheses, encompassing covariate-effect analysis, analysis of variance, and model comparisons. The challenge that accompanies a large number of tests is the ubiquitous presence of heavy-tailed and/or highly skewed measurement noise, which is the main reason for the failure of conventional least squares-based methods. For large-scale multivariate regression, we develop a set of robust inference methods to explore data features such as heavy tailedness and skewness, which are not visible to least squares methods. The new testing procedure is based on the data-adaptive Huber regression and a new covariance estimator of regression estimates. Under mild conditions, we show that our methods produce consistent estimates of the false discovery proportion. Extensive numerical experiments and an empirical study on quantitative linguistics demonstrate the advantage of the proposed method over many state-of-the-art methods when the data are generated from heavy-tailed and/or skewed distributions.

Key words and phrases: General linear hypotheses, heavy-tailed and/or skewed regression errors, Huber loss, large-scale multiple testing, multivariate regression, quantitative linguistics.

1. Introduction

Multivariate regression is a fundamental statistical tool for data analysis with applications in fields including biology, financial economics, linguistics, psychology, and social science. By modeling thousands or tens of thousands of responses and covariates or experimental factors, it provides statistical decisions on individual levels by simultaneously testing many general linear hypotheses, including the covariate-effect analysis, analysis of variance, and model comparisons, among others. For example, multivariate regression has become a standard tool in dif-

Corresponding author: Wen Zhou, Department of Statistics, Colorado State University, Fort Collins, CO 80523, USA. E-mail: riczw@stat.colostate.edu.

ferential expression analyses in genomics (Ritchie et al. (2015)), and is commonly used in corpus linguistics for word usage comparisons (Khany and Tazik (2019)). See Cai and Sun (2017) for a comprehensive review of relevant applications.

To simultaneously test many general linear hypotheses, the conventional practice is to compute individual p-values based on F-tests or likelihood ratio tests, and then to employ multiple testing procedures to control the false discovery rate (FDR, see Benjamini and Hochberg (1995); Storey (2002)). However, this standard approach and its theoretical validity often rely on strong distributional assumptions, such as a normality/sub-Gaussianity or symmetry condition on the error distribution. Its effectiveness in terms of FDR control and power may be compromised when dealing with heavy-tailed and/or skewed data with large scales, such as microarray data (Purdom and Holmes (2005)) and text data (Zipf (1949)).

Overcoming this challenge requires a procedure that is robust against heavy-tailed and/or skewed error distributions. Heavy tailedness increases the chance of observing data that are more extreme than the majority. We refer to these outlying data points as stochastic outliers. A procedure that is robust against such outliers, as evidenced by its better finite-sample performance than that of a non-robust method, is called a *tail-robust procedure* (Ke et al. (2019)). In contrast to the conventional robustness under Huber's ϵ-contamination model (Huber (1964)) or the regularization-based robustness for detecting and removing outliers (Kong, Bondell and Wu (2018)), the notion of tail-robustness focuses on the challenge that methods that minimize the empirical risk perform poorly because the empirical risk is not uniformly close to the population risk, given the heavy-tailed and/or skewed errors (Prasad et al. (2020)). Although several new methods and estimation theory under heavy-tailed models have been developed (Catoni (2012); Minsker (2018); Sun, Zhou and Fan (2020)), fewer studies have focused on inference, especially in a large-scale setting (Fan et al. (2019); Minsker (2019)).

Building on the idea of the *adaptive Huber regression*, we develop a robust multiple testing procedure to test many general linear hypotheses in the presence of heavy-tailed and/or skewed errors. First, we employ the adaptive Huber regression to estimate the multivariate regression coefficients, based on which, we construct a robust test statistic and compute approximate p-values to estimate the false discovery proportion (FDP). Next, we apply Storey's FDR controlling procedure (Storey (2002)) to determine a threshold, below which the p-values lead to the corresponding hypotheses being rejected. By allowing the robustification parameter to diverge with the sample size, the adaptive Huber regression

estimator admits a tight non-asymptotic deviation bound and is asymptotically efficient (Sun, Zhou and Fan (2020)). Theoretically, the non-asymptotic Bahadur representation is a crucial step for establishing the limiting distribution of the estimator or its functionals. Practically, the proposed method is fully data-driven (Wang et al. (2021)), and therefore computationally attractive and applicable to large-scale problems.

The main contributions of this study are as follows. Methodologically, we develop a tail-robust multiple testing procedure to simultaneously draw inferences on large-scale multivariate regressions in the presence of heavy-tailed and/or skewed errors. This general framework includes the large-scale simultaneous mean testing problem as a special case. Compared with the traditional approach in multivariate and high-dimensional statistics, our method imposes very mild moment conditions on the data, and the number of hypotheses/responses is allowed to grow exponentially fast with the sample size. These features make our method particularly advantageous and appealing for conducting an inference on large-scale multivariate regression models with heavy-tailed and/or asymmetric errors, which is corroborated by our comprehensive simulation studies. Furthermore, motivated by Huber (1973), we propose a novel covariance estimator of the adaptive Huber regression estimate, and derive an interesting new exponential-type deviation bound that is of independent interest. The theoretical analysis of the new procedure is nontrivial. For that, we explore and develop interesting new technical results, by which we show that the proposed method controls the FDP and FDR asymptotically under mild moment and correlation conditions on the error vector. Computationally, our method is fast by taking advantage of the computational efficiency of the data-adaptive Huber regression (Wang et al. (2021)). In addition to numerical experiments, we apply our method to analyze text data from the Standardized Gutenberg Project Corpus (Gerlach and Font-Clos (2020)). We identify genre-representative words in works of William Shakespeare, and investigate the differences between the works of Lewis Carroll, Charles Dickens, and Arthur Conan Doyle. This empirical study demonstrates that our method is a useful addition to the existing toolkit for modeling and analyzing text data in quantitative linguistics.

The rest of the paper proceeds as follows. In Section 2, we revisit testing general linear hypotheses based on multivariate regressions, and introduce our procedure based on the adaptive Huber regression. In particular, we introduce a novel Huber-type estimator of the covariance of the regression coefficients in Section 2.2. We establish the statistical guarantees in Section 3. Section 4 presents our simulations. In Section 5, we apply our method to a well-known quantitative

linguistics data set, namely the Gutenberg Project. Extensions of our method are discussed in Section 6. All proofs and additional numerical results are provided in the Supplemental Material.

2. Model and Methodology

Throughout the paper, we write $\|\mathbf{u}\| = (\sum_{i=1}^{d} u_i^2)^{1/2}$ as the ℓ_2-norm of the vector $\mathbf{u} = (u_1, u_d)^{\mathrm{T}} \in \mathbb{R}^d$. Let $\langle \mathbf{u}, \mathbf{w} \rangle$ be the inner product of vectors $\mathbf{u}$ and $\mathbf{w}$ and $\|\mathbf{u}\|^2 = \langle \mathbf{u}, \mathbf{u} \rangle$. Denote $\mathbb{S}^{d-1} = \{\mathbf{u} \in \mathbb{R}^d : \|\mathbf{u}\| = 1\}$ as the unit sphere in $\mathbb{R}^d$. For the matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$, denote $\|\mathbf{A}\| = \sup_{\mathbf{u} \in \mathbb{S}^{d-1}} \|\mathbf{A}\mathbf{u}\|$, $\lambda_{\max}(\mathbf{A})$, and $\lambda_{\min}(\mathbf{A})$ as the spectral norm, maximum eigenvalue, and minimum eigenvalue, respectively. Let $\Phi(z) := \mathbb{P}(U < z)$, with $U \sim N(0, 1)$, be the cumulative distribution function of the standard normal distribution. Denote $\mathbb{I}(\cdot)$ as the indicator function.

Consider independent data $\{(\mathbf{Y}_i, \mathbf{X}_i)\}_{i=1}^n$, where $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{ip})^{\mathrm{T}}$, $\mathbf{X}_i = (X_{i1}, \ldots, X_{id})^{\mathrm{T}}$, with $d \geq 1$ and $d/n \to 0$ as $n \to \infty$. For each $1 \leq j \leq p$, the conditional expectation of Y_{ij} given $\mathbf{X}_i$ is modeled by $\mathbb{E}(Y_{ij}|\mathbf{X}_i) = \mu_j + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_j$. Define the data matrices $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)^{\mathrm{T}} \in \mathbb{R}^{n \times p}$ and $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^{\mathrm{T}} \in \mathbb{R}^{n \times d}$. The multivariate regression of interest is

$$\mathbf{Y} = \mathbf{1}_n \boldsymbol{\mu}^{\mathrm{T}} + \mathbf{X}\mathbf{B} + \boldsymbol{\Xi}, \tag{2.1}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^{\mathrm{T}}$ is the intercept vector, $\mathbf{1}_n = (1, \ldots, 1)^{\mathrm{T}} \in \mathbb{R}^n$, $\mathbf{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p) \in \mathbb{R}^{d \times p}$ consists of the slope coefficients, and $\boldsymbol{\Xi} = (\boldsymbol{\epsilon}_1^{\mathrm{T}}, \ldots, \boldsymbol{\epsilon}_n^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{n \times p}$, with $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{ip})^{\mathrm{T}}$. Independent of $\mathbf{X}_i$, the p-dimensional residual errors $\boldsymbol{\epsilon}_i$ are independent and identically distributed (i.i.d.), with mean zero and covariance matrix $\boldsymbol{\Sigma}_\epsilon = (\sigma_{\epsilon,jk})_{1 \leq j,k \leq p}$. For ease of notation, let $\boldsymbol{\theta}_j = (\mu_j, \boldsymbol{\beta}_j^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{d+1}$ and $\mathbf{Z}_i = (1, \mathbf{X}_i^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{d+1}$, and define the parameter and design matrix as $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p) \in \mathbb{R}^{(d+1) \times p}$ and $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_n)^{\mathrm{T}}$, respectively, so that (2.1) reduces to $\mathbf{Y} = \mathbf{Z}\boldsymbol{\Theta} + \boldsymbol{\Xi}$. Based on (2.1), we are interested in performing a simultaneous inference on the p hypotheses

$$H_{0j} : \mathbf{C}\boldsymbol{\theta}_j = \mathbf{c}_{0j} \quad \text{versus} \quad H_{1j} : \mathbf{C}\boldsymbol{\theta}_j \neq \mathbf{c}_{0j} \quad \text{for } j = 1, \ldots, p, \tag{2.2}$$

where the matrix $\mathbf{C} \in \mathbb{R}^{q \times (d+1)}$, the vectors $\mathbf{c}_{0j} \in \mathbb{R}^q$ are prescribed, and $\mathrm{rank}(\mathbf{C}) = q \leq d + 1$. The hypotheses in (2.2) encompass a variety of important applications, including inferences on contrasts in an analysis of variance and testing for treatment effects. Likelihood-based or least squares-based methods have been employed under the assumption that the covariates and/or errors follow ei-

ther normal or light-tailed symmetric distributions (Friguet, Kloareg and Causeur (2009)). With a large p, the underlying distributions, by chance alone, may have quite different scales, and can be highly skewed and heavy tailed. Therefore, outliers occur more frequently, challenging the efficacy of the standard methods. We make no parametric distributional assumptions, such as normality or elliptical symmetry. Instead, we define moment parameters $v_{j,\delta} = \{\mathbb{E}(|\epsilon_{1j}|^{2+\delta})\}^{1/(2+\delta)}$, for $\delta > 0$. Specifically, set $v_j = v_{j,2}$.

To test the linear hypotheses in (2.2), we first estimate the model parameters robustly in the presence of heavy-tailed and/or skewed errors. For $j = 1, \ldots, p$, define the Huber-type M-estimators $\widehat{\boldsymbol{\theta}}_j$ as

$$\widehat{\boldsymbol{\theta}}_j := (\widehat{\mu}_j, \widehat{\boldsymbol{\beta}}_j^{\mathrm{T}})^{\mathrm{T}} = \underset{\mu\in\mathbb{R},\boldsymbol{\beta}\in\mathbb{R}^d}{\operatorname{argmin}} \sum_{i=1}^{n} \ell_{\tau_j}(Y_{ij} - \mu - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}), \tag{2.3}$$

where $\ell_\tau(x) = (x^2/2)\mathbb{I}(|x| \le \tau) + (\tau|x| - \tau^2/2)\mathbb{I}(|x| > \tau)$ is the Huber loss (Huber (1964)), parameterized by $\tau > 0$. Our theoretical analysis suggests that with $\tau_j \asymp n^{1/(2+\delta)}\{\log(np) + d\}^{-1/(2+\delta)}$, for some $\delta > 0$, the estimators $\widehat{\boldsymbol{\theta}}_j$ are close to $\boldsymbol{\theta}_j$ uniformly over $j = 1, \ldots, p$ with high probability, even when p grows exponentially fast with n. Here, the divergence of τ_j guarantees $\widehat{\theta}_j$ to be sub-Gaussian, even if the error admits only a $(2 + \delta)$th finite moment. More importantly, the order of τ_j grants the desired approximation error of the Bahadur representation to $\widehat{\boldsymbol{\theta}}_j$ (Proposition 1), as well as the uniform non-asymptotic bounds of the estimated covariance of $\widehat{\boldsymbol{\theta}}_j$ (Theorem 2). As noted in the literature (Catoni (2012); Fan et al. (2019); Sun, Zhou and Fan (2020); Wang et al. (2021)), the divergent τ_j is necessary to balance the bias and robustness in the presence of heavy-tailed and/or skewed errors. On the other hand, the order of τ_j in our setting differs from those of earlier studies on adaptive Huber regressions. For example, with a finite $(1 + \epsilon)$th moment of the error, Sun, Zhou and Fan (2020) focused on estimating the adaptive Huber regression that corresponds to $p = 1$ in our setting, and considered $\tau_j = O(n^{\max\{1/(1+\epsilon),1/2\}}(d + \log n)^{-\max\{1/(1+\epsilon),1/2\}})$. Fan et al. (2019) used $\tau_j = O(n^{1/2}\{\log(np)\}^{-1/2})$ to test p-dimensional mean vectors under the assumption of a finite fourth moment of the errors, which corresponds to $d = 1$ in our setting. In practice, τ_j can be chosen using either cross-validation or the recent data-driven method of Wang et al. (2021). The latter avoids a grid search for each j, and hence is computationally appealing, especially for large p. Using these robust estimates $\widehat{\boldsymbol{\theta}}_j$, we then construct test statistics with approximated p-values for (2.2) that are obtained under the null. Together with the Benjamini−Hochberg (BH) method (Benjamini and Hochberg (1995)) or its

variants, for example, Storey (2002), we develop a robust procedure to simultaneously test the p hypotheses in (2.2).

2.1. Test procedure for general linear hypotheses

We now describe our test procedure for (2.2). Given the estimators $\widehat{\boldsymbol{\theta}}_j$ obtained from (2.3), with $\tau_j = \tau_{0j} n^{1/(2+\delta)} \{\log(np) + d\}^{-1/(2+\delta)}$ for $\tau_{0j} \geq v_{j,\delta}$ and $\delta \in (0, 2]$, we consider the following test statistic:

$$V_j = n(\mathbf{C}\widehat{\boldsymbol{\theta}}_j - \mathbf{c}_{0j})^{\mathrm{T}} (\mathbf{C}\widehat{\boldsymbol{\Sigma}}_j \mathbf{C}^{\mathrm{T}})^{-1} (\mathbf{C}\widehat{\boldsymbol{\theta}}_j - \mathbf{c}_{0j}), \tag{2.4}$$

for each j, where $\widehat{\boldsymbol{\Sigma}}_j$ is an estimate of $\boldsymbol{\Sigma}_j := \mathrm{cov}(n^{1/2}\widehat{\boldsymbol{\theta}}_j)$; see Sections 2.2 and 3.2. In (2.2), H_{0j} is rejected for large V_j. As we will show, the V_j are asymptotically χ^2_q-distributed under H_{0j} uniformly in j. Leveraging this, we can estimate the FDP to determine the rejection threshold that bounds the estimated FDP by a prespecified level $\alpha \in (0, 1)$.

Let $\mathcal{H}_0 = \{j : 1 \leq j \leq p, H_{0j} \text{ is true}\}$ and $p_0 := |\mathcal{H}_0|$. Denote the number of discoveries and false discoveries by $R(z) = \sum_{j=1}^{p} \mathbb{I}(V_j \geq z)$ and $V(z) = \sum_{j \in \mathcal{H}_0} \mathbb{I}(V_j \geq z)$, respectively, for the threshold $z > 0$. The FDP is defined as $\mathrm{FDP}(z) = V(z)/\max\{R(z), 1\}$. According to the law of large numbers, $V(z)$ should be close to $p_0 \mathbb{P}(\chi^2_q > z)$, whereas the number of nulls p_0 is not accessible, in general. When both p and p_0 are large and $p_1 = p - p_0 = o(p)$ is small, which is known as a sparse setting in the high-dimensional regime, the approximated FDP $\mathrm{AFDP}(z) = \widehat{V}(z)/\max\{R(z), 1\}$, with $\widehat{V}(z) = p\,\mathbb{P}(\chi^2_q > z)$, is a reasonable and slightly conservative surrogate for the asymptotic approximation $p_0 \mathbb{P}(\chi^2_q > z)/\max\{R(z), 1\}$ and $\mathrm{FDP}(z)$. Using $\mathrm{AFDP}(z)$, we can determine the threshold $\widehat{z}_\alpha = \inf\{z \geq 0 : \mathrm{AFDP}(z) \leq \alpha\}$ for the nominal level α. For $j = 1, \ldots, p$, H_{0j} is rejected whenever $V_j \geq \widehat{z}_\alpha$. Essentially, our procedure is based on the BH method, with the input p-values obtained from robustified/Huberized test statistics. Similar ideas are also adopted in Cai and Liu (2016) and Cai et al. (2019). The main difference is that the latter test statistics have closed-form expressions, whereas our statistics are based on M-estimators.

Note that if $\pi_0 = p_0/p$ is bounded away from one as $p \to \infty$, $\mathrm{AFDP}(z)$ may overestimate $\mathrm{FDP}(z)$. To improve the power, we may combine existing estimations of π_0 in the literature with our procedure to calibrate the threshold of rejection in a more adaptive fashion. For example, Storey (2002) estimates $V(z)$ by $p\widehat{\pi}_0(\eta)\,\mathbb{P}(\chi^2_q > z)$ for a predetermined $\eta \in [0, 1)$, where $\widehat{\pi}_0(\eta) = \{(1 - \eta)p\}^{-1} \sum_{j=1}^{p} \mathbb{I}(P_j > \eta)$ and P_j is the p-value associated with the jth test statistic. Storey and Tibshirani (2003) suggest $\eta = 0.5$, and Blanchard and Roquain

(2009) recommend $\eta = \alpha$ for dependent hypotheses. Using this estimate of $V(z)$, our threshold of rejection can be refined accordingly as $\widehat{z}_\alpha^\eta = \inf\{z \geq 0 : p\widehat{\pi}_0(\eta)\,\mathbb{P}(\chi_q^2 > z)/R(z) \leq \alpha\}$.

2.2. A refined Huber-type estimator of $\boldsymbol{\Sigma_j}$

A naive estimator of $\boldsymbol{\Sigma}_j = \mathrm{cov}(n^{1/2}\widehat{\boldsymbol{\theta}}_j)$ for conducting our test is $\widetilde{\sigma}_{\epsilon,jj}\widehat{\boldsymbol{\Sigma}}_Z^{-1}$, where $\widetilde{\sigma}_{\epsilon,jj}$ is an estimate of $\sigma_{\epsilon,jj}$, and $\widehat{\boldsymbol{\Sigma}}_Z = n^{-1}\sum_{i=1}^n \mathbf{Z}_i\mathbf{Z}_i^{\mathrm{T}}$. When only $\boldsymbol{\mu}$ is present, that is, $d = 0$, Fan et al. (2019) proposed a U-statistic-based variance estimator and an adaptive Huber-type estimator of the second moment, which, combined with the mean estimator, is used to estimate the variance. The computational complexity of their estimator is $O(n^2(d+1))$, and hence grows fast with d. For the latter estimator, because the squared data is severely right skewed, the Huber-type truncation inevitably leads to an underestimation of the second moment, and therefore the variance. Motivated by the classical theory of the Huber regression (Section 7.6 in Huber and Ronchetti (2009)), we propose an estimator $\widehat{\boldsymbol{\Sigma}}_j$ based on the asymptotic covariance of the conventional Huber regression estimator.

Given $\tau > 0$, the classical Huber regression estimator $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ admits that $n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ converges to $N(0, \boldsymbol{\Sigma}_\tau)$ in distribution, where $\boldsymbol{\Sigma}_\tau = \{\mathbb{P}(|\epsilon| < \tau)\}^{-2}\,\mathbb{E}\{\ell_\tau'(\epsilon)^2\}\boldsymbol{\Sigma}_Z^{-1}$ and $\boldsymbol{\Sigma}_Z = \mathbb{E}(\mathbf{Z}\mathbf{Z}^{\mathrm{T}}) \in \mathbb{R}^{(d+1)\times(d+1)}$ (Huber (1973)). Resembling $\boldsymbol{\Sigma}_\tau$, our estimator $\widehat{\boldsymbol{\Sigma}}_j$ consists of three Huber-type estimates and uses the tapering function (Cai, Zhang and Zhou (2010))

$$\mathbb{I}_\tau^*(x) = \mathbb{I}(|x| \leq \tau) + h_n^{-1}(\tau + h_n - |x|)\mathbb{I}(\tau < |x| \leq \tau + h_n), \tag{2.5}$$

which is h_n^{-1}-Lipschitz continuous. Given a robustification parameter $\tau_j > 0$ and the corresponding estimate $\widehat{\boldsymbol{\theta}}_j$ from (2.3), define $\mathbf{W}_j = n^{-1}\sum_{i=1}^n \mathbb{I}_{\tau_j}^*(e_{ij})\mathbf{Z}_i\mathbf{Z}_i^{\mathrm{T}}$ and $m_j = n^{-1}\sum_{i=1}^n \mathbb{I}_{\tau_j}^*(e_{ij})$, where $e_{ij} = Y_{ij} - \mathbf{Z}_i^{\mathrm{T}}\widehat{\boldsymbol{\theta}}_j$. Here $\mathbf{W}_j$ and m_j are estimates of $\mathbb{P}(|\epsilon_{1j}| \leq \tau_j)\boldsymbol{\Sigma}_Z$ and $\mathbb{P}(|\epsilon_{1j}| \leq \tau_j)$, respectively. Recall that $\widehat{\boldsymbol{\Sigma}}_Z = n^{-1}\sum_{i=1}^n \mathbf{Z}_i\mathbf{Z}_i^{\mathrm{T}}$. Inspired by (7.83) in Huber and Ronchetti (2009), we define the covariance estimator $\widehat{\boldsymbol{\Sigma}}_j$ in (2.4) as

$$\widehat{\boldsymbol{\Sigma}}_j = \left[\sum_{i=1}^n \{\ell_{\tau_j}'(e_{ij})\}^2\right]\{(n-d-1)K_j\}^{-1}\mathbf{W}_j^{-1}\widehat{\boldsymbol{\Sigma}}_Z\mathbf{W}_j^{-1}, \tag{2.6}$$

where $K_j = 1 + (nm_j)^{-1}(d+1)(1-m_j)$ is a correction factor that benefits the finite-sample performance.

For the conventional Huber regression with fixed $\tau > 0$, it can be shown that,

with $\mathbb{I}^*_\tau(x)$ replaced by $\mathbb{I}(|x| \le \tau)$, $\widehat{\mathbf{\Sigma}}_j$ converges in probability to $\mathbf{\Sigma}_\tau$ as $n \to \infty$. To legitimize using V_j to test (2.2), we show in Section 3.2 that with the adaptive τ_j, the covariance estimator $\widehat{\mathbf{\Sigma}}_j$ in (2.6) is close to $\mathbf{\Sigma}_j$ uniformly over j, with high probability. In addition, because h_n is aligned with $\tau = \tau_0 a(n, p, d)$ for some function a in n, p, d, to make it scale invariant, a more adaptive approach is to consider ch_n, where c can be set as τ_0, which is determined similarly to τ (Wang et al. (2021)), or as a minimum absolute deviation estimator of the variance using the fitted residuals. Refer to Section S5.3 in the Supplement Material for a numerical experiment that examines the stability of our method on the choice of h_n.

2.3. Related works

Our method generalizes the robust large-scale simultaneous mean testing procedure considered by Fan et al. (2019). In addition to the robust multiple inference, Fan et al. (2019) focused on modeling $\mathbf{\Xi}$ in (2.1) using a latent factor model to improve the power, without which their problem can be viewed as a special case of (2.1). Methodologically, to draw multiple inferences on $\mathbf{B}$ in (2.1) with $p \gg n$, an easily computable and accurate estimate of the covariance of the adaptive Huber regression coefficient is needed for all p regressions. Such an estimator dictates a careful exploitation of the design $\mathbf{Z}$, whereas Fan et al. (2019) considered only $\mathbf{Z} = \mathbf{1} \in \mathbb{R}^{n\times 1}$, which is not trivially extendable to the problem we consider here.

Our estimator in (2.6) bridges the gap, and consists of two parts: the first part $(n-d-1)^{-1}\sum_{i=1}^n \{\ell'_{\tau_j}(e_{ij})\}^2$ provides a robust estimate of $\sigma_{\epsilon,jj}$, and the second part $K_j^{-1}\mathbf{W}_j^{-1}\widehat{\mathbf{\Sigma}}_Z\mathbf{W}_j^{-1}$ offers a robustification of the inverse Gram matrix $\left(n^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}$. This can be naturally considered as a robustification of the covariance of the least squares estimator. In addition, by using the tapering function $\mathbb{I}^*_\tau(x)$ as a smoothed version of the second-order derivative of Huber's loss to specify $\mathbf{W}_j$ and K_j, our estimator is continuous, which is crucial for the uniform consistency of $\widehat{\mathbf{\Sigma}}_j$ across j. The uniform consistency of $\widehat{\mathbf{\Sigma}}_j$ leads to the FDP control of our robust multiple test for large-scale multivariate regressions. In contrast, in addition to the fact that the procedure of Fan et al. (2019) is not able to exploit $\mathbf{Z}$ when $d \ge 1$, their variance estimator of $\sigma_{\epsilon,jj}$, which is the difference between a (restricted) robust second-order moment and a squared robust first-order moment of the error, may suffer from bias when d is large, as discussed in Section 2.2. Moreover, it requires extra tuning parameters to robustly estimate the second-order moment. A numerical experiment is reported in Section S5.5 of

the Supplementary Material to verify the above discussion.

3. Statistical Guarantees

In this section, we establish theoretical guarantees of our method by first assuming a known $\mathbf{\Sigma}_j$, and then exploring the closeness between $\mathbf{\Sigma}_j$ and $\widehat{\mathbf{\Sigma}}_j$ in (2.6). Hereafter, we focus on $\mathbf{Z}_i$ being random (except for the first coordinate), and report the results under fixed designs in the Supplementary Material.

3.1. Approximation of FDP with known $\mathbf{\Sigma}_j$

Assume the covariance matrix $\mathbf{\Sigma}_j$ is known for each j. Consider the oracle test statistic $V_j^\circ = n(\mathbf{C}\widehat{\boldsymbol{\theta}}_j - \mathbf{c}_{0j})^{\mathrm{T}}(\mathbf{C}\mathbf{\Sigma}_j\mathbf{C}^{\mathrm{T}})^{-1}(\mathbf{C}\widehat{\boldsymbol{\theta}}_j - \mathbf{c}_{0j})$. Given $z \geq 0$, write $R^\circ(z) = \sum_{j=1}^p \mathbb{I}(V_j^\circ > z)$, $V^\circ(z) = \sum_{j\in\mathcal{H}_0} \mathbb{I}(V_j^\circ > z)$, and $\mathrm{FDP}^\circ(z) = V^\circ(z)/R^\circ(z)$. Heuristically, V_j° is approximately χ_q^2-distributed under H_{0j}, so that we can approximate $\mathrm{FDP}^\circ(z)$ by

$$\mathrm{AFDP}^\circ(z) = \{p_0\,\mathbb{P}(\chi_q^2 > z)\}\{R^\circ(z)\}^{-1}. \tag{3.1}$$

To show that $\mathrm{AFDP}^\circ(z)$ provides a valid asymptotic (pointwise) approximation of $\mathrm{FDP}^\circ(z)$, we impose the following technical conditions. Denote $\mathbf{R}_\epsilon = (r_{\epsilon,jk})_{1\leq j,k\leq p}$ as the correlation matrix of $\boldsymbol{\epsilon}_1 = (\epsilon_{11},\ldots,\epsilon_{1p})^{\mathrm{T}}$, that is, $\mathbf{R}_\epsilon = \mathbf{D}_\epsilon^{-1}\mathbf{\Sigma}_\epsilon\mathbf{D}_\epsilon^{-1}$, with $\mathbf{D}_\epsilon^2 = \mathrm{diag}(\sigma_{\epsilon,11},\ldots,\sigma_{\epsilon,pp})$.

Condition 1. *(i) $p = p(n) \to \infty$ and $\log(p) = o(n^{1/2})$ as $n \to \infty$; (ii) the error vectors $\epsilon_1,\ldots,\epsilon_n$ are independent, and satisfy $\mathbb{E}(\epsilon_{ij}|\mathbf{Z}_i) = 0$, $\mathbb{E}(\epsilon_{ij}^2|\mathbf{Z}_i) = \sigma_{\epsilon,jj}$; (iii) there exist $\delta \in (0,2]$, $c_\epsilon > 0$, and $C_\epsilon > 0$, such that $c_\epsilon \leq \min_{1\leq j\leq p}\sigma_{\epsilon,jj}^{1/2} \leq \max_{1\leq j\leq p} v_{j,\delta} \leq C_\epsilon$; and (iv) there exist $\kappa_0 \in (0,1)$ and $\kappa_1 > 0$ such that $\max_{1\leq j\neq k\leq p}|r_{\epsilon,jk}| \leq \kappa_0$ and $p^{-2}\sum_{1\leq j\neq k\leq p}|r_{\epsilon,jk}| = O(p^{-\kappa_1})$.*

In Condition 1, (i) is a commonly assumed asymptotic regime for (n,p) in high-dimensional statistical inference; (ii) is standard for linear regression models; compared with traditional settings that presume a finite fourth or higher-order moments of errors, (iii) assumes only the uniform boundedness of the $(2+\delta)$th moments; and (iv) allows weak dependence among $\epsilon_{11},\ldots,\epsilon_{1p}$. In addition, we impose the following conditions on $\mathbf{Z}_i$. Denote $\widetilde{\mathbf{Z}}_i = \mathbf{\Sigma}_Z^{-1/2}\mathbf{Z}_i$, where $\mathbf{\Sigma}_Z = \mathbb{E}(\mathbf{Z}\mathbf{Z}^{\mathrm{T}})$ is assumed to be positive definite.

Condition 2. *The predictors $\{\mathbf{Z}_i\}_{i=1}^n$ are sub-Gaussian, that is, for some $A_0 > 0$, $\mathbb{P}(|\langle\mathbf{u},\widetilde{\mathbf{Z}}_i\rangle| \geq A_0\|\mathbf{u}\|t) \leq 2\exp(-t^2)$, for any $\mathbf{u}\in\mathbb{R}^{d+1}$ and $t \geq 0$.*

Refer to Vershynin (2018) for an overview of sub-Gaussian vectors. Under

Conditions 1 and 2, Proposition 1 shows that AFDP$^\circ$ in (3.1) consistently estimates FDP$^\circ$. It also provides a guideline to establish the FDP control and serves as the cornerstone of the guarantees of our method.

Proposition 1. *Assume Conditions* 1 *and* 2 *hold, and* $p_0 \geq ap$, *for some* $a \in (0,1)$. *Let* $\tau_j = \tau_{0j} n^{1/(2+\delta)} \{\log(np) + d\}^{-1/(2+\delta)}$, *with* $\tau_{0j} \geq v_{j,\delta}$ *and* $\delta \in (0,2]$. *Then, for any* $z \geq 0$, $|\mathrm{FDP}^\circ(z) - \mathrm{AFDP}^\circ(z)| = o_{\mathbb{P}}(1)$ *as* $n, p \to \infty$.

We conclude this subsection with two remarks. If we strengthen Condition 1 (iii) to uniformly bounded kth moments for $k \geq 4$, Proposition 1 remains valid, with $\tau_j = \tau_{0j} n^{1/(2+\delta)} \{\log(np)+d\}^{-1/(2+\delta)}$ and $\delta \in (0, k-2]$. In addition, to prove Proposition 1, we show that $|\mathrm{FDP}^\circ(z) - \mathrm{AFDP}^\circ(z)| = O_{\mathbb{P}}\{p^{-\kappa_1} q^{1/2} + q^{7/4} n^{-1/2} + q\{\log(np) + d\}^{\delta/(2+\delta)} n^{-\delta/(2+\delta)}\}$. This explicit rate is nontrivial and reveals how the parameter q, which corresponds to the dimension of the hypothesis, affects the difficulty of testing (2.2). We revisit this in our numerical studies in Section 4.

3.2. Statistical guarantees with estimated covariance input $\widehat{\mathbf{\Sigma}}_j$

Next, we establish the statistical guarantee of our method using the estimated covariance matrices $\widehat{\mathbf{\Sigma}}_j$ in (2.6). To this end, Theorem 1 provides a mild condition on the accuracy of the estimated covariances that lead to the consistency of the approximated FDP. Let $\widetilde{\mathbf{\Sigma}}_j$ be a generic estimator of $\mathbf{\Sigma}_j$ for each j. The corresponding FDP and its approximation are $\widetilde{\mathrm{FDP}}(z) = \widetilde{V}(z)/\widetilde{R}(z)$ and $\widetilde{\mathrm{AFDP}}(z) = p_0\, \mathbb{P}(\chi^2_q > z)/\widetilde{R}(z)$, for $z \geq 0$, where $\widetilde{V}(z) = \sum_{j \in H_0} \mathbb{I}(\widetilde{V}_j > z)$, $\widetilde{R}(z) = \sum_{j=1}^p \mathbb{I}(\widetilde{V}_j > z)$, and $\widetilde{V}_j = n(\mathbf{C}\widehat{\boldsymbol{\theta}}_j - \mathbf{c}_{0j})^{\mathrm{T}} (\mathbf{C}\widetilde{\mathbf{\Sigma}}_j \mathbf{C}^{\mathrm{T}})^{-1} (\mathbf{C}\widehat{\boldsymbol{\theta}}_j - \mathbf{c}_{0j})$.

Theorem 1. *Suppose that the conditions of Proposition* 1 *hold. As long as the estimated covariances* $\{\widetilde{\mathbf{\Sigma}}_j\}_{j=1}^p$ *satisfy* $\max_{1 \leq j \leq p} \|\widetilde{\mathbf{\Sigma}}_j - \mathbf{\Sigma}_j\| = o_{\mathbb{P}}\{(\log(np) + d)^{-1}\}$, *we have* $|\widetilde{\mathrm{FDP}}(z) - \widetilde{\mathrm{AFDP}}(z)| = o_{\mathbb{P}}(1)$, *for any* $z > 0$, *as* $n, p \to \infty$.

By verifying that $\widehat{\mathbf{\Sigma}}_j$ in (2.6) satisfy the required accuracy in Theorem 1, together with Proposition 1, Theorem 2 acquires the convergence in probability of the approximated FDP to the true FDP, for any $z > 0$, as $n, p \to \infty$.

Theorem 2. *Suppose that the conditions of Proposition* 1 *hold. For each* $\mathbf{\Sigma}_j = \mathrm{cov}(n^{1/2}\widehat{\boldsymbol{\theta}}_j)$, *for* $j = 1, \ldots, p$, *let* $\widehat{\mathbf{\Sigma}}_j$ *be the corresponding estimators given in* (2.6), *with* $\tau_j = \tau_{0j} n^{1/(2+\delta)} \{\log(np) + d\}^{-1/(2+\delta)}$ *and* $\tau_{0j} \geq v_{j,\delta}$, *for* $\delta \in (0,2]$. *Then, with probability at least* $1 - 16n^{-1}$,

$$\max_{1 \leq j \leq p} \left\| \widehat{\mathbf{\Sigma}}_j - \mathbf{\Sigma}_j \right\| \leq C_1 \max \left[\left\{ \frac{\log(np) + d}{n} \right\}^{\delta/(2+\delta)}, \frac{\Delta}{h_n} \right], \tag{3.2}$$

where $\Delta = \{d^{1/2} + (2\log n)^{1/2}\}[n^{-1}\{\log(np) + d\}]^{1/2}$, *and* $C_1 > 0$ *depends only on* $\lambda_{\max}(\Sigma_Z)$, A_0, *and* $v_{j,\delta}$.

Theorem 2 implies that the required accuracy in Theorem 1, that is, $\max_{1\le j\le p} \|\widehat{\mathbf{\Sigma}}_j - \mathbf{\Sigma}_j\| = o_{\mathbb{P}}\{(\log(np) + d)^{-1}\}$, is met if $\log(p) + d = o(n^{\delta/(2+2\delta)})$ and $\Delta/h_n = o\{(\log(np) + d)^{-1}\}$, such as $h_n = n^{-1/4}$. So far, we have focused on $\widehat{\mathbf{\Sigma}}_j$ in (2.6). In fact, the conclusion in Theorem 2 remains valid for some variants of $\widehat{\mathbf{\Sigma}}_j$, such as $\widehat{\mathbf{\Sigma}}_j^{(1)} = \sum_{i=1}^n \{\ell'_{\tau_j}(e_{ij})\}^2\{(n-d-1)m_j\}^{-1}K_j\mathbf{W}_j^{-1}$.

4. Simulation Studies

4.1. Model settings

To examine the finite-sample performance of our procedure, we consider the following methods: (i) our method that employs the data-adaptive Huber regression (Wang et al. (2021)); (ii) our method with τ_j selected using five-fold cross-validation (Sun, Zhou and Fan (2020)); (iii) a least squares-based multiple testing method; (iv) an empirical Bayes-based multiple testing method implemented using `limma` (Ritchie et al. (2015)); (v) `limma`, with the traditional robust M-estimation instead of the least squares; and (vi) an empirical Bayes-based multiple testing method for count data, implemented using `edgeR` (Robinson, McCarthy and Smyth (2010)). Both `limma` and `edgeR` are widely used to analyze a large number of regression models, and serve as benchmarks in genomics studies. `limma` employs empirical Bayes methods to shrink individual variances toward a common value better control the FDR. `edgeR` models count data with large variations using the negative binomial model. To implement `edgeR`, we round the response Y_{ij} to its nearest integer. For our method, we set $\delta = 2$ in (2.3) (i.e., we assume the errors have finite fourth moments) and $h_n = n^{-1/4}$ in (2.5). For (ii), we set $\tau_j = c\widehat{v}_j n^{1/4}\{\log(np)+d\}^{-1/4}$, with $\widehat{v}_j^4 = n^{-1}\sum_{i=1}^n (Y_{ij} - \bar{Y}_{\cdot j})^4$, and choose c from $\{0.25, 0.5, 0.75, 1, 1.25, 1.5\}$ based on cross-validation that minimizes the mean-squared prediction error. For (i)–(iii), we use the FDR controlling procedure of Storey (2002) to determine the threshold.

We generate data from model (2.1) for $n = 85, 120, 150$, $p = 1000, 2000$, $p_1 = 50$, and $d = 6, 8$. Entries of $\mathbf{X} \in \mathbb{R}^{n\times d}$ are drawn independently from $N(0,1)$, and each column is standardized to have a zero mean and unit variance. We consider three heavy-tailed and highly skewed error distributions: (a) Pareto (scale $= 1$, shape $= 4$), (b) log-normal ($\mu = 0, \sigma = 1$), and (c) a mixture of the log-normal in (b) and a t_2 distribution with proportions 0.7 and 0.3, respectively. Setting (c) reflects more challenging scenarios in practice, because t_2 does not have a finite second moment. Under each setting, we first

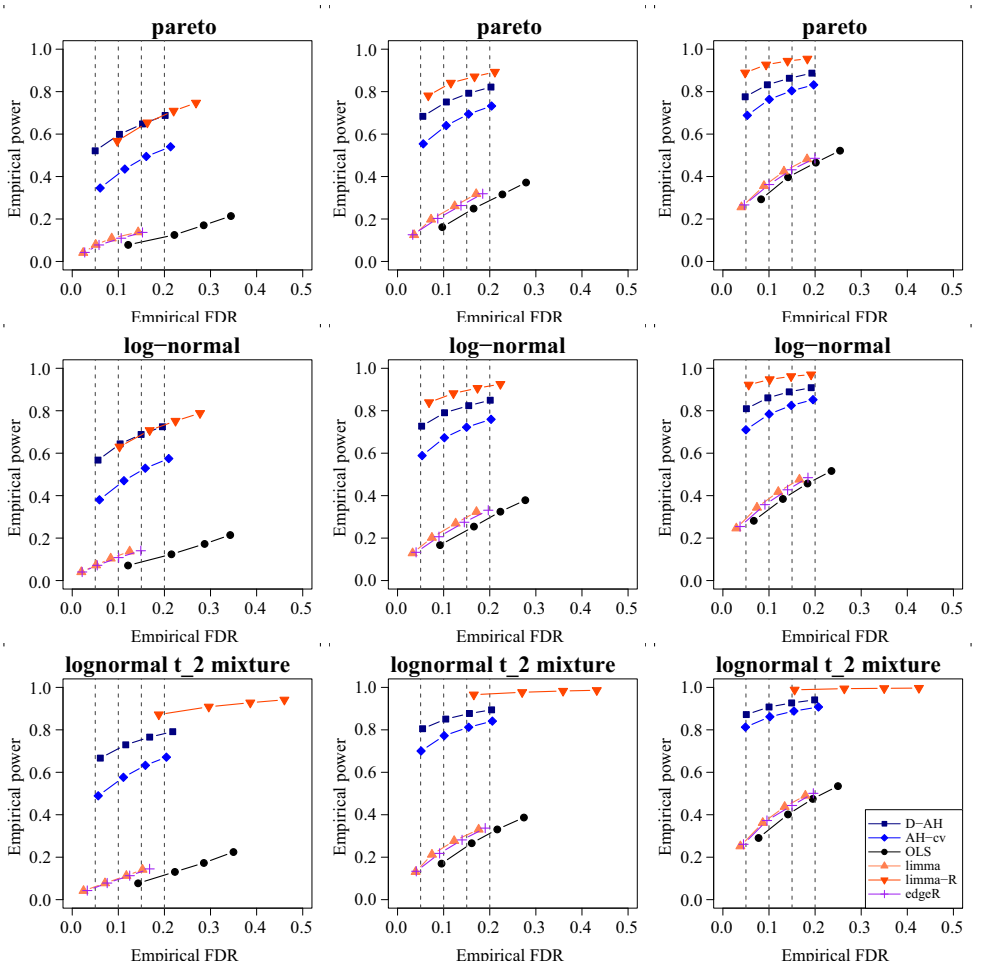
generate $\mathbf{E} = (\epsilon_{ij})_{1\le i\le n, 1\le j\le p}$ with i.i.d. entries. To incorporate dependence, set $\Xi = 100\mathbf{R}_\epsilon^{1/2}\mathbf{E}$, where the correlation matrix $\mathbf{R}_\epsilon = (r_{\epsilon,jk})_{1\le j,k\le p}$ admits one of the following three structures: *Model* 1, the identity matrix; *Model* 2, $r_{\epsilon,ij} = r_{\epsilon,ji}$ drawn independently from $0.3 \times \mathrm{Ber}(0.1)$, for $i \neq j$; and *Model* 3, $r_{\epsilon,j,j+1} = r_{\epsilon,j+1,j} = 0.3$, $r_{\epsilon,j,j+2} = r_{\epsilon,j+2,j} = 0.1$, and $r_{\epsilon,j,j+k} = r_{\epsilon,j+k,j} = 0$, for $k \ge 3$. Note that *Model* 2 does not satisfy Condition 3.2 (iv). Together with the results in Section S5.4 of the Supplementary Material, the results for *Model* 2 show that our method is reliable even when Condition 3.2 (iv) is mildly violated.

For each $j = 1, \ldots, p$, we set $\mu_j = 5000$ and consider two hypotheses: *Hypothesis* 1, $H_{0j} : 1^{\mathrm{T}}\boldsymbol{\beta}_j = 0$ versus $H_{aj} : 1^{\mathrm{T}}\boldsymbol{\beta}_j \neq 0$, where $q = 1$; and *Hypothesis* 2, $H_{0j} : \boldsymbol{\beta}_j = 0 \in \mathbb{R}^d$ versus $H_{aj} : \boldsymbol{\beta}_j \neq 0$, where $q = d$. For *Hypothesis* 1, let $\beta_{jk} \sim \mathrm{Unif}(-150, 150)$, for $1 \le j \le p$ and $1 \le k \le d-1$, $\beta_{jd} = -\sum_{k=1}^{d-1}\beta_{jk}$, for $1 \le j \le p - p_1$, so that $1^{\mathrm{T}}\boldsymbol{\beta}_j = 0$, and $\beta_{jd} = \delta d^{1/2} W_j - \sum_{k=1}^{d-1}\beta_{jk}$, for $p - p_1 + 1 \le j \le p$, where W_j are Rademacher variables. For *Hypothesis* 2, let $\boldsymbol{\beta}_j = 0$ for $1 \le j \le p - p_1$, and $\beta_{jk} = (2d^{-1})^{1/2}\delta W_{jk}$ for $p - p_1 + 1 \le j \le p$ and $1 \le k \le d$, where W_{jk} are Rademacher variables. We take $\delta = 75\eta$ and $\eta = 0.3$, which determine the signal strength.

4.2. Numerical performance

We use the nominal FDR level $\alpha \in \{0.05, 0.1, 0.15, 0.2\}$, and carry out 250 Monte Carlo simulations at each α. Figures 1 and 2 report the empirical FDR and power under *Model* 2 with $p = 1000$ and $d = 6$. The results under other settings are documented in Section S5 of the Supplementary Material. Each point corresponds to a nominal level (marked as a vertical gray dashed line), with the x- and y-axes representing, the empirical FDR and the power, respectively. Therefore, the closer the point is to the corresponding vertical line, the more the empirical and nominal FDRs coincide.

From Figures 1 and 2, across different error settings and hypotheses, our method, with either the data-driven Huber regression or cross-validation, controls the FDR well, in general, and maintains high power. The competitors are either too conservative, with a notable power loss, or too liberal to control the FDR, especially for small n. The advantage of our method is more substantial when $q > 1$ (Figure 2). The numerical evidence favors using the data-adaptive Huber regression over cross-validation in terms of both statistical accuracy and computational efficiency. Both `limma` and `edgeR` are fairly conservative, suggesting that researchers should be careful when using them for heavy-tailed and skewed data. Method (v) is comparable to our method when n is large, but completely fails to control the FDR for errors from the mixture of the log-normal and

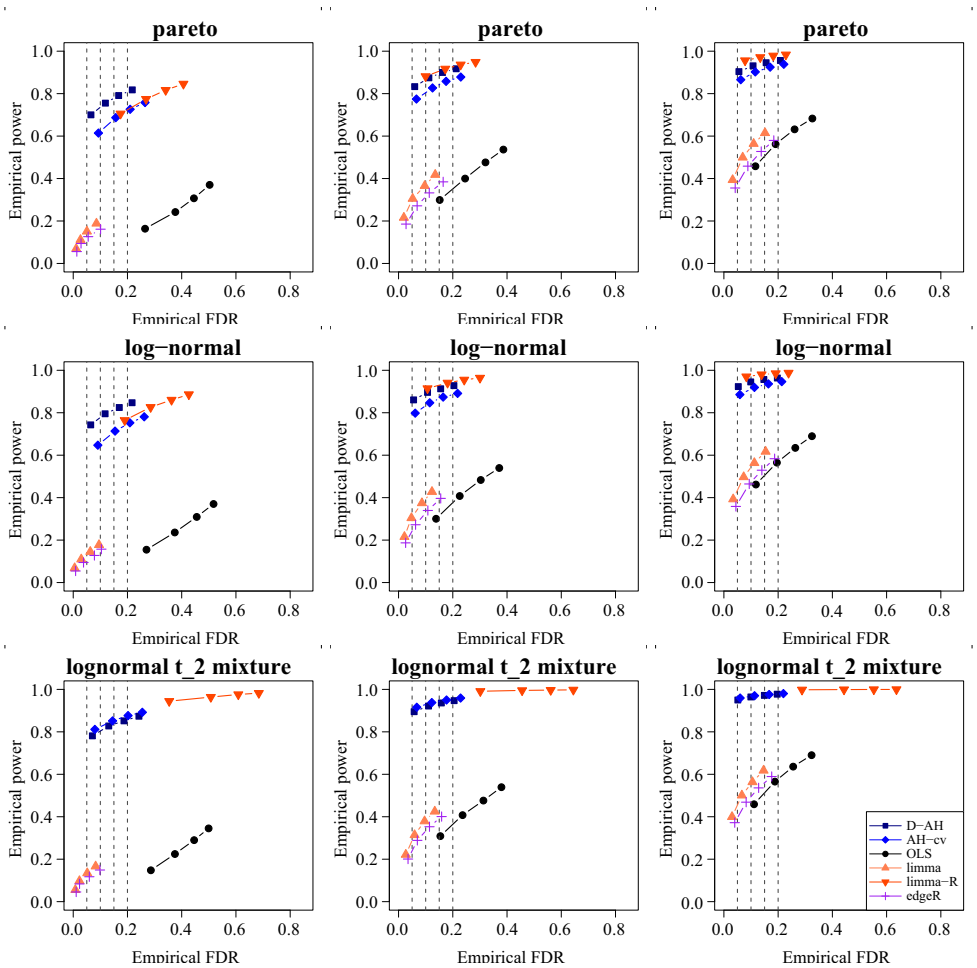


(a) $n = 85$ (b) $n = 120$ (c) $n = 150$

Figure 1. Empirical FDR and power for testing *Hypothesis* 1 under *Model* 2 with $p = 1000$ and $d = 6$ using six methods: the proposed method with the data-adaptive Huber regression (D-AH, ■); the proposed method with cross-validation (AH-cv, ⧫); the least squares method (OLS, •); `limma` (▲); `limma` with a robust regression (`limma`-R, ▼); and `edgeR` (+). Each point corresponds to a nominal FDR level (marked as a vertical dashed line), with the x- and y- axes denoting the empirical FDR and the power, respectively.

t_2. Overall, the power of all methods increases with n, and drops for larger p, see Figures S1−S11 in the Supplementary Material. Because the intrinsic difficulty of the testing problem elevates with q, the power of all methods shrinks when $q = d = 8$ (Figures S3 and S4).

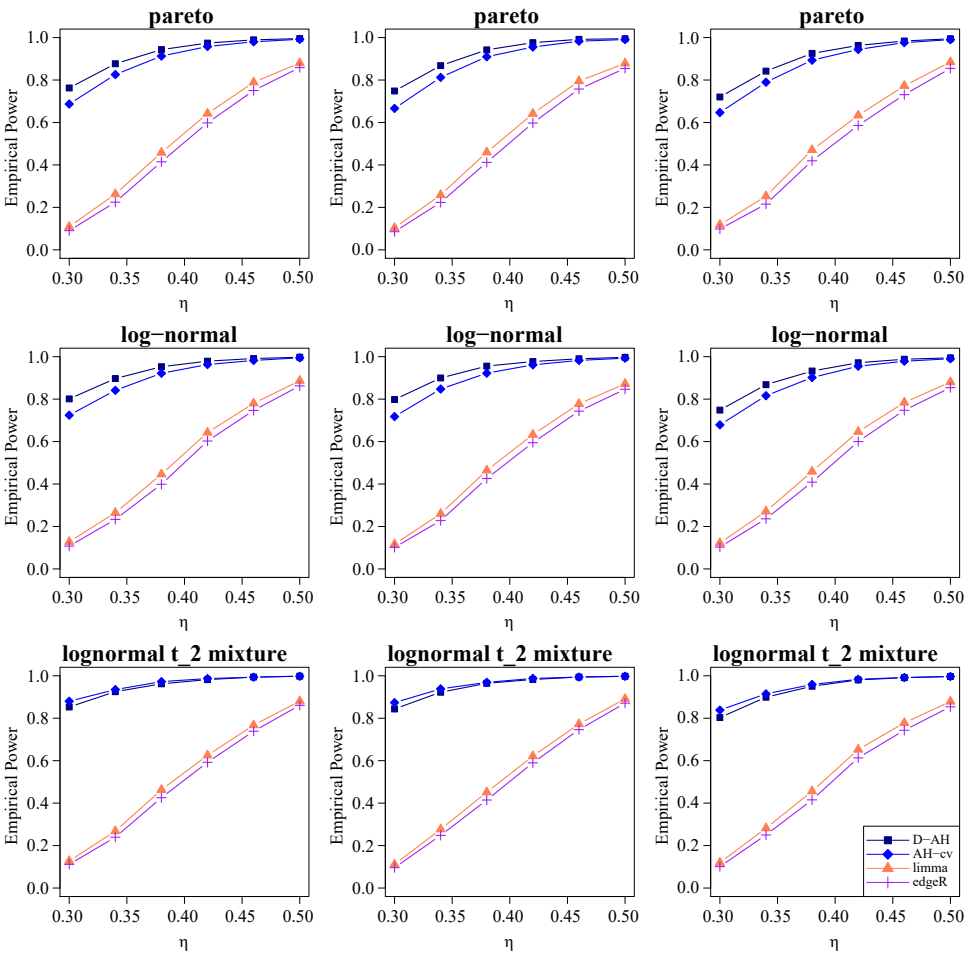
We further examine the power with varying signal strengths, determined by η. We exclude methods (iii) and (v), because they fail to control the FDR. In



(a) $n = 85$ (b) $n = 120$ (c) $n = 150$

Figure 2. Empirical FDR and power for testing *Hypothesis* 2 under *Model* 2 with $p = 1000$ and $d = 6$ using six methods: the proposed method with the data-adaptive Huber regression (D-AH, ■); the proposed method with cross-validation (AH-cv, ⧫); the least squares method (OLS, •); `limma` (▲); `limma` with a robust regression (`limma`-R, ▼); and `edgeR` (+). Each point corresponds to a nominal FDR level (marked as a vertical dashed line), with the x- and y- axes denoting the empirical FDR and the power, respectively.

the above settings, we take $n = 100$, $p = 1000$, $d = 6$, and choose equally spaced η within $[0.3, 0.7]$ for *Hypothesis* 1 and within $[0.3, 0.5]$ for *Hypothesis* 2. From Figure 3, we see that the proposed methods outperform the competitors across all error settings. The gains in power are considerable when the error is both heavy tailed and skewed. Again, for our method, the data-adaptive approach is preferable to cross-validation. With heavier tails (mixture of log-normal and t_2),



(a) *Model* 1 (b) *Model* 2 (c) *Model* 3

Figure 3. Plots of the empirical power for testing *Hypothesis* 2 with $n = 100$, $p = 1000$, $d = 6$, and $\eta \in \{0.30, 0.34, \ldots, 0.46, 0.5\}$ using four methods: the proposed method with the data-adaptive Huber regression (D-AH, ■); the proposed method with cross-validation (AH-cv, ⧫); `limma` (▲); and `edgeR` (+).

the power decreases slightly for all methods.

5. Real-Data Analysis: The Gutenberg Project

Inference on large-scale text data from literary publications has drawn increased attention, and has provided novel and revealing discoveries in a variety of fields, including sociology (O'Connor, Bamman and Smith (2011)), political science (Wilkerson and Casas (2017); Baum, Cohen and Zhukov (2018)), crim-

inology (Caines et al. (2018)), and linguistics. A major task in text analysis is to identify word markers to distinguish or identify different authors, cultures, resources, and so on. These word markers are usually identified by small p-values from testing regression coefficients used to model subject effects on the word frequency, or from model comparisons among multiple groups. In computational linguistics, for example, Marsden et al. (2013) compared 168 plays from the Shakespearean era to identify word markers for authorship classification. Here, we consider hypotheses that help identify the distinctive word markers, referred to as "differentially represented" (DR) words, to identify authors or different writing styles of a particular author.

As a well-known publicly accessible digital library of literary publications, the Project Gutenberg was founded in 1971, offering 60,156 e-books, as of September 03, 2019. The Standardized Project Gutenberg Corpus (SPGC, Gerlach and Font-Clos (2020)) is a text corpus of Project Gutenberg, and provides a static version of the corpus (`https://doi.org/10.5281/zenodo.2422560`). It consists of three data types: raw text, sequences of word-tokens, and word counts. It also contains metadata about books, such as author information, subject categories, and book types.

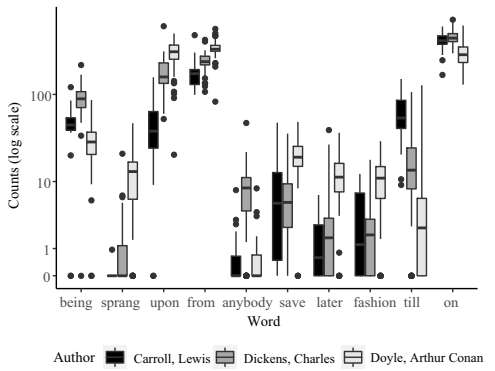
We apply our method to word counts from SPGC to identify idiosyncratic word markers that represent an author or a category of works. Specifically, we consider two problems: a comparison of the works of Lewis Carroll, Charles Dickens, and Arthur Conan Doyle, and a study of the works of William Shakespeare. See a snapshot of the raw data in the Supplementary Material. From the histograms of the empirical kurtosis of the word counts (Figure S20), the data are heavy tailed both book-wise and word-wise. For pre-processing, we first merge the word counts across books, and then remove those words with a total count of less than half the number of books or those that appear in less than 20% of the books under consideration. Finally, we normalize the filtered word counts by the total counts (Bullard et al. (2010)). The details are deferred to the Supplementary Material.

For the first problem, the three British authors are all from the mid-19th to early 20th century, and share similar writing structures and backgrounds. On the other hand, we also observe separations of their 167 works based on word usage in Figure S20 in the Supplementary Material. To identify DR words in their works, we use model (2.1) with $\mathbf{X}_i = (1, 1, 0)^{\mathrm{T}}$ if the ith book is written by Carroll, $\mathbf{X}_i = (1, 0, 1)^{\mathrm{T}}$ if it is written by Dickens, and $\mathbf{X}_i = (1, -1, -1)^{\mathrm{T}}$ if it is written by Conan Doyle, for $1 \le i \le 167$ books, and $\boldsymbol{\beta}_j = (\mu_j, \alpha_{1j}, \alpha_{2j})^{\mathrm{T}}$, for $1 \le j \le 6839$ words. We consider the following linear hypotheses: (Hypothesis CDD1) H_{0j} :

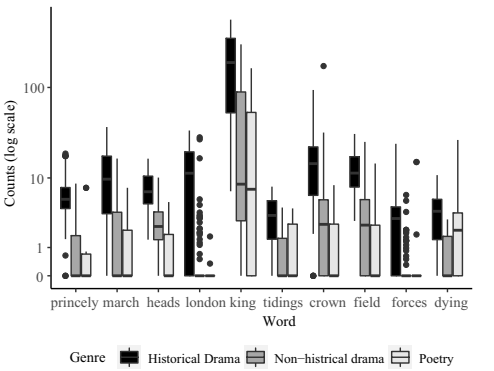
$[(0\ 1\ 0)^{\mathrm{T}}\ (0\ 0\ 1)^{\mathrm{T}}]^{\mathrm{T}}\boldsymbol{\beta}_j = 0$ versus $H_{aj} : [(0\ 1\ 0)^{\mathrm{T}}\ (0\ 0\ 1)^{\mathrm{T}}]^{\mathrm{T}}\boldsymbol{\beta}_j \neq 0$; (Hypothesis CDD2) $H_{0j} : \alpha_{1j} = 0$ versus $H_{aj} : \alpha_{1j} \neq 0$; (Hypothesis CDD3) $H_{0j} : \alpha_{2j} = 0$ versus $H_{aj} : \alpha_{2j} \neq 0$; and (Hypothesis CDD4) $H_{0j} : (0,1,1)^{\mathrm{T}}\boldsymbol{\beta}_j = 0$ versus $H_{aj} : (0,1,1)^{\mathrm{T}}\boldsymbol{\beta}_j \neq 0$. Hypothesis CDD1 compares the three authors together, whereas the other hypotheses compare one author with the remaining two. Using a nominal level of 0.5%, our method detects 2595, 419, 1388, and 1445 DR words, respectively for each hypothesis. The top 10 DR words for the three authors, such as *being* and *sprang* are displayed in Figure 4(a). The overall comparison is reported in Figure S21 in the Supplementary Material. Note that Conan Doyle favored the word *sprang*, whereas Carroll and Dickens rarely used it. In Figure 4(c), we further report the percentages of DR words and non-differentially represented (NDR) words within each speech category (Nguyen et al. (2016)). DR words among the three authors have higher percentages in adjectives, adverbs, and pronouns than do NDRs. In contrast, DR words have lower percentages of nouns, proper nouns, and verbs.

Next, we investigate the genre difference between the works of Shakespeare based on three subject groups: poetry, non-historical drama, and historical drama. We model the normalized word counts by (2.1), with $\mathbf{X}_i = (1,0,0)^{\mathrm{T}}$ if the ith book is poetry, $\mathbf{X}_i = (1,1,0)^{\mathrm{T}}$ if it is non-historical drama, and $\mathbf{X}_i = (1,1,1)^{\mathrm{T}}$ if it is historical drama, for $i = 1,\ldots,176$ books, and $\boldsymbol{\beta}_j = (\mu_j, \alpha_j, \gamma_j)^{\mathrm{T}}$, for $j = 1,\ldots,4122$ words. We consider (Hypothesis WS1) $H_{0j} : (0,0,1)^{\mathrm{T}}\boldsymbol{\beta}_j = 0$ versus $H_{aj} : (0,0,1)^{\mathrm{T}}\boldsymbol{\beta}_j \neq 0$, which compares the non-historical and historical dramas, and (Hypothesis WS2) $H_{0j} : (0,2,1)^{\mathrm{T}}\boldsymbol{\beta}_j = 0$ versus $H_{aj} : (0,2,1)^{\mathrm{T}}\boldsymbol{\beta}_j \neq 0$, which distinguishes poetry and dramas. With a nominal level of 0.5%, our method identifies 724 and 225 DR words for each hypothesis. Because many of Shakespeare's historical dramas are about kings of the Kingdom of England, the words *princely*, *London*, *king*, and *crown* appear more often in the historical dramas (Figure 4(b)). In addition, Shakespeare used words such as *march*, *forces*, *army*, and *battle* more frequently in the historical dramas than in the non-historical dramas. Interestingly, the love story-related lexicons, such as *love* and *marry*, appear more often in his non-historical dramas. From Figure 4(d), the DR words between Shakespeare's historical and non-historical dramas have higher percentages of nouns, pronouns, and proper nouns, whereas their percentages are lower for adjectives, adverbs, and verbs.

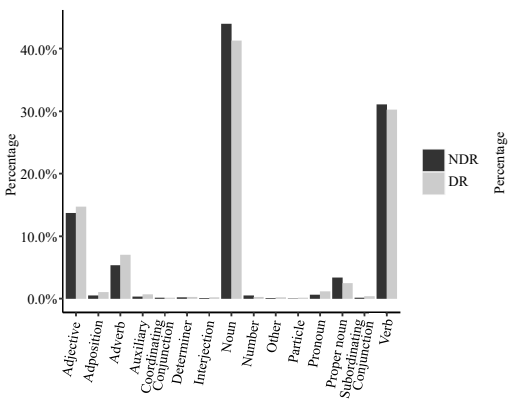
In summary, our method provides a reliable addition to the existing toolkit in corpus linguistics and text/literature analysis. It can be used to analyze a large volume of individual words, extending current methods that focus on the overall distribution of word counts. An interesting follow-up work is to investigate how



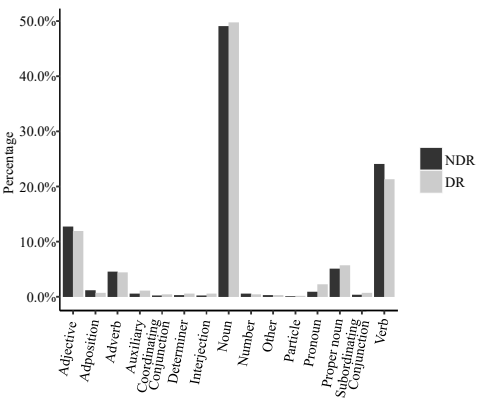
(a) Hypothesis CDD1



(b) Hypothesis WS1



(c) Hypothesis CDD1



(d) Hypothesis WS1

Figure 4. Panels (a) and (b): The top 10 DR words placed in ascending order by their p-values (from left to right) for hypotheses CDD1 and WS1, respectively, where the vertical axis shows the counts under a log-scale. Panels (c) and (d): Percentages of DR and NDR words within each speech category (`https://universaldependencies.org/u/pos/all.html`) for hypotheses CDD1 and WS1, respectively. The nominal FDR level is 0.5%.

stopping words, such as *upon* affect the results, and whether their removal alter the discovery.

6. Conclusion

We conclude this article by discussing several open issues. First, our inference method is based on a normal approximation, which works well for a moderate sample size. For a relatively smaller sample, the bootstrap may provide

better performance (Cai and Liu (2016)). The pioneering work of Chernozhukov Chetverikov and Kato (2013) on the Gaussian approximation to the functional of high-dimensional empirical processes sheds light on applying the multiplier bootstrap to the adaptive Huber regression. Although the validity of the multiplier bootstrap for the adaptive Huber regression can be established similarly, the computational demand is more challenging.

In addition, our framework can be generalized for potentially heavy-tailed designs. In practice, in the mediation analysis involving RNA-sequencing data, for example, both the responses and the entries in the design are heavy tailed. To address this challenge, we may replace the entries in the design by their trimmed versions $X_i^{\bar{\omega}} = (\varphi_{\bar{\omega}}(x_{i1}) \ldots, \varphi_{\bar{\omega}}(x_{id}))^{\mathrm{T}}$, where $\varphi_{\bar{\omega}}(u) = \min\{\max(-\bar{\omega}, u), \bar{\omega}\}$, with the tuning parameter $\bar{\omega} > 0$. This is similar to the approach of filtering entries in the design using some thresholds (Pensia, Jog and Loh (2021)). Here, the data-driven selection on $\bar{\omega}$ is largely unknown, and cross-validation is therefore unavoidable for the implementation. At the cost of an extra tuning parameter $\bar{\omega}$ and an additional $\log(np)$ term in the orders of both τ and $\bar{\omega}$, results similar to Proposition 1 can be established, although the theoretical guarantee on $\widehat{\mathbf{\Sigma}}_j$ is more involved.

Finally, it would be challenging, yet interesting to perform a power analysis of our method to seek potential power improvement. Two approaches are possible in addition to the adaptive calibration discussed in Section 2.1. The first relies on recovering the latent common factors, in addition to the observed covariates (Fan et al. (2019)). That is, we consider a mixed-effects model $\mathbf{Y}_i = \mathbf{\Theta}\mathbf{Z}_i + \mathbf{A}\mathbf{f}_i + \boldsymbol{\epsilon}_i$, where $\mathbf{A} \in \mathbb{R}^{p\times K}$ is the loading matrix and $\mathbf{f}_i \in \mathbb{R}^K$ are zero-mean latent common factors that are unobserved. Because the common factors contribute to the common variance, the signal-to-noise ratio can therefore increase using a factor adjustment, which, in turn, improves the power. The second approach employs a more subtly designed multiple testing framework than the BH procedure. For example, Cai, Sun and Wang (2019) proposed a new covariate-assisted ranking and screening (CARS) approach that incorporates a carefully constructed auxiliary variable to improve the power. Proposition 6 in Cai, Sun and Wang (2019) indicates the applicability of the CARS approach to non-normal data. The finite fourth-moment assumption is adequate for the asymptotic normality of their statistics, but not enough for the uniform convergence of the sample means when the number of hypotheses outnumbers the sample size. An interesting future direction would be to determine whether the robustification/Huberization can be incorporated into the CARS approach to handle heavy-tailed and/or skewed data. We leave these topics for future work.

Supplementary Material

The online Supplementary Material contains the proofs of all the theoretical results in the main text, as well as additional numerical results.

Acknowledgments

We thank the editor, associate editor, and two referees for their insightful comments. Wen Zhou's research was supported by DOE DE-SC0018344, NSF Grant IOS-1922701, and NIH Grant R01GM144961. Wen-Xin Zhou's research was supported by NSF Grants DMS-1811376 and DMS-2113409.

References

Baum, M., Cohen, D. and Zhukov, Y. (2018). Does rape culture predict rape? Evidence from U.S. newspapers, 2000–2013. *Quarterly Journal of Political Science* **13**, 263–289.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289–300.

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

Bullard. J. H., Purdom, E., Hansen, K. D. and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11**, Article 94.

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

Cai, T., Cai, T. T., Liao, K. and Liu, W. (2019). Large-scale simultaneous testing of cross-covariance matrices with applications to PheWAS. *Statistica Sinica* **29**, 983–1005.

Cai, T. T. and Liu, W. (2016). Large-scale multiple testing of correlations. *Journal of the American Statistical Association* **111**, 229–240.

Cai, T. T. and Sun, W. (2017). Large-scale global and simultaneous inference: Estimation and testing in very high dimensions. *Annual Review of Economics* **9**, 411–439.

Cai, T. T., Sun, W. and Wang, W. (2019). Covariate-assisted ranking and screening for large-scale two-sample inference. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **81**, 187–234.

Cai, T. T., Zhang, C. H. and Zhou, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics* **38**, 2118–2144.

Caines, A., Pastrana, S., Hutchings, A. and Buttery, P. J. (2018). Automatically identifying the function and intent of posts in underground forums. *Crime Science* **7**, Article 19.

Chernozhukov, V., Chetverikov, D. and Kato, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics* **41**, 2786–2819.

Fan, J., Ke, Y., Sun, Q. and Zhou, W.-X. (2019). FarmTest: Factor-adjusted robust multiple testing with approximate false discovery control. *Journal of the American Statistical Association* **114**, 1880–1893.

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.

Gerlach, M. and Font-Clos, F. (2020). A standardized Project Gutenberg corpus for statistical analysis of natural language and quantitative linguistics. *Entropy* **22**, 1–14.

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

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *The Annals of Statistics* **1**, 799–821.

Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statistics*. 2nd Edition. John Wiley & Sons, Hoboken.

Khany, R. and Tazik, K. (2019). Levels of statistical use in applied linguistics research articles: From 1986 to 2015. *Journal of Quantitative Linguistics* **26**, 48–65.

Ke, Y., Minsker, S., Ren, Z., Sun, Q. and Zhou, W.-X. (2019). User-friendly covariance estimation for heavy-tailed distributions. *Statistical Science* **34**, 454–471.

Kong, D., Bondell, H. and Wu, Y. (2018). Fully efficient robust estimation, outlier detection and variable selection via penalized regression. *Statistica Sinica* **28**, 1031–1052.

Marsden, J., Budden, D., Craig, H. and Moscato, P. (2013). Language individuation and marker words: Shakespeare and his Maxwell's demon. *PLoS ONE* **8**, e66813.

Minsker, S. (2018). Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries. *The Annals of Statistics* **46**, 2871–2903.

Minsker, S. (2019). Distributed statistical estimation and rates of convergence in normal approximation. *Electronic Journal of Statistics* **13**, 5213–5252.

Nguyen, D. Q., Nguyen, D. Q., Pham, D. D. and Pham, S. B. (2016). A robust transformation-based learning approach using ripple down rules for part-of-speech tagging. *AI Communications* **29**, 409–422.

O'Connor, B., Bamman, D. and Smith, N. A. (2011). Computational text analysis for social science: Model assumptions and complexity. In *Proceedings of the NIPS Workshop on Computational Social Science and the Wisdom of Crowds*.

Pensia, A., Jog, V. and Loh, P.-L. (2020). Robust regression with covariate filtering: Heavy tails and adversarial contamination. Web: `https://arxiv.org/abs/2009.12976`.

Prasad, A., Suggala, A. S., Balakrishnan, S. and Ravikumar, P. (2020). Robust estimation via robust gradient estimation. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **82**, 601–627.

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

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W. et al.(2015). `limma` powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **43**, e47.

Robinson, M. D., McCarthy, D. J. and Smyth, G. K. (2010). `edgeR`: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140.

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

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences* **100**, 9440–9445.

Sun, Q., Zhou, W.-X. and Fan, J. (2020). Adaptive Huber regression. *Journal of the American Statistical Association* **115**, 254–265.

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, Cambridge.

Wang, L., Zheng, C., Zhou, W. and Zhou, W.-X. (2021). A new principle for tuning-free Huber regression. *Statistica Sinica* **31**, 2153–2177.

Wilkerson, J. and Casas, A. (2017). Large-scale computerized text analysis in political science: Opportunities and challenges. *Annual Review of Political Science* **20**, 529–544.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology.* Addison-Wesley Press, Cambridge.

Youngseok Song

Institute of Mathematics, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland.

E-mail: youngseok.song@epfl.ch

Wen Zhou

Department of Statistics, Colorado State University, Fort Collins, CO 80523, USA.

E-mail: riczw@stat.colostate.edu

Wen-Xin Zhou

Department of Mathematics, University of California San Diego, La Jolla, CA 92093, USA.

E-mail: wez243@ucsd.edu

(Received January 2021; accepted September 2021)