Statistica Sinica Preprint No: SS-2021-0003	
Title	Large-Scale Inference of Multivariate Regression for Heavy-Tailed and Asymmetric Data
Manuscript ID	SS-2021-0003
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202021.0003
Complete List of Authors	Youngseok Song, Wen Zhou and Wen-Xin Zhou
Corresponding Author	Wen Zhou
E-mail	riczw@stat.colostate.edu
Notice: Accepted version subject to English editing.	

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Statistica Sinica

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*

[♯]*University of California San Diego*

Abstract: Large-scale multivariate regression is a fundamental statistical tool that finds applications in a wide range of areas. This paper considers the problem of simultaneously testing a large number of general linear hypotheses, encompassing covariate-effect analysis, analysis of variance, and model comparisons. The new challenge that comes along with the overwhelmingly 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 invisible to the scope of least squares. The new testing procedure is built on 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, along with an empirical study on quantitative linguistics, demonstrate the advantage of our proposal compared to many state-of-the-art methods when the

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

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 in various fields ranging from biology, financial economics, linguistics, psychology, to social science. By modeling thousands or tens of thousands of responses and covariates or experimental factors, it provides statistical decisions on the individual levels by simultaneously testing many general linear hypotheses, including covariate-effect analysis, analysis of variance, model comparisons, etc. For example, multivariate regression has become a standard tool in the differential expression analysis in genomics (Ritchie et al., 2015), and has also been commonly used in corpus linguistics for the word usage comparison (Khany and Tazik, 2019). We refer to Cai and Sun (2017) for a more comprehensive review on relevant applications.

To simultaneously test many general linear hypotheses, a conventional practice is to compute individual p-values based on F-tests or likelihood ratio tests, and then employ multiple testing procedures to control the false

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

discovery rate (FDR, see Benjamini and Hochberg (1995); Storey (2002)). This standard approach and its theoretical validity, however, often rely on strong distributional assumptions, such as the 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 the microarray data (Purdom and Holmes, 2005) and text data (Zipf, 1949).

To overcome the above challenge, a procedure that is robust against heavy-tailed and/or skewed error distribution is desired. 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, evidenced by its better finite sample performance than a non-robust method is called a *tail-robust procedure* (Ke et al., 2019). Different from the conventional robustness under Huber's ϵ-contamination model (Huber, 1964) or the regularization-based robustness for detecting and removing outliers (Kong et al., 2018), the notion of tail-robustness focuses on the challenge that methods minimizing the empirical risk perform poorly as the empirical risk is not uniformly close to the population risk given heavy-tailed and/or skewed errors (Prasad et al., 2020). Lately, a variety of new methods and estimation theory under

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

heavy-tailed models have been developed (Catoni, 2012; Minsker, 2018; Sun et al., 2020), while less progress has been documented in terms of inference, especially in a large-scale setting (Fan et al., 2019; Minsker, 2019).

Building on the idea of *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 the approximated 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 will lead the corresponding hypotheses rejected. By allowing the robustification parameter to diverge with the sample size, the adaptive Huber regression estimator admits tight non-asymptotic deviation bound and is asymptotically efficient (Sun et al., 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 can be fully data-driven (Wang et al., 2021), and therefore is computationally attractive and applicable to real large-scale problems.

The main contributions of this paper are as follows. Methodologically,

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

we develop a tail-robust multiple testing procedure to simultaneously draw inference 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 to the traditional practice in multivariate and high-dimensional statistics, our method imposes very mild moment conditions on the data, while 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 inference on large-scale multivariate regression models with heavy-tailed and/or asymmetric errors, which is corroborated by the comprehensive simulation studies. Also, 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 a couple of interesting new technical results, by which we show that the proposed method controls the false discovery proportion and rate asymptotically under mild moment and correlation conditions on the error vector. Computationally, our method is fast by taking advantage of the computational efficiency of data-adaptive Huber regression (Wang et al., 2021). In addition to numer-

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

ical experiments, we apply our method to analyze the text data from the Standardized Gutenberg Project Corpus (Gerlach and Font-Clos, 2018). We identify the genre representative words in works of William Shakespeare, and also investigate the differences among 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. Particularly, 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 is devoted to simulations. In Section 5, we apply our method to the well-known quantitative linguistics data set, the Gutenberg Project. Extensions of our method are discussed in Section 6. All the 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 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 $\mathbf{w}$

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

and $\|\mathbf{u}\|^2 = \langle \mathbf{u}, \mathbf{u} \rangle$. Denote $\mathbb{S}^{d-1} = \{\mathbf{u} \in \mathbb{R}^d : \|\mathbf{u}\| = 1\}$ the unit sphere in $\mathbb{R}^d$. For 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 $\lambda_{\min}(\mathbf{A})$ the spectral norm, the maximum eigenvalue, and the minimum eigenvalue, respectively. Let $\Phi(z) := \mathbb{P}(U < z)$ with $U \sim N(0,1)$ be the cumulative distribution function of standard normal. Denote $\mathbb{I}(\cdot)$ the indicator function.

Suppose we observe 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 $d \geqslant 1$ and $d/n \to 0$ as $n \to \infty$. For each $j = 1, \ldots, p$, 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 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 $\mathbf{X}_i$'s, the p-dimensional residual errors $\boldsymbol{\epsilon}_i$'s are independent and identically distributed (i.i.d.) with mean zero and covariance matrix $\boldsymbol{\Sigma}_\epsilon = (\sigma_{\epsilon,jk})_{1 \leqslant j,k \leqslant p}$. To ease the 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 $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_n)^{\mathrm{T}}$, so that (2.1) reduces to $\mathbf{Y} = \mathbf{Z}\boldsymbol{\Theta} + \boldsymbol{\Xi}$. Based on (2.1), we are interested in

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

making simultaneous inference on the p general linear 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 matrix $\mathbf{C} \in \mathbb{R}^{q\times(d+1)}$ and vectors $\mathbf{c}_{0j} \in \mathbb{R}^q$ are prescribed, and $\text{rank}(\mathbf{C}) = q \leqslant d+1$. Hypotheses in (2.2) encompass a variety of important applications, including the inference on contrasts in the 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 either normal or light-tailed symmetric distributions (Friguet et al., 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 will occur more frequently, challenging the efficacy of standard methods. Throughout this paper, we will not make any 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 Huber-type M-estimators $\widehat{\boldsymbol{\theta}}_j$ as

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

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

where $\ell_\tau(x) = (x^2/2)\mathbb{I}(|x| \leqslant \tau) + (\tau|x| - \tau^2/2)\mathbb{I}(|x| > \tau)$ 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 the error only admits $(2 + \delta)$th finite moment, and more importantly, the order of τ_j grants the desired approximation error of 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 noticed in the literature (Catoni, 2012; Fan et al., 2019; Sun et al., 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 is different from the earlier studies on the adaptive Huber regressions. For example, with the finite $(1 + \epsilon)$th moment of error, Sun et al. (2020) focused on estimating the adaptive Huber regression that corresponds to $p = 1$ in our setting and considered $\tau_j = O(n^{\max\{1/(1+\epsilon),1/2\}}(d + \log n)^{-\max\{1/(1+\epsilon),1/2\}})$, while Fan et al. (2019) used $\tau_j = O(n^{1/2}\{\log(np)\}^{-1/2})$ for testing p-dimensional mean vectors under the assumption of finite fourth moment of errors, which corresponds to $d = 1$ in our setting. In practice, τ_j can be chosen by either the

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

cross-validation or the recent data-driven method by 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 $\widehat{\boldsymbol{\theta}}_j$'s, we then construct test statistics whose approximated p-values for (2.2) are obtained under the null. Partnered with the Benjamini-Hochberg (BH) method (Benjamini and Hochberg, 1995) or its variants, e.g., 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 are in position to detail our test procedure for (2.2). Given 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 $\tau_{0j} \geqslant v_{j,\delta}$ 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 $\boldsymbol{\Sigma}_j := \mathrm{cov}(n^{1/2}\widehat{\boldsymbol{\theta}}_j)$ as we will discuss in Sections 2.2 and 3.2. In (2.2), H_{0j} will be rejected for large V_j. As we will show, V_j's are asymptotically χ^2_q-distributed under H_{0j} uniformly in j. Leveraging this, we can estimate the FDP, so as to determine the rejection threshold that bounds the estimated FDP by a pre-specified level $\alpha \in (0, 1)$.

Let $\mathscr{H}_0 = \{j : 1 \leqslant j \leqslant p, H_{0j} \text{ is true}\}$ and $p_0 := |\mathscr{H}_0|$. Denote the number of discoveries and false discoveries by $R(z) = \sum_{j=1}^{p} \mathbb{I}(V_j \geqslant z)$ and

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

$V(z) = \sum_{j\in\mathcal{H}_0} \mathbb{I}(V_j \geqslant z)$, respectively, for threshold $z > 0$. The false discovery proportion 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 $p_0\mathbb{P}(\chi_q^2 > z)$ while 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 the sparse setting in the high-dimensional regime, the approximated false discovery proportion $\mathrm{AFDP}(z) = \widehat{V}(z)/\max\{R(z), 1\}$ with $\widehat{V}(z) = p\,\mathbb{P}(\chi_q^2 > z)$ is a reasonable and slightly conservative surrogate for the asymptotic approximation $p_0\mathbb{P}(\chi_q^2 > z)/\max\{R(z), 1\}$ and $\mathrm{FDP}(z)$. Using $\mathrm{AFDP}(z)$, we can determine threshold $\widehat{z}_\alpha = \inf\{z \geqslant 0 : \mathrm{AFDP}(z) \leqslant \alpha\}$ for the nominal level α. For $j = 1, \ldots, p$, H_{0j} will be rejected whenever $V_j \geqslant \widehat{z}_\alpha$. Essentially, our procedure is build upon the BH method with input p-values obtained from robustified/Huberized test statistics. Similar ideas have also been adopted in Cai and Liu (2016) and Cai et al. (2019). The main difference is that the test statistics used in Cai and Liu (2016) and Cai et al. (2019) have closed-form expressions, while our statistics are based on M-estimators.

Of note, if $\pi_0 = p_0/p$ is bounded away from 1 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)

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

estimates $V(z)$ by $p\widehat{\pi}_0(\eta)\,\mathbb{P}(\chi_q^2 > z)$ 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. Among a few studies, 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 by $\widehat{z}_\alpha^{\eta} = \inf\{z \geqslant 0 : p\widehat{\pi}_0(\eta)\,\mathbb{P}(\chi_q^2 > z)/R(z) \leqslant \alpha\}$.

2.2 A refined Huber-type estimator of $\mathbf{\Sigma}_j$

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

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

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

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

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

$$\hat{\mathbf{\Sigma}}_j = \frac{\sum_{i=1}^n \{\ell'_{\tau_j}(e_{ij})\}^2}{(n-d-1)K_j}\mathbf{W}_j^{-1}\hat{\mathbf{\Sigma}}_Z\mathbf{W}_j^{-1}, \tag{2.6}$$

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

For the conventional Huber regression with fixed $\tau > 0$, it can be shown that, with $\mathbb{I}^*_\tau(x)$ replaced by $\mathbb{I}(|x| \leqslant \tau)$, $\hat{\mathbf{\Sigma}}_j$ converges in probability to $\mathbf{\Sigma}_\tau$ as $n \to \infty$. To legitimize the use of V_j for testing (2.2), we will show in Section 3.2 that with adaptive τ_j, the covariance estimator $\hat{\mathbf{\Sigma}}_j$ in (2.6) is

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

close to $\mathbf{\Sigma}_j$ uniformly over j with high probability. In addition, as 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 that is similarly determined in τ (Wang et al., 2021) or as a minimum absolute deviation estimator of the variance using the fitted residuals. We refer to Section S5.3 in the supplement 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). Besides the robust multiple inference, Fan et al. (2019) focused more 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 inference 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 design $\mathbf{Z}$, whereas Fan et al. (2019) only considered $\mathbf{Z} = \mathbf{1} \in \mathbb{R}^{n\times 1}$, which is not trivially extendable to the problem under our consideration.

Our estimator in (2.6) bridges the gap, and it consists of two parts: the

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

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 $K_j^{-1}\mathbf{W}_j^{-1}\hat{\mathbf{\Sigma}}_Z\mathbf{W}_j^{-1}$ offers a robustification of inverse Gram matrix $\left(n^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}$. It 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 the Huber's loss to specify $\mathbf{W}_j$ and K_j, our estimator is continuous, which is crucial for the uniform consistency of $\hat{\mathbf{\Sigma}}_j$ across j. The uniform consistency of $\hat{\mathbf{\Sigma}}_j$'s leads to the FDP control of our robust multiple test for large-scale multivariate regressions. In contrast, additional to the fact that the procedure by Fan et al. (2019) is not able to exploit $\mathbf{Z}$ when $d \geqslant 1$, the variance estimator of $\sigma_{\epsilon,jj}$ in Fan et al. (2019), 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 and moreover, it requires extra tuning parameters for robustly estimating the second order moment. A numerical experiment is reported in Section S5.5 in the supplement to verify the above discussions.

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

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

$\hat{\mathbf{\Sigma}}_j$ in (2.6). Hereafter, we focus on $\mathbf{Z}_i$ being random (except for the first coordinate), and leave the results under fixed designs to the supplement.

3.1 Approximation of FDP with known Σ_j

Assume the covariance matrix $\mathbf{\Sigma}_j$ is known for each j. Consider the oracle test statistic $V_j^\circ = n(\mathbf{C}\hat{\boldsymbol{\theta}}_j - \mathbf{c}_{0j})^{\mathrm{T}}(\mathbf{C}\mathbf{\Sigma}_j\mathbf{C}^{\mathrm{T}})^{-1}(\mathbf{C}\hat{\boldsymbol{\theta}}_j - \mathbf{c}_{0j})$. Given $z \geqslant 0$, write $R^\circ(z) = \sum_{j=1}^p \mathbb{I}(V_j^\circ > z)$, $V^\circ(z) = \sum_{j\in\mathscr{H}_0} \mathbb{I}(V_j^\circ > z)$, 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

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

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 $\mathbf{R}_\epsilon = (r_{\epsilon,jk})_{1\leqslant j,k\leqslant p}$ 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 constants $\delta \in (0,2]$ and $c_\epsilon, C_\epsilon > 0$ such that $c_\epsilon \leqslant \min_{1\leqslant j\leqslant p}\sigma_{\epsilon,jj}^{1/2} \leqslant \max_{1\leqslant j\leqslant p} \upsilon_{j,\delta} \leqslant C_\epsilon$; and (iv) there exist $\kappa_0 \in (0,1)$ and $\kappa_1 > 0$ such that $\max_{1\leqslant j\neq k\leqslant p}|r_{\epsilon,jk}| \leqslant \kappa_0$ and $p^{-2}\sum_{1\leqslant j\neq k\leqslant p}|r_{\epsilon,jk}| = O(p^{-\kappa_1})$.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

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 to the traditional settings that presumes the finite fourth or higher order moments of errors, (iii) only assumes 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 i.i.d. predictors $\{\mathbf{Z}_i\}_{i=1}^n$ are sub-Gaussian, i.e., for some $A_0 > 0$, $\mathbb{P}(|\langle \mathbf{u}, \widetilde{\mathbf{Z}}_i \rangle| \geqslant A_0\|\mathbf{u}\|t) \leqslant 2\exp(-t^2)$ for any $\mathbf{u} \in \mathbb{R}^{d+1}$ and $t \geqslant 0$.

We refer to Vershynin (2018) for an overview about sub-Gaussian vectors. Under Conditions 1 and 2, Proposition 1 shows that AFDP° in (3.1) consistently estimates FDP°, and it provides the guideline to establish the FDP control and serves as the cornerstone to the guarantees of our method.

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

We conclude this subsection with two remarks. If we strengthen Condition 1 (iii) to uniformly bounded k-th moments for $k \geqslant 4$, 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

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

addition, to prove Proposition 1, we will 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 non-trivial and reveals how the parameter q, which corresponds to the dimension of the hypothesis, affects the difficulty of testing (2.2). We will revisit this via numerical studies in Section 4.

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

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

Theorem 1. *Suppose that the conditions of Proposition 1 hold. As long as the estimated covariances* $\{\tilde{\mathbf{\Sigma}}_j\}_{j=1}^p$ *satisfy* $\max_{1 \leqslant j \leqslant p} \|\tilde{\mathbf{\Sigma}}_j - \mathbf{\Sigma}_j\| = o_{\mathbb{P}}\{(\log(np) + d)^{-1}\}$, *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 $\hat{\mathbf{\Sigma}}_j$'s in (2.6) satisfy the required accuracy in Theorem 1, together with Proposition 1, theorem below acquires the convergence in probability of approximated FDP to true FDP for any $z > 0$ as $n, p \to \infty$.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

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

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

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

Theorem 2 implies that the required accuracy in Theorem 1, that is, $\max_{1 \leqslant j \leqslant p} \|\widehat{\boldsymbol{\Sigma}}_j - \boldsymbol{\Sigma}_j\| = o_{\mathbb{P}}\{(\log(np) + d)^{-1}\}$, 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 $\widehat{\boldsymbol{\Sigma}}_j$ in (2.6). In fact, the conclusion in Theorem 2 remains valid for some variant of $\widehat{\boldsymbol{\Sigma}}_j$, such as $\widehat{\boldsymbol{\Sigma}}_j^{(1)} = \sum_{i=1}^n \{\ell'_{\tau_j}(e_{ij})\}^2 \{(n - d - 1)m_j\}^{-1} K_j \mathbf{W}_j^{-1}$.

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's selected via five-fold cross-validation (Sun et al., 2020); (iii) least squares based mul-

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

tiple testing method; (iv) empirical Bayes based multiple testing method implemented via `limma` (Ritchie et al., 2015); (v) `limma` with the traditional robust M-estimation instead of the least squares; and (vi) empirical Bayes based multiple testing method for count data implemented via `edgeR` (Robinson, McCarthy, and Smyth, 2010). Both `limma` and `edgeR` are widely-used software to analyze a large number of regression models, and serve as benchmarks in genomics studies. `limma` employs empirical Bayes methods to shrink individual variances towards a common value in the hope of better controlling the FDR. `edgeR` models count data with large variations via the negative binomial model. To implement `edgeR`, we round response Y_{ij} to its nearest integer. For our method, we set $\delta = 2$ in (2.3) (i.e., 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 employed the FDR controlling procedure by 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 independently drawn from $N(0, 1)$, and each column is standardized to have zero mean and unit variance. We consider three heavy-tailed and highly skewed error distribu-

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

tions: (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 proportion 0.7 and 0.3, respectively. Setting (c) reflects more challenging scenarios in practice as t_2 does not have finite second moment. Under each setting, we first generate $\mathbf{E} = (\epsilon_{ij})_{1\leqslant i\leqslant n, 1\leqslant j\leqslant p}$ with i.i.d. entries. To incorporate dependence, set $\mathbf{\Xi} = 100\mathbf{R}_\epsilon^{1/2}\mathbf{E}$, where the correlation matrix $\mathbf{R}_\epsilon = (r_{\epsilon,jk})_{1\leqslant j,k\leqslant p}$ admits one of the three structures: *Model* 1, the identity matrix; *Model* 2, $r_{\epsilon,ij} = r_{\epsilon,ji}$ independently drawn from $0.3 \times \text{Ber}(0.1)$ 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 $k \geqslant 3$. Notice that *Model* 2 does not satisfy Condition 1 (iv). Together with results in Section S5.4 in the supplement, results for *Model* 2 show that our method is reliable even when Condition 1 (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 $\beta_{jk} \sim \text{Unif}(-150, 150)$ for $1 \leqslant j \leqslant p$ and $1 \leqslant k \leqslant d-1$, $\beta_{jd} = -\sum_{k=1}^{d-1}\beta_{jk}$ for $1 \leqslant j \leqslant p - p_1$ 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 $p - p_1 + 1 \leqslant j \leqslant p$, where W_j are Rademacher variables. For *Hypothesis* 2, let $\boldsymbol{\beta}_j = 0$ for $1 \leqslant j \leqslant p - p_1$, and $\beta_{jk} = (2d^{-1})^{1/2}\delta W_{jk}$ for $p - p_1 + 1 \leqslant j \leqslant p$ and $1 \leqslant k \leqslant d$, where W_{jk} are Rademacher variables.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

We take $\delta = 75\eta$ and $\eta = 0.3$, which determine the signal strength.

4.2 Numerical performance

We take 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 the power under *Model* 2 with $p = 1000$ and $d = 6$. Results under other settings are documented in Section S5 in the supplementary material. Each point corresponds to a nominal level (marked as a vertical gray dashed line) with x- and y-axes representing, respectively, the empirical FDR and the power. 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, control the FDR well in general and maintain 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 the use of 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 take precautions

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

when use 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 log-normal and t_2. Overall, the power of all methods increases with n, and drops for larger p, see Figures S1-S11 in the supplement. As 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) as they fail to control the FDR. In 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 $[0.3, 0.5]$ 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 dominates cross-validation. With heavier tails (mixture of log-normal and t_2), the power slightly decreases for all methods.

5. Real Data Analysis: The Gutenberg Project

Inference on large-scale text data from literary publications has drawn growing attention and it has provided novel and revealing discoveries in a variety of fields such as sociology (O'Connor et al., 2011), political science (Wilker-

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

son and Casas, 2017; Baum et al., 2018), criminology (Caines et al., 2018) as well as linguistics. A major task in text analysis is to identify word markers to distinguish or identify different authors, cultures, or resources, etc. These word markers are usually identified by small p-values from testing regression coefficients that are 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 the authorship classification. Here, we consider hypotheses that help identify the distinctive word markers, which are referred as "differentially represented" words, to distinguish authors or different writing styles of a particular author.

As a well-known public accessible digital library to literary publications, the Project Gutenberg is founded in 1971, and offers 60156 e-books as of September 03, 2019. The Standardized Project Gutenberg Corpus (SPGC, Gerlach and Font-Clos (2018)) 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. Also, it contains metadata about books, such as the author information, subject categories, and book types.

We apply our method to word counts from SPGC to find the idiosyn-

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

cratic word markers to represent an author or a category of publications. Specifically, we consider two problems: a comparison of works of Lewis Carroll, Charles Dickens, and Arthur Conan Doyle, and the study of works of William Shakespeare. Table S1 in the supplementary material provides a snapshot of the raw data. From the histograms of empirical kurtosis of word counts (Figure S20), the data is heavy-tailed in both book-wise and word-wise. For pre-processing, we first merge word counts across books, and then remove the words whose total count is less than half the number of books or those only 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). More details are deferred to the supplementary files.

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 the word usage in Figure S20 in the supplementary files. To identify differentially represented 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 authored by Dickens, and $\mathbf{X}_i = (1, -1, -1)^{\mathrm{T}}$ if it is written by Conan Doyle for $i = 1, \ldots, 167$ books, and $\boldsymbol{\beta}_j = (\mu_j, \alpha_{1j}, \alpha_{2j})^{\mathrm{T}}$ for $j = 1, \ldots, 6839$ words. We consider the following linear hypotheses: (Hypothesis CDD1) H_{0j} :

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

$[(0\ 1\ 0)^{\mathrm{T}}\ (0\ 0\ 1)^{\mathrm{T}}]^{\mathrm{T}}\boldsymbol{\beta}_j = 0$ 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 altogether, while the other hypotheses compare one author with the remaining two. With a nominal level 0.5%, our method detects 2595, 419, 1388, and 1445 differentially represented words for each hypothesis. The top 10 differentially represented words for the three authors, such as "being" and "sprang", are displayed in Figure 4(a), while the overall comparison is reported in a Venn diagram in Figure S21 in the supplement. It is worth noticing that Conan Doyle favored "sprang" while Carroll and Dickens barely used it. In Figure 4(c), we further report the percentages of differentially represented (DR) words and non-differentially represented (NDR) words within each speech category (Nguyen et al., 2016). Differentially represented words among the three authors have higher percentages in adjectives, adverbs, and pronouns than non-differentially represented words. In contrast, differentially represented words have lower percentages in nouns, proper nouns, and verbs.

Next, we investigate the genre difference among works of Shakespeare based on three subject groups: poetry, non-historical drama, and historical

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

drama. We model the normalized word counts by (2.1) with $\mathbf{X}_i = (1,0,0)^{\mathrm{T}}$ if the ith book is a poetry, $\mathbf{X}_i = (1,1,0)^{\mathrm{T}}$ if it is a non-historical drama, and $\mathbf{X}_i = (1,1,1)^{\mathrm{T}}$ if it is a 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 0.5%, our method identifies 724 and 225 DR words for each hypothesis. As a vast amount of historical dramas of Shakespeare are about kings of the Kingdom of England, the words "princely", "London", "king", and "crown" appear more in the historical dramas (Figure 4(b)). In addition, Shakespeare used vocabularies 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 in his non-historical dramas. From Figure 4(d), the DR words between historical and non-historical dramas of Shakespeare have higher percentages in nouns, pronouns, and proper nouns, whereas their percentages are lower in adjectives, adverbs, and verbs.

In summary, our method provides a reliable addition to the existing

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

toolkit in corpus linguistics and text/literature analysis. It can be used to analyze a large volume of individual words, which extends the current state-of-art that focuses on the overall distribution of word counts. An interesting follow-up work is to investigate how do the stopping words, such as "upon", affect the results and whether their removal will alter the discovery.

6. Discussions

We conclude this article by discussing several open issues. First, our inference method is based on the normal approximation, which works well for a moderate sample size. For a relatively smaller sample, the bootstrap may provide better performances (Cai and Liu, 2016). The pioneering work of Chernozhukov et al. (2013) on the Gaussian approximation to the functional of high dimensional empirical processes sheds light on the application of multiplier bootstrap to the adaptive Huber regression. While the validity of multiplier bootstrap for the adaptive Huber regression can be established similarly, the computational demand is more challenge.

In addition, our framework can be generalized for potentially heavy-tailed designs. In practice, take the mediation analysis involving the RNA-sequencing data for example, both the responses and entries in the design are heavy-tailed. To tackle this challenge, we may replace the entries in the

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

design by its trimmed version $X_i^{\bar{\omega}} = (\varphi_{\bar{\omega}}(x_{i1}) \ldots, \varphi_{\bar{\omega}}(x_{id}))^{\mathrm{T}}$, where $\varphi_{\bar{\omega}}(u) = \min\{\max(-\bar{\omega}, u), \bar{\omega}\}$ with tuning parameter $\bar{\omega} > 0$. This is similar to the approach of filtering entries in the design using some thresholds (Pensia et al., 2021). Here, the data driven selection on $\bar{\omega}$ is largely unknown and cross-validation is therefore unavoidable for implementations. At the cost of 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 while the theoretical guarantee on $\hat{\mathbf{\Sigma}}_j$ is more involved.

Finally, it is challenging yet interesting to perform power analysis of our method to seek for the 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. As the common factors contribute to the common variance, the signal-to-noise ratio can therefore increase via factor-adjustment, which in turns improves the power. The second approach employs a more subtly designed multiple testing framework than the BH procedure. For example, recently, Cai et al. (2019) has proposed a new covariate-assisted ranking and screening (CARS)

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

approach, which incorporates a carefully constructed auxiliary variable to improve the power. Proposition 6 in Cai et al. (2019) indicates the applicability of CARS 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 sample means when the number of hypotheses outnumbers the sample size. An interesting future direction is to see if the robustification/Huberization can be incorporated in CARS to handle the heavy-tailed and/or skewed data. We leave these for future work.

Supplementary Materials

The supplementary materials contain the proofs of all the theoretical results in the main text and additional numerical results.

Acknowledgements

We thank the editor, the associate editor and two referees for insightful comments that have substantially improved this article. Wen Zhou's research is supported by the Department of Energy, NSF Grant IOS-1922701, and NIH Grant R01GM144961. Wen-Xin Zhou's research is supported by NSF Grants DMS-1811376 and DMS-2113409.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

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* **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.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

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* **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**, 1–14.

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.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

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. (2018). 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. *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 ed.). John Wiley & Sons, Inc., Hoboken, NJ.

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: Newly accepted Paper
(accepted author-version subject to English editing)

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.

Maronna, R. A., Martin, R. D., Yohai, V. J., and Salibiàn-Barrera, M. (2019). *Robust Statistics: Theory and Methods (with R)*. John Wiley & Sons, Inc., Hoboken, NJ.

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 *Second Workshop on Computational Social Science and the Wisdom of Crowds (NIPS 2011)*.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Pensia, A., Jog, V., and Loh, P.-L. (2020). Robust regression with covariate filtering: Heavy tails and adversarial contamination. *Preprint*. `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* **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., and Smyth, G. K. (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* **64**, 479–498.

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

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

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, United Kingdom.

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.

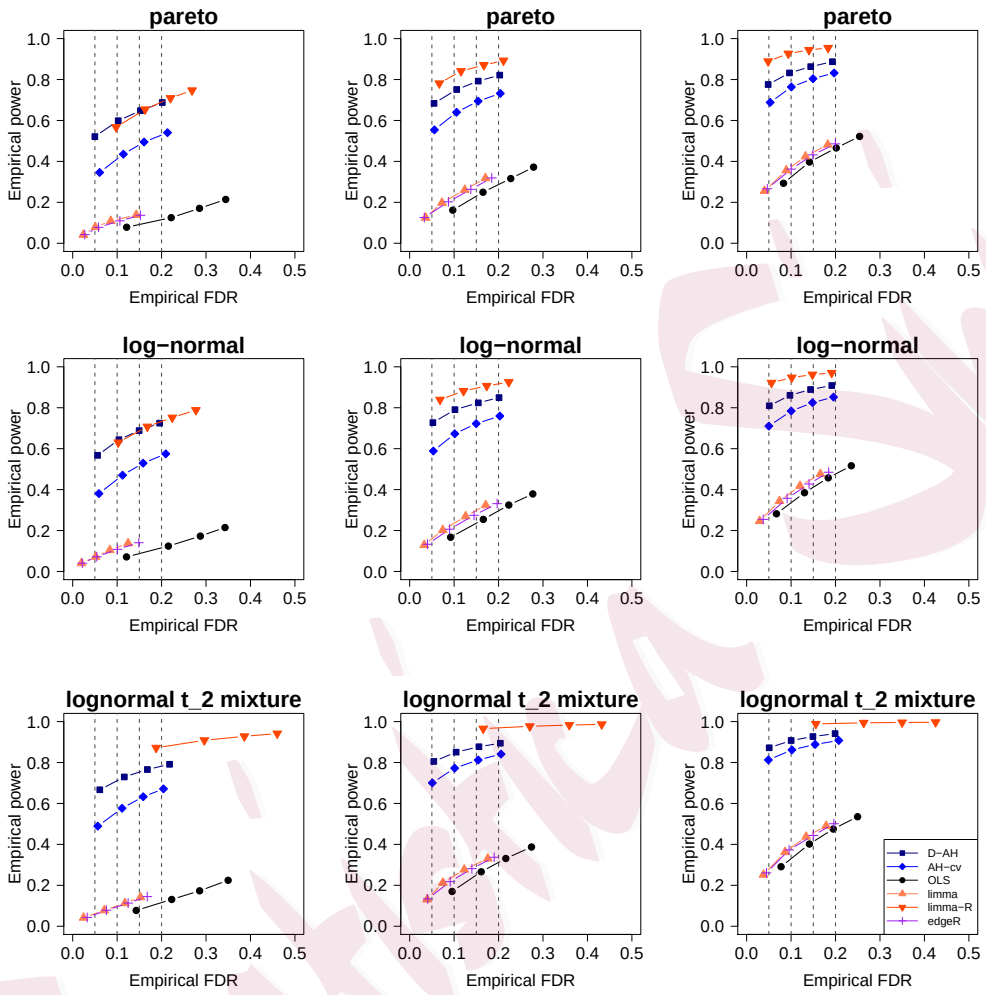
Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Ravenio Books.

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, Colorado 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)

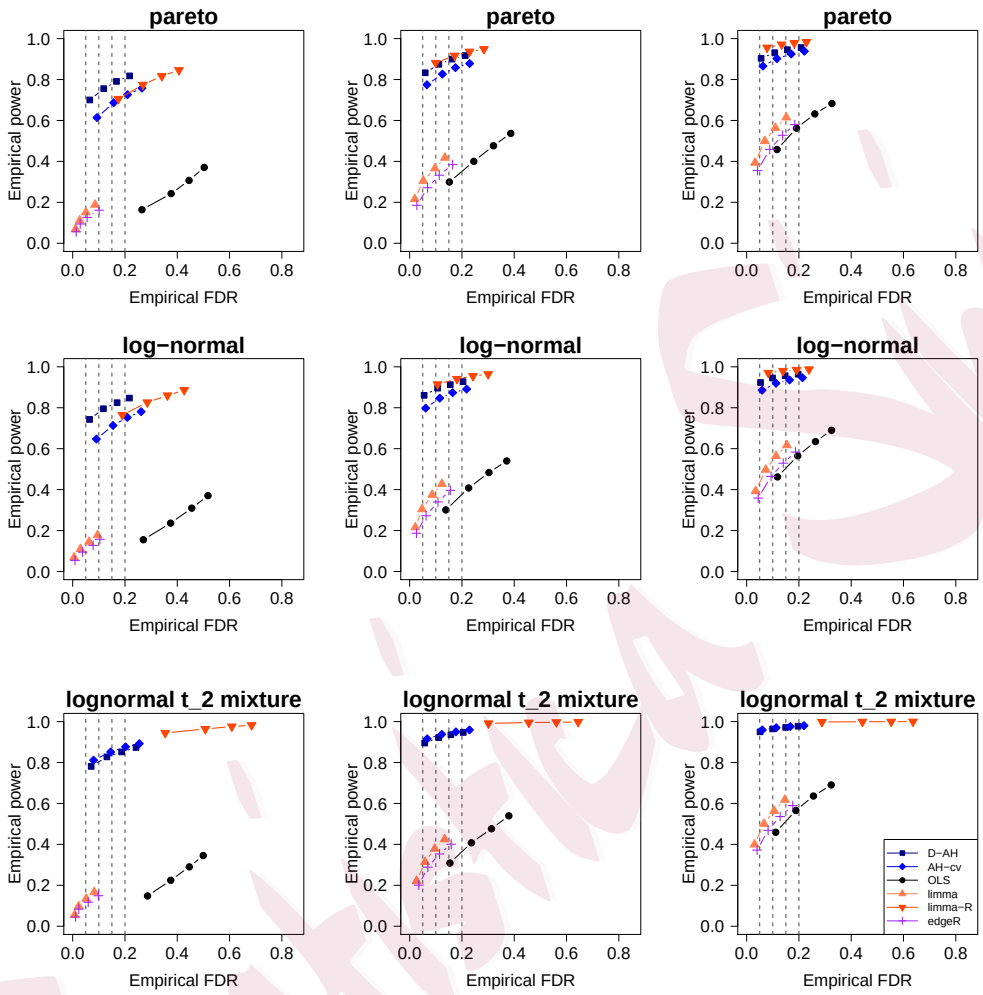
Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)



(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 $d = 6$ by six methods: the proposed method with data-adaptive Huber regression (D-AH, ■); the proposed method with cross-validation (AH-cv, ⧫); the least squares method (OLS, •); `limma` (▲); `limma` with robust regression (`limma`-R, ▾); and `edgeR` (+). Each point corresponds to a nominal FDR level (marked as a vertical gray dashed line) with x-axis representing the empirical FDR and y-axis denoting the power.

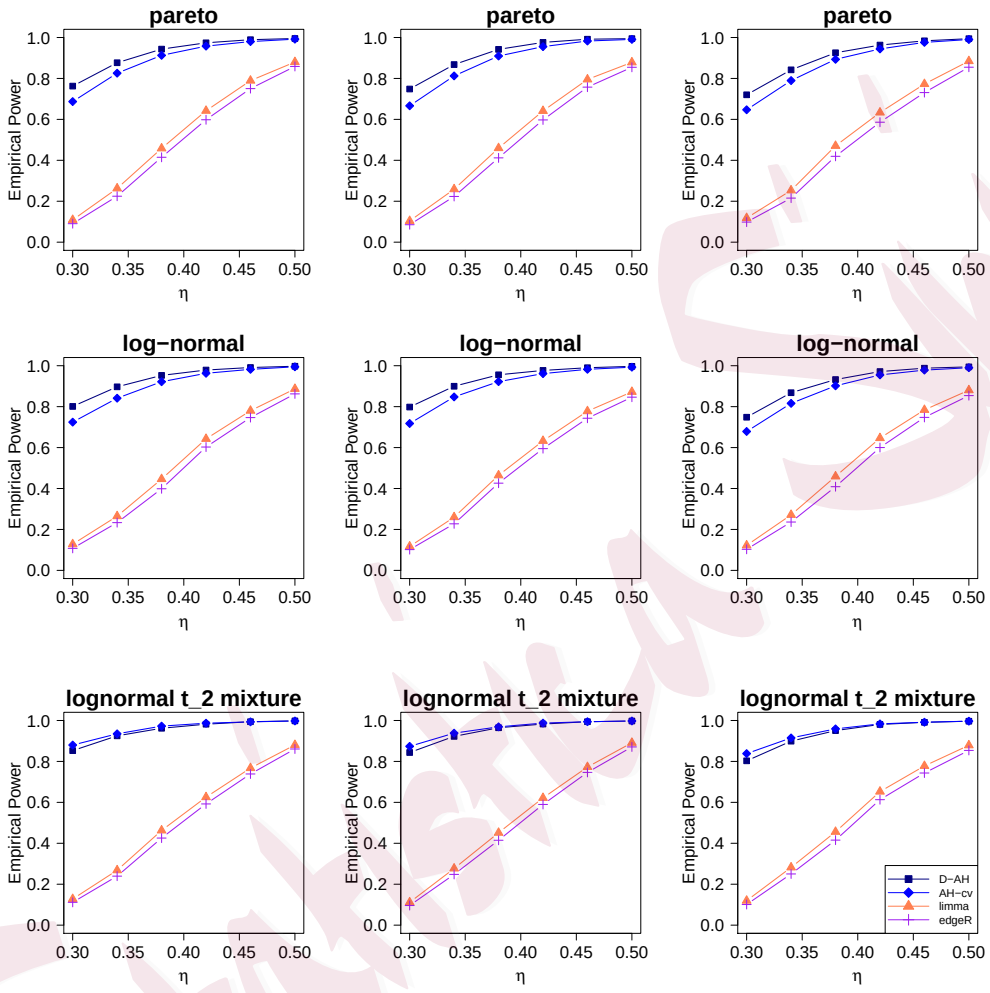
Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)



(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 $d = 6$ by six methods: the proposed method with data-adaptive Huber regression (D-AH, ■); the proposed method with cross-validation (AH-cv, ⧫); the least squares method (OLS, •); `limma` (▲); `limma` with robust regression (`limma`-R, ▾); and `edgeR` (+). Each point corresponds to a nominal FDR level (marked as a vertical gray dashed line) with x-axis representing the empirical FDR and y-axis denoting the power.

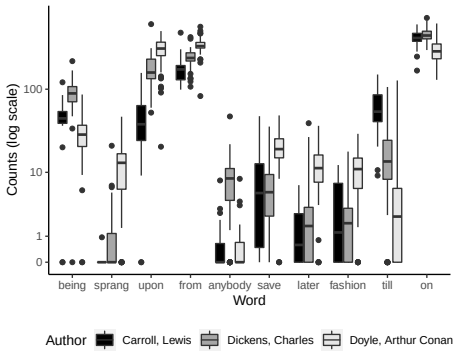
Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)



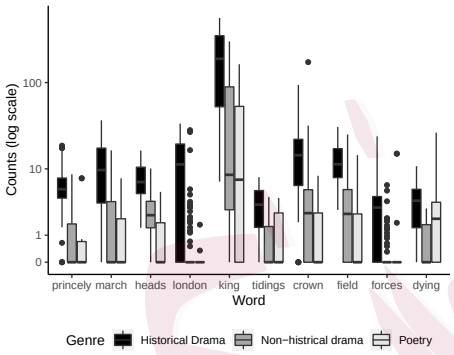
(a) *Model* 1 (b) *Model* 2 (c) *Model* 3

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

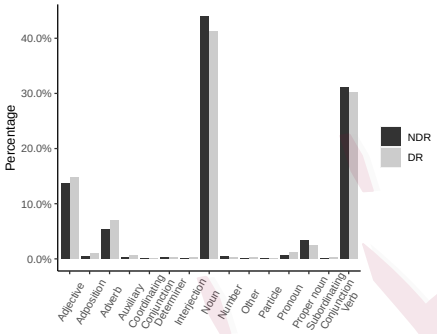
Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)



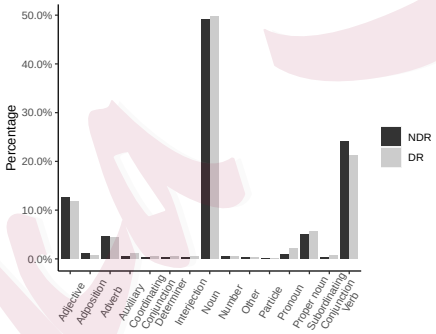
(a) Hypothesis CDD1



(b) Hypothesis WS1



(c) Hypothesis CDD1



(d) Hypothesis WS1

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