The Annals of Statistics
2017, Vol. 45, No. 5, 1863–1894
DOI: 10.1214/16-AOS1511

CONFOUNDER ADJUSTMENT IN MULTIPLE HYPOTHESIS TESTING

BY JINGSHU WANG*, QINGYUAN ZHAO*, TREVOR HASTIE†,[1] AND ART B. OWEN†,[2]

*University of Pennsylvania** *and Stanford University*†

We consider large-scale studies in which thousands of significance tests are performed simultaneously. In some of these studies, the multiple testing procedure can be severely biased by latent confounding factors such as batch effects and unmeasured covariates that correlate with both primary variable(s) of interest (e.g., treatment variable, phenotype) and the outcome. Over the past decade, many statistical methods have been proposed to adjust for the confounders in hypothesis testing. We unify these methods in the same framework, generalize them to include multiple primary variables and multiple nuisance variables, and analyze their statistical properties. In particular, we provide theoretical guarantees for RUV-4 [Gagnon-Bartsch, Jacob and Speed (2013)] and LEAPP [*Ann. Appl. Stat.* **6** (2012) 1664–1688], which correspond to two different identification conditions in the framework: the first requires a set of "negative controls" that are known a priori to follow the null distribution; the second requires the true nonnulls to be sparse. Two different estimators which are based on RUV-4 and LEAPP are then applied to these two scenarios. We show that if the confounding factors are strong, the resulting estimators can be asymptotically as powerful as the oracle estimator which observes the latent confounding factors. For hypothesis testing, we show the asymptotic z-tests based on the estimators can control the type I error. Numerical experiments show that the false discovery rate is also controlled by the Benjamini–Hochberg procedure when the sample size is reasonably large.

1. Introduction. Multiple hypothesis testing has become an important statistical problem for many scientific fields, where tens of thousands of tests are typically performed simultaneously. Traditionally, the tests are assumed to be independent of each other, so the false discovery rate (FDR) can be easily controlled by, for example, the Benjamini–Hochberg procedure [8]. Recent years have witnessed an extensive investigation of multiple hypothesis testing under dependence, ranging from permutation tests [33, 60], positive dependence [9], weak dependence [14, 56], accuracy calculation under dependence [18, 44] to mixture models [19,

Received August 2015; revised January 2016.
[1]Supported in part by NSF Grant DMS-14-07548 and NIH Grant 5R01-EB-001988-21.
[2]Supported in part by NSF Grant DMS-15-21145.
MSC2010 subject classifications. Primary 62J15; secondary 62H25.
Key words and phrases. Empirical null, surrogate variable analysis, unwanted variation, batch effect, robust regression.

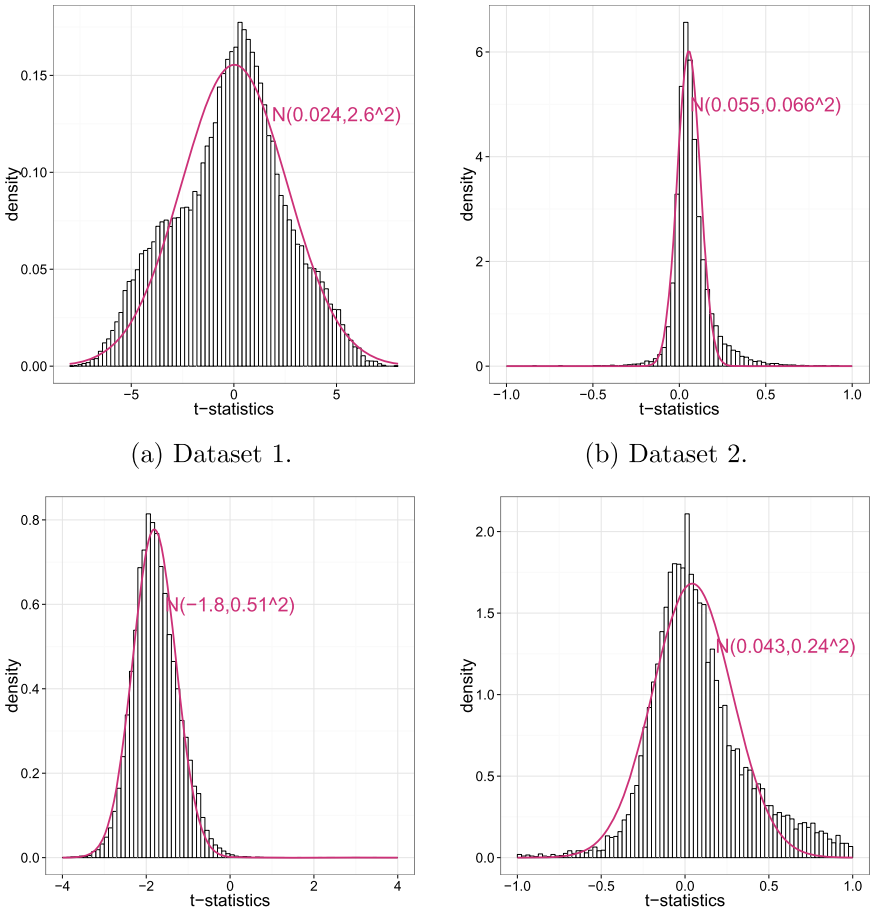
57] and latent factor models [20, 21, 35]. Many of these works provide theoretical guarantees for FDR control under the assumption that the individual test statistics are valid and may even be correlated.

In this paper, we investigate a more challenging setting. The test statistics may be correlated with each other due to latent factors and those latent factors may also be correlated with the variable of interest. As a result, the test statistics are not only correlated but are also confounded. We use the phrase "confounding" to emphasize that these latent factors can significantly bias the individual p-values, therefore, this problem is fundamentally different from the literature in the previous paragraph and poses an immediate threat to the reproducibility of the discoveries. Many confounder adjustment methods have already been proposed for multiple testing over the last decade [25, 39, 49, 59]. Our goal is to unify these methods in the same framework and study their statistical properties.

The confounding problem. We start with three real data examples to illustrate the confounding problem. The first microarray data [Figure 1(a)] is used by Singh et al. [55] to identify candidate genes associated with a chronic lung disease called emphysema. The second [Figure 1(b) and (d)] and third [Figure 1(c)] data are used by Gagnon-Bartsch, Jacob and Speed [25] to study the performance of various confounder adjustment methods. For each dataset, we plot the histogram of t-statistics of a simple linear model that regresses the gene expression on the variable of interest (disease status for the first and gender for the second and third datasets). These statistics are commonly used in genome-wide association studies (GWAS) to find potentially interesting genes. See Section 6.2.1 for more detail of these datasets.

The histograms of t-statistics in Figure 1 clearly depart from the approximate theoretical null distribution N(0, 1). The bulk of the test statistics can be skewed [Figure 1(a) and (b)], overdispersed [Figure 1(a)], underdispersed [Figure 1(b) and (d)] or noncentered [Figure 1(c)]. In these cases, neither the theoretical null N(0, 1), nor even the empirical null as shown in the histograms, look appropriate for measuring significance. Schwartzman [52] proved that a largely overdispersed histogram like Figure 1(a) cannot be explained by correlation alone, and is possibly due to the presence of confounding factors. For a sneak preview of the confounder adjustment, the reader can find the histograms after our confounder adjustment in Figure 3 at the end of this paper. The p-values of our test of confounding (Section 3.3.2) in Table 2 indicate that all the three datasets suffer from confounding latent factors.

Other common sources of confounding in gene expression profiling include systematic ancestry differences [49], environmental changes [22, 27] and surgical manipulation [41]. See [36] for a survey. In many studies, especially for observational clinical research and human expression data, the latent factors, either genetic or technical, are confounded with primary variables of interest due to the observational nature of the studies and heterogeneity of samples [50, 51]. Similar confounding problems also occur in other high-dimensional datasets such as brain imaging [53] and metabonomics [15].



(a) Dataset 1. (b) Dataset 2.

(c) Dataset 3. (d) Dataset 2 (batch correction).

FIG. 1. *Dataset* 1 *is the COPD dataset* [55]. *Dataset* 2 *and* 3 *are from* [25]. *Histograms of regression t-statistics in three microarray studies show clear departure from the theoretical null distribution* N(0, 1). *The mean and standard deviation of the normal approximation are obtained from the median and median absolute deviation of the statistics. See Section* 6.2 *for the empirical distributions after confounder adjustment.*

Previous methods. As early as [1], principal component analysis has been suggested to estimate the confounding factors. This approach can work reasonably well if the confounders clearly stand out. For example, in population genetics, [49] proposed a procedure called EIGENSTRAT that removes the largest few principal components from their SNP genotype data, claiming they closely resemble the ancestry difference. In gene expression data, however, it is often unrealistic to assume they always represent the confounding factors. The largest principal com-

ponent may also correlate with the primary effects of interest. Therefore, directly removing them can result in loss of statistical power.

More recently, an emerging literature considers the confounding problem in similar statistical settings and a variety of methods have been proposed for confounder adjustment [24–26, 38, 39, 59]. These statistical methods are shown to work better than the EIGENSTRAT procedure for gene expression data. However, little is known about their theoretical properties. Indeed, the authors did not focus on model identifiability and rely on impressive heuristic calculations to derive their estimators. In this paper, we address the identifiability problem, rederive the estimators in [25, 59] in a more principled way and provide theoretical guarantees for them.

Before describing the modeling framework, we want to clarify our terminology. The confounding factors or confounders considered in the present paper are referred to by different names in the literature, such as "surrogate variables" [38], "latent factors" [24], "batch effects" [37], "unwanted variation" [26] and "latent effects" [59]. We believe they are all describing the same phenomenon: that there exist some unobserved variables that correlate with both the primary variable(s) of interest and the outcome variables (e.g., gene expression). This problem is generally known as confounding [23, 32]. A famous example is Simpson's paradox. The term "confounding" has multiple meanings in the literature. We use the meaning from [28]: "a mixing of effects of extraneous factors (called confounders) with the effect of interest."

Statistical model of confounding. Most of the confounder adjustment methods mentioned above are built around the following model:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}^T + \mathbf{Z}\boldsymbol{\Gamma}^T + \mathbf{E}. \tag{1.1}$$

Here, $\mathbf{Y}$ is a $n \times p$ observed matrix (e.g., gene expression); $\mathbf{X}$ is an $n \times 1$ observed primary variable of interest (e.g., treatment-control, phenotype, health trait); $\mathbf{Z}$ is an $n \times r$ latent confounding factor matrix; $\mathbf{E}$ is often assumed to be a Gaussian noise matrix. The $p \times 1$ vector $\boldsymbol{\beta}$ contains the primary effects we want to estimate.

Model (1.1) is very general for multiple testing dependence. Leek and Storey [39], Proposition 1, suggest that multiple hypothesis tests based on linear regression can always be represented by (1.1) using sufficiently many factors. However, equation (1.1) itself is not enough to model confounded tests. To elucidate the concept of confounding, we need to characterize the relationship between the latent variables $\mathbf{Z}$ and the primary variable $\mathbf{X}$. To be more specific, we assume the regression of $\mathbf{Z}$ on $\mathbf{X}$ also follows a linear relationship:

$$\mathbf{Z} = \mathbf{X}\boldsymbol{\alpha}^T + \mathbf{W}, \tag{1.2}$$

where $\mathbf{W}$ is a $n \times r$ random noise matrix independent of $\mathbf{X}$ and $\mathbf{E}$ and the $r \times 1$ vector $\boldsymbol{\alpha}$ characterizes the extent of confounding in this data. By plugging (1.2) in

(1.1), the linear regression of $\mathbf{Y}$ on $\mathbf{X}$ gives an unbiased estimate of the marginal effects

$$\boldsymbol{\tau} = \boldsymbol{\beta} + \boldsymbol{\Gamma}\boldsymbol{\alpha}. \tag{1.3}$$

When $\boldsymbol{\alpha} \neq \mathbf{0}$, $\boldsymbol{\tau}$ is not the same as $\boldsymbol{\beta}$ by (1.3). In this case, the data $(\mathbf{X}, \mathbf{Y})$ are confounded by $\mathbf{Z}$. Since the confounding factors $\mathbf{Z}$ are data artifacts in this model, the statistical inference of $\boldsymbol{\beta}$ is much more interesting than that of $\boldsymbol{\tau}$. See Section 5.2 for more discussion on the marginal and the direct effects.

Following LEAPP [59], we use a QR decomposition to decouple the estimation of $\boldsymbol{\Gamma}$ from $\boldsymbol{\beta}$. The inference procedure splits into the following two steps:

Step 1. By regressing out $\mathbf{X}$ in (1.1), $\boldsymbol{\Gamma}$ is the loading matrix in a factor analysis model and can be efficiently estimated by maximum likelihood.

Step 2. Equation (1.3) can be viewed as a linear regression of the marginal effects $\boldsymbol{\tau}$ on the factor loadings $\boldsymbol{\Gamma}$. To estimate $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, we replace $\boldsymbol{\tau}$ by its observed value and $\boldsymbol{\Gamma}$ by its estimate in Step 1.

As mentioned before, other existing confounder adjustment methods including SVA [39] and RUV-4 [25] can be unified in this two-step statistical procedure. See Section 5.3 for a detailed discussion of these methods.

Contributions. Our first contribution in Section 2 is to establish identifiability for the confounded multiple testing model. In the first step of estimating factor loadings $\boldsymbol{\Gamma}$, identifiability is well studied in classical multivariate statistics. However, the second step of estimating the effects $\boldsymbol{\beta}$ is not identifiable without additional constraints. We consider two different sufficient conditions for global identifiability. The first condition assumes the researcher has a "negative control" variable set for which there should be no direct effect. This negative control set often serves as a quality control precaution in microarray studies [26], but they can also be used to adjust for the confounding factors. The second identification condition assumes at least half of the true effects are zero, that is, the true alternative hypotheses are sparse. These two identification conditions correspond to the approaches of RUV-4 [25] and LEAPP [59], respectively.

Our second contribution in Section 3 is to derive valid and efficient statistical methods under these identification conditions in the second step. In order to estimate the effects, it is essential to estimate the coefficients $\boldsymbol{\alpha}$ relating the primary variable to the confounders. Under the two different identification conditions, we study two different regression methods which are analytically tractable and equally well performing alternatives to RUV-4 and LEAPP. For the negative control (NC) scenario, $\hat{\boldsymbol{\alpha}}^{\mathrm{NC}}$ and $\hat{\boldsymbol{\beta}}^{\mathrm{NC}}$ are obtained by generalized least squares using the negative controls. For the sparsity scenario, $\hat{\boldsymbol{\alpha}}^{\mathrm{RR}}$ and $\hat{\boldsymbol{\beta}}^{\mathrm{RR}}$ are obtained by using a simpler and more analytically tractable robust regression (RR) than the one used in LEAPP.

When the factors are strong (as large as the noise magnitude), for both scenarios we find that the resulting estimators of $\boldsymbol{\beta}$ are asymptotically as efficient as the oracle estimator which is allowed to observe the confounding factors. It is surprising that no essential loss of efficiency is incurred by searching for the confounding variables. Our asymptotic analysis relies on some recent theoretical results for factor analysis due to [3]. The asymptotic regime we consider has both n, the number of observations, and p, the number of outcome variables (e.g., genes), going to infinity. The most important condition that we require for asymptotic efficiency in the negative control scenario is that the number of negative controls increases to infinity; in the sparsity scenario, we need the L_1 norm of the effects to satisfy $\|\boldsymbol{\beta}\|_1\sqrt{n}/p \to 0$. The fact that $p \gg n$ in many multiple hypothesis testing problems plays an important role in these asymptotics.

Next, in Section 3, we show that the asymptotic z-statistics based on the efficient estimators of $\boldsymbol{\beta}$ can control the type I error. This is not a trivial corollary from the asymptotic distribution of the test statistics because the size of $\boldsymbol{\beta}$ is growing and the z-statistics are weakly correlated. Proving FDR control is more technically demanding and is beyond the scope of this paper. Instead, we use numerical simulations to study the empirical performance (including FDR) of our tests. We also give a significance test of confounding (null hypothesis $\boldsymbol{\alpha} = \mathbf{0}$) in Section 3. This test can help the experimenter to determine if there is any hidden confounder in the design or the experiment process.

In Section 4, we generalize the confounder adjustment model to include multiple primary variables and multiple nuisance covariates. We show the statistical methods and theory for the single primary variable regression problem (1.1) can be smoothly extended to the multiple regression problem.

Outline. Section 2 introduces the model and describes the two identification conditions. Section 3 studies the statistical inference. Section 4 extends our framework to a linear model with multiple primary variables and multiple known controlling covariates. Section 5 discusses our theoretical analysis in the context of previous literature, including the existing procedures for debiasing the confounders and existing theoretical results of multiple hypothesis testing under dependence (but no confounding). Section 6 studies the empirical behavior of our estimators in simulations and real data examples. Technical proofs of the results are provided in the supplementary material [62].

To help the reader follow this paper and compare our methods and theory with existing approaches, Table 1 summarizes some related publications with more detailed discussion in Section 5.

Notation. Throughout the article, we use bold upper-case letters for matrices and lower-case letters for vectors. We use Latin letters for random variables and Greek letters for model parameters. Subscripts of matrices are used to indicate row(s) whenever possible. For example, if $\mathcal{C}$ is a set of indices, then $\boldsymbol{\Gamma}_{\mathcal{C}}$ is the

TABLE 1
Selected literature in multiple hypothesis testing under dependence. The categorization is partially subjective as some authors do not use exactly the same terminology

	Noise conditional on latent factors	
	Independent	**Correlated**
Positive or weak dependence	Benjamini and Yekutieli [9] Storey, Taylor and Siegmund [56] Clarke and Hall [14]	
Unconfounding factors	Friguet, Kloareg and Causeur [24] Desai and Storey [16]	Fan, Han and Gu [21] Lan and Du [35] *Discussed in Sections* 5.1 and 5.2
Confounding factors	Leek and Storey [38, 39] Gagnon-Bartsch and Speed [26] Sun, Zhang and Owen [59] *Studied in Sections* 2–4 *Discussed in Section* 5.3	*Discussed in Section* 5.4 (*future research*)

corresponding rows of $\mathbf{\Gamma}$. The L_0 norm of a vector is defined as the number of nonzero entries: $\|\boldsymbol{\beta}\|_0 = |\{1 \le j \le p : \beta_j \ne 0\}|$. A random matrix $\mathbf{E} \in \mathbb{R}^{n\times p}$ is said to follow a *matrix normal* distribution with mean $\mathbf{M} \in \mathbb{R}^{n\times p}$, row covariance $\mathbf{U} \in \mathbb{R}^{n\times n}$ and column covariance $\mathbf{V} \in \mathbb{R}^{p\times p}$, abbreviated as $\mathbf{E} \sim \mathrm{MN}(\mathbf{M}, \mathbf{U}, \mathbf{V})$, if the vectorization of $\mathbf{E}$ by column follows the multivariate normal distribution $\mathrm{vec}(\mathbf{E}) \sim \mathrm{N}(\mathrm{vec}(\mathbf{M}), \mathbf{V} \otimes \mathbf{U})$. When $\mathbf{U} = \mathbf{I}_n$, this means the rows of $\mathbf{E}$ are i.i.d. $\mathrm{N}(0, \mathbf{V})$. We use the usual notation in asymptotic statistics that a random variable is $O_p(1)$ if it is bounded in probability, and $o_p(1)$ if it converges to 0 in probability. Bold symbols $\mathbf{O}_p(1)$ or $\mathbf{o}_p(1)$ mean each entry of the vector is $O_p(1)$ or $o_p(1)$.

2. The model.

2.1. *Linear model with confounders.* We consider a single primary variable of interest in this section. It is common to add intercepts and known confounder effects (such as lab and batch effects) in the regression model. This extension to multiple linear regression does not change the main theoretical results in this paper and is discussed in Section 4.

For simplicity, all the variables in this section are assumed to have mean 0 marginally. Our model is built on equation (1.1) that is already widely used in the existing literature and we rewrite it here:

$$\mathbf{Y}_{n\times p} = \mathbf{X}_{n\times 1}\boldsymbol{\beta}_{p\times 1}^T + \mathbf{Z}_{n\times r}\mathbf{\Gamma}_{p\times r}^T + \mathbf{E}_{n\times p}. \tag{2.1a}$$

As mentioned earlier, it is also crucial to model the dependence of the confounders $\mathbf{Z}$ and the primary variable $\mathbf{X}$. We assume a linear relationship as in (1.2)

$$\mathbf{Z} = \mathbf{X}\boldsymbol{\alpha}^T + \mathbf{W}, \tag{2.1b}$$

and in addition some distributional assumptions on $\mathbf{X}$, $\mathbf{W}$ and the noise matrix $\mathbf{E}$

$$X_i \overset{\text{i.i.d.}}{\sim} \text{mean } 0, \qquad \text{variance } 1, \qquad i = 1, \ldots, n, \tag{2.1c}$$

$$\mathbf{W} \sim \mathrm{MN}(\mathbf{0}, \mathbf{I}_n, \mathbf{I}_r), \qquad \mathbf{W} \perp\!\!\!\perp \mathbf{X}, \tag{2.1d}$$

$$\mathbf{E} \sim \mathrm{MN}(\mathbf{0}, \mathbf{I}_n, \mathbf{\Sigma}), \qquad \mathbf{E} \perp\!\!\!\perp (\mathbf{X}, \mathbf{Z}). \tag{2.1e}$$

The parameters in the model (2.1a)–(2.1e) are $\boldsymbol{\beta} \in \mathbb{R}^{p\times 1}$ the primary effects we are most interested in, $\mathbf{\Gamma} \in \mathbb{R}^{p\times r}$ the influence of confounding factors on the outcomes, $\boldsymbol{\alpha} \in \mathbb{R}^{r\times 1}$ the association of the primary variable with the confounding factors, and $\mathbf{\Sigma} \in \mathbb{R}^{p\times p}$ the noise covariance matrix. We assume $\mathbf{\Sigma}$ is diagonal $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$, so the noise for different outcome variables is independent. We discuss possible ways to relax this independence assumption in Section 5.4.

In (2.1c), X_i is not required to be Gaussian or even continuous. For example, a binary or categorical variable after normalization also meets this assumption. As mentioned in Section 1, the parameter vector $\boldsymbol{\alpha}$ measures how severely the data are confounded. For a more intuitive interpretation, consider an oracle procedure of estimating $\boldsymbol{\beta}$ when the confounders $\mathbf{Z}$ in (2.1a) are observed. The best linear unbiased estimator in this case is the ordinary least squares $(\hat{\beta}_j^{\mathrm{OLS}}, \hat{\mathbf{\Gamma}}_j^{\mathrm{OLS}})$, whose variance is $\sigma_j^2 \operatorname{Var}(X_i, \mathbf{Z}_i)^{-1}/n$. Using (2.1b) and (2.1d), it is easy to show that $\operatorname{Var}(\hat{\beta}_j^{\mathrm{OLS}}) = (1 + \|\boldsymbol{\alpha}\|_2^2)\sigma_j^2/n$ and $\operatorname{Cov}(\hat{\beta}_j{}^{\mathrm{OLS}}, \hat{\beta}_k{}^{\mathrm{OLS}}) = 0$ for $j \neq k$. In summary,

$$\operatorname{Var}\big(\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}\big) = \frac{1}{n}\big(1 + \|\boldsymbol{\alpha}\|_2^2\big)\mathbf{\Sigma}. \tag{2.2}$$

Notice that in the unconfounded linear model in which $\mathbf{Z} = \mathbf{0}$, the variance of the OLS estimator of $\boldsymbol{\beta}$ is $\mathbf{\Sigma}/n$. Therefore, $1 + \|\boldsymbol{\alpha}\|_2^2$ represents the relative loss of efficiency when we add observed variables $\mathbf{Z}$ to the regression which are correlated with $\mathbf{X}$. In Section 3.2, we show that the oracle efficiency (2.2) can be asymptotically achieved even when $\mathbf{Z}$ is unobserved.

Let $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{\Gamma}, \mathbf{\Sigma})$ be all the parameters and $\mathbf{\Theta}$ be the parameter space. Without any constraint, the model (2.1a)–(2.1e) is unidentifiable. In Sections 2.3 and 2.4, we show how to restrict $\mathbf{\Theta}$ to ensure identifiability.

2.2. *Rotation*. Following [59], we introduce a transformation of the data to make the identification issues clearer. Consider the Householder rotation matrix $\mathbf{Q}^T \in \mathbb{R}^{n\times n}$ such that $\mathbf{Q}^T\mathbf{X} = \|\mathbf{X}\|_2\mathbf{e}_1 = (\|\mathbf{X}\|_2, 0, 0, \ldots, 0)^T$. Left-multiplying $\mathbf{Y}$ by $\mathbf{Q}^T$, we get $\tilde{\mathbf{Y}} = \mathbf{Q}^T\mathbf{Y} = \|\mathbf{X}\|_2\mathbf{e}_1\boldsymbol{\beta}^T + \tilde{\mathbf{Z}}\mathbf{\Gamma}^T + \tilde{\mathbf{E}}$, where

$$\tilde{\mathbf{Z}} = \mathbf{Q}^T\mathbf{Z} = \mathbf{Q}^T\big(\mathbf{X}\boldsymbol{\alpha}^T + \mathbf{W}\big) = \|\mathbf{X}\|_2\mathbf{e}_1\boldsymbol{\alpha}^T + \tilde{\mathbf{W}}, \tag{2.3}$$

and $\tilde{\mathbf{W}} = \mathbf{Q}^T\mathbf{W} \overset{d}{=} \mathbf{W}$, $\tilde{\mathbf{E}} = \mathbf{Q}^T\mathbf{E} \overset{d}{=} \mathbf{E}$. As a consequence, the first and the rest of the rows of $\tilde{\mathbf{Y}}$ are

$$\tilde{\mathbf{Y}}_1 = \|\mathbf{X}\|_2\boldsymbol{\beta}^T + \tilde{\mathbf{Z}}_1\mathbf{\Gamma}^T + \tilde{\mathbf{E}}_1 \sim \mathrm{N}\big(\|\mathbf{X}\|_2(\boldsymbol{\beta} + \mathbf{\Gamma}\boldsymbol{\alpha})^T, \mathbf{\Gamma}\mathbf{\Gamma}^T + \mathbf{\Sigma}\big), \tag{2.4}$$

$$\tilde{\mathbf{Y}}_{-1} = \tilde{\mathbf{Z}}_{-1}\mathbf{\Gamma}^T + \tilde{\mathbf{E}}_{-1} \sim \mathrm{MN}\big(\mathbf{0}, \mathbf{I}_{n-1}, \mathbf{\Gamma}\mathbf{\Gamma}^T + \mathbf{\Sigma}\big). \tag{2.5}$$

Here, $\tilde{\mathbf{Y}}_1$ is a $1 \times p$ vector, $\tilde{\mathbf{Y}}_{-1}$ is a $(n-1) \times p$ matrix, and the distributions are conditional on $\mathbf{X}$.

The parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ only appear in (2.4), so their inference (step 1 in our procedure) can be completely separated from the inference of $\boldsymbol{\Gamma}$ and $\boldsymbol{\Sigma}$ (step 2 in our procedure). In fact, $\tilde{\mathbf{Y}}_1 \perp\!\!\!\perp \tilde{\mathbf{Y}}_{-1} | \mathbf{X}$ because $\tilde{\mathbf{E}}_1 \perp\!\!\!\perp \tilde{\mathbf{E}}_{-1}$, so the two steps use mutually independent information. This in turn greatly simplifies the theoretical analysis.

We intentionally use the symbol $\mathbf{Q}$ to resemble the QR decomposition of $\mathbf{X}$. In Section 4, we show how to use the QR decomposition to separate the primary effects from confounder and nuisance effects when $\mathbf{X}$ has multiple columns. Using the same notation, we discuss how SVA and RUV decouple the problem in a slightly different manner in Section 5.3.1.

2.3. *Identifiability of* $\boldsymbol{\Gamma}$. Equation (2.5) is just the exploratory factor analysis model, thus $\boldsymbol{\Gamma}$ can be easily identified up to some rotation under some mild conditions. Here, we assume a classical sufficient condition for the identification of $\boldsymbol{\Gamma}$ ([2], Theorem 5.1).

LEMMA 2.1. *Let* $\boldsymbol{\Theta} = \boldsymbol{\Theta}_0$ *be the parameter space such that*:

1. *If any row of* $\boldsymbol{\Gamma}$ *is deleted, there remain two disjoint submatrices of* $\boldsymbol{\Gamma}$ *of rank* r, *and*
2. $\boldsymbol{\Gamma}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\Gamma}/p$ *is diagonal and the diagonal elements are distinct, positive and arranged in decreasing order.*

Then $\boldsymbol{\Gamma}$ *and* $\boldsymbol{\Sigma}$ *are identifiable in the model* (2.1a)–(2.1e).

In Lemma 2.1, condition (1) requires that $p \geq 2r + 1$. Condition (1) identifies $\boldsymbol{\Gamma}$ up to a rotation which is sufficient to identify $\boldsymbol{\beta}$. To see this, we can reparameterize $\boldsymbol{\Gamma}$ and $\boldsymbol{\alpha}$ to $\boldsymbol{\Gamma}\mathbf{U}$ and $\mathbf{U}^T\boldsymbol{\alpha}$ using an $r \times r$ orthogonal matrix $\mathbf{U}$. This reparameterization does not change the distribution of $\tilde{\mathbf{Y}}_1$ in (2.4) if $\boldsymbol{\beta}$ remains the same. Condition (2) identifies the rotation uniquely but is not necessary for our theoretical analysis in later sections.

2.4. *Identifiability of* $\boldsymbol{\beta}$. The parameters $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ cannot be identified from (2.4) because they have in total $p + r$ parameters while $\tilde{\mathbf{Y}}_1$ is a length p vector. If we write $\mathcal{P}_{\boldsymbol{\Gamma}}$ and $\mathcal{P}_{\boldsymbol{\Gamma}^\perp}$ as the projection onto the column space and orthogonal space of $\boldsymbol{\Gamma}$ so that $\boldsymbol{\beta} = \mathcal{P}_{\boldsymbol{\Gamma}}\boldsymbol{\beta} + \mathcal{P}_{\boldsymbol{\Gamma}^\perp}\boldsymbol{\beta}$, it is impossible to identify $\mathcal{P}_{\boldsymbol{\Gamma}}\boldsymbol{\beta}$ from (2.4).

This suggests that we should further restrict the parameter space $\boldsymbol{\Theta}$. We will reduce the degrees of freedom by restricting at least r entries of $\boldsymbol{\beta}$ to equal 0. We consider two different sufficient conditions to identify $\boldsymbol{\beta}$:

Negative control $\boldsymbol{\Theta}_1 = \{(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}) : \boldsymbol{\beta}_{\mathcal{C}} = \mathbf{0}, \operatorname{rank}(\boldsymbol{\Gamma}_{\mathcal{C}}) = r\}$ for a known negative control set $|\mathcal{C}| \geq r$.

Sparsity $\boldsymbol{\Theta}_2(s) = \{(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}) : \|\boldsymbol{\beta}\|_0 \le \lfloor (p-s)/2 \rfloor, \operatorname{rank}(\boldsymbol{\Gamma}_{\mathcal{C}}) = r, \forall \mathcal{C} \subset \{1, \ldots, p\}, |\mathcal{C}| = s\}$ for some $r \le s \le p$.

PROPOSITION 2.1. *If* $\boldsymbol{\Theta} = \boldsymbol{\Theta}_0 \cap \boldsymbol{\Theta}_1$ *or* $\boldsymbol{\Theta} = \boldsymbol{\Theta}_0 \cap \boldsymbol{\Theta}_2(s)$ *for some* $r \le s \le p$, *the parameters* $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma})$ *in the model* (2.1a)–(2.1e) *are identifiable*.

PROOF. Since $\boldsymbol{\Theta} \subset \boldsymbol{\Theta}_0$, we know from Lemma 2.1 that $\boldsymbol{\Gamma}$ and $\boldsymbol{\Sigma}$ are identifiable. Now consider two combinations of parameters $\boldsymbol{\theta}^{(1)} = (\boldsymbol{\alpha}^{(1)}, \boldsymbol{\beta}^{(1)}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma})$ and $\boldsymbol{\theta}^{(2)} = (\boldsymbol{\alpha}^{(2)}, \boldsymbol{\beta}^{(2)}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma})$ both in the space $\boldsymbol{\Theta}$ and inducing the same distribution in the model (2.1a)–(2.1e), that is, $\boldsymbol{\beta}^{(1)} + \boldsymbol{\Gamma}\boldsymbol{\alpha}^{(1)} = \boldsymbol{\beta}^{(2)} + \boldsymbol{\Gamma}\boldsymbol{\alpha}^{(2)}$.

Let $\mathcal{C}$ be the set of indices such that $\boldsymbol{\beta}^{(1)}_{\mathcal{C}} = \boldsymbol{\beta}^{(2)}_{\mathcal{C}} = \mathbf{0}$. If $\boldsymbol{\Theta} = \boldsymbol{\Theta}_0 \cap \boldsymbol{\Theta}_1$, we already know $|\mathcal{C}| \ge r$. If $\boldsymbol{\Theta} = \boldsymbol{\Theta}_0 \cap \boldsymbol{\Theta}_2(s)$, it is easy to show that $|\mathcal{C}| \ge s$ is also true because both $\boldsymbol{\beta}^{(1)}$ and $\boldsymbol{\beta}^{(2)}$ have at most $\lfloor (p-s)/2 \rfloor$ nonzero entries. Along with the rank constraint on $\boldsymbol{\Gamma}_{\mathcal{C}}$, this implies that $\boldsymbol{\Gamma}_{\mathcal{C}}\boldsymbol{\alpha}^{(1)} = \boldsymbol{\Gamma}_{\mathcal{C}}\boldsymbol{\alpha}^{(2)}$. However, the conditions in $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$ ensure that $\boldsymbol{\Gamma}_{\mathcal{C}}$ has full rank, so $\boldsymbol{\alpha}^{(1)} = \boldsymbol{\alpha}^{(2)}$, and hence $\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^{(2)}$. □

REMARK 2.1. The condition (2) in Lemma 2.1 that uniquely identifies $\boldsymbol{\Gamma}$ is not necessary for the identification of $\boldsymbol{\beta}$. This is because for any set $|\mathrm{C}| \ge r$ and any orthogonal matrix $\mathbf{U} \in \mathbb{R}^{r \times r}$, we always have $\operatorname{rank}(\boldsymbol{\Gamma}_{\mathcal{C}}) = \operatorname{rank}(\boldsymbol{\Gamma}_{\mathcal{C}})\mathbf{U}$. Therefore, $\boldsymbol{\Gamma}$ only needs to be identified up to a rotation.

REMARK 2.2. Almost all dense matrices of $\boldsymbol{\Gamma} \in \mathbb{R}^{p \times r}$ satisfy the conditions. However, for $\boldsymbol{\Theta}_2(s)$ the sparsity of $\boldsymbol{\Gamma}$ allowed depends on the sparsity of $\boldsymbol{\beta}$. The condition $\boldsymbol{\Theta}_2(s)$ rules out some too sparse $\boldsymbol{\Gamma}$. In this case, one may consider using confirmatory factor analysis instead of exploratory factor analysis to model the relationship between confounders and outcomes. For some recent identification results in confirmatory factor analysis, see [29, 34].

REMARK 2.3. The maximum allowed $\|\boldsymbol{\beta}\|_0$ in $\boldsymbol{\Theta}_2$, $\lfloor (p-r)/2 \rfloor$, is exactly the maximum breakdown point of a robust regression with p observations and r predictors [42]. Indeed, we use a standard robust regression method to estimate $\boldsymbol{\beta}$ in this case in Section 3.2.2.

REMARK 2.4. To the best of our knowledge, the only existing literature that explicitly addresses the identifiability issue is [58], Chapter 4.2, where the author gives sufficient conditions for *local* identifiability of $\boldsymbol{\beta}$ by viewing (2.1a) as a "sparse plus low rank" matrix decomposition problem. See [13], Section 3.3, for a more general discussion of the local and global identifiability for this problem. Local identifiability refers to identifiability of the parameters in a neighborhood of the true values. In contrast, the conditions in Proposition 2.1 ensure that $\boldsymbol{\beta}$ is *globally* identifiable in the restricted parameter space.

3. Statistical inference. As mentioned earlier in Section 1, the statistical inference consists of two steps: the factor analysis (Section 3.1) and the linear regression (Section 3.2).

3.1. *Inference for* $\mathbf{\Gamma}$ *and* $\mathbf{\Sigma}$. The most popular approaches for factor analysis are principal component analysis (PCA) and maximum likelihood (ML). Bai and Ng [6] derived a class of estimators of r by principal component analysis using various information criteria. The estimators are consistent under Assumption 3 in this section and some additional technical assumptions in [6]. Due to this reason, we assume the number of confounding factors r is known in this section. See [45], Section 3, for a comprehensive literature review of choosing r in practice.

We are most interested in the asymptotic behavior of factor analysis when both $n, p \to \infty$. In this case, PCA cannot consistently estimate the noise variance $\mathbf{\Sigma}$ [3]. For theoretical analysis, we use the quasi maximum likelihood estimate in [3] to get $\hat{\mathbf{\Gamma}}$ and $\hat{\mathbf{\Sigma}}$. This estimator is called "quasi"-MLE because it treats the factors $\tilde{\mathbf{Z}}_{-1}$ as fixed quantities. Since the confounders $\mathbf{Z}$ in our model (2.1a)–(2.1e) are random variables, we introduce a rotation matrix $\mathbf{R} \in \mathbb{R}^{r\times r}$ and let $\tilde{\mathbf{Z}}_{-1}^{(0)} = \tilde{\mathbf{Z}}_{-1}(\mathbf{R}^{-1})^T$, $\mathbf{\Gamma}^{(0)} = \mathbf{\Gamma}\mathbf{R}$ be the target factors and factor loadings that are studied in [3].

To make $\tilde{\mathbf{Z}}_{-1}^{(0)}$ and $\mathbf{\Gamma}^{(0)}$ identifiable, [3] consider five different identification conditions. However, the parameter of interest in model (2.1a)–(2.1e) is $\boldsymbol{\beta}$ instead of $\mathbf{\Gamma}$ or $\mathbf{\Gamma}^{(0)}$. As we have discussed in Section 2.4, we only need the column space of $\mathbf{\Gamma}$ to estimate $\boldsymbol{\beta}$, which gives us some flexibility of choosing the identification condition. In our theoretical analysis, we use the third condition (IC3) in [3], which imposes the constraints that $(n-1)^{-1}(\tilde{\mathbf{Z}}_{-1}^{(0)})^T\tilde{\mathbf{Z}}_{-1}^{(0)} = \mathbf{I}_r$ and $p^{-1}\tilde{\mathbf{\Gamma}}^{(0)T}\mathbf{\Sigma}^{-1}\mathbf{\Gamma}^{(0)}$ is diagonal. Therefore, the rotation matrix $\mathbf{R}$ satisfies $\mathbf{R}\mathbf{R}^T = (n-1)^{-1}\tilde{\mathbf{Z}}_{-1}^T\tilde{\mathbf{Z}}_{-1}$.

The quasi-log-likelihood being maximized in [3] is

$$-\frac{1}{2p}\log\det\bigl(\mathbf{\Gamma}^{(0)}(\mathbf{\Gamma}^{(0)})^T + \mathbf{\Sigma}\bigr) - \frac{1}{2p}\operatorname{tr}\bigl\{\mathbf{S}\bigl[\mathbf{\Gamma}^{(0)}(\mathbf{\Gamma}^{(0)})^T + \mathbf{\Sigma}\bigr]^{-1}\bigr\}, \tag{3.1}$$

where $\mathbf{S}$ is the sample covariance matrix of $\tilde{\mathbf{Y}}_{-1}$.

The theoretical results in this section rely heavily on recent findings in [3]. They use these three assumptions.

ASSUMPTION 1. The noise matrix $\mathbf{E}$ follows the matrix normal distribution $\mathbf{E} \sim \mathrm{MN}(\mathbf{0}, \mathbf{I}_n, \mathbf{\Sigma})$ and $\mathbf{\Sigma}$ is a diagonal matrix.

ASSUMPTION 2. There exists a positive constant D such that $\|\mathbf{\Gamma}_j\|_2 \le D$, $D^{-2} \le \sigma_j^2 \le D^2$ for all j, and the estimated variances $\hat{\sigma}_j^2 \in [D^{-2}, D^2]$ for all j.

ASSUMPTION 3. The limits $\lim_{p\to\infty} p^{-1}\mathbf{\Gamma}^T\mathbf{\Sigma}^{-1}\mathbf{\Gamma}$ and $\lim_{p\to\infty}\sum_{j=1}^p \sigma_j^{-4} \times (\mathbf{\Gamma}_j \otimes \mathbf{\Gamma}_j)(\mathbf{\Gamma}_j^T \otimes \mathbf{\Gamma}_j^T)$ exist and are positive definite matrices.

LEMMA 3.1 (Bai and Li [3]). *Under Assumptions 1–3, the maximizer* $(\hat{\mathbf{\Gamma}}, \hat{\mathbf{\Sigma}})$ *of the quasi-log-likelihood* (3.1) *satisfies*

$$\sqrt{n}(\hat{\mathbf{\Gamma}}_j - \mathbf{\Gamma}_j^{(0)}) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \sigma_j^2 \mathbf{I}_r) \quad \textit{and} \quad \sqrt{n}(\hat{\sigma}_j^2 - \sigma_j^2) \xrightarrow{d} \mathrm{N}(0, 2\sigma_j^4).$$

In the supplementary material [62], we prove some strengthened technical results of Lemma 3.1 that are used in the proof of subsequent theorems.

REMARK 3.1. Assumption 2 is Assumption D from [3]. It requires that the diagonal elements of the quasi-MLE $\hat{\mathbf{\Sigma}}$ be uniformly bounded away from zero and infinity. We would prefer boundedness to be a consequence of some assumptions on the distribution of the data, but at present we are unaware of any other results like Lemma 3.1 which do not use this assumption. In practice, the quasi-likelihood problem (3.1) is commonly solved by the Expectation–Maximization (EM) algorithm. Similar to [3, 4], we do not find it necessary to impose an upper or lower bound for the parameters in the EM algorithm in the numerical experiments.

3.2. *Inference for* $\boldsymbol{\alpha}$ *and* $\boldsymbol{\beta}$. The estimation of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is based on the first row of the rotated outcome $\tilde{\mathbf{Y}}_1$ in (2.4), which can be rewritten as

$$\tilde{\mathbf{Y}}_1^T / \|\mathbf{X}\|_2 = \boldsymbol{\beta} + \mathbf{\Gamma}(\boldsymbol{\alpha} + \tilde{\mathbf{W}}_1 / \|\mathbf{X}\|_2) + \tilde{\mathbf{E}}_1^T / \|\mathbf{X}\|_2, \tag{3.2}$$

where $\tilde{\mathbf{W}}_1 \sim \mathrm{N}(0, \mathbf{I}_p)$ is from (2.3) and $\tilde{\mathbf{W}}_1$ is independent of $\tilde{\mathbf{E}}_1 \sim \mathrm{N}(0, \mathbf{\Sigma})$. Note that $\tilde{\mathbf{Y}}_1 / \|\mathbf{X}\|_2$ is proportional to the sample covariance between $\mathbf{Y}$ and $\mathbf{X}$. All the methods described in this section first try to find a good estimator $\hat{\boldsymbol{\alpha}}$. They then use $\hat{\boldsymbol{\beta}} = \tilde{\mathbf{Y}}_1^T / \|\mathbf{X}\|_2 - \hat{\mathbf{\Gamma}}\hat{\boldsymbol{\alpha}}$ to estimate $\boldsymbol{\beta}$.

To reduce variance, we choose to estimate (3.2) conditional on $\tilde{\mathbf{W}}_1$. Also, to use the results in Lemma 3.1, we replace $\mathbf{\Gamma}$ by $\mathbf{\Gamma}^{(0)}$. Then we can rewrite (3.2) as

$$\tilde{\mathbf{Y}}_1^T / \|\mathbf{X}\|_2 = \boldsymbol{\beta} + \mathbf{\Gamma}^{(0)} \boldsymbol{\alpha}^{(0)} + \tilde{\mathbf{E}}_1^T / \|\mathbf{X}\|_2, \tag{3.3}$$

where $\mathbf{\Gamma}^{(0)} = \mathbf{\Gamma}\mathbf{R}$ and $\boldsymbol{\alpha}^{(0)} = \mathbf{R}^{-1}(\boldsymbol{\alpha} + \tilde{\mathbf{W}}_1 / \|\mathbf{X}\|_2)$. Notice that the random $\mathbf{R}$ only depends on $\tilde{\mathbf{Y}}_{-1,}$ and thus is independent of $\tilde{\mathbf{Y}}_1$. In the proof of the results in this section, we first consider the estimation of $\boldsymbol{\beta}$ for fixed $\tilde{\mathbf{W}}_1$, $\mathbf{R}$ and $\mathbf{X}$, and then show the asymptotic distribution of $\hat{\boldsymbol{\beta}}$ indeed does not depend on $\tilde{\mathbf{W}}_1$, $\mathbf{R}$ or $\mathbf{X}$, and thus also holds unconditionally.

3.2.1. *Negative control scenario*. If we know a set $\mathcal{C}$ such that $\boldsymbol{\beta}_{\mathcal{C}} = 0$ (so $\mathbf{\Theta} \subset \mathbf{\Theta}_1$), then $\tilde{\mathbf{Y}}_1$ can be correspondingly separated into two parts:

$$\begin{aligned} \tilde{\mathbf{Y}}_{1,\mathcal{C}}^T / \|\mathbf{X}\|_2 &= \mathbf{\Gamma}_{\mathcal{C}}^{(0)} \boldsymbol{\alpha}^{(0)} + \tilde{\mathbf{E}}_{1,\mathcal{C}}^T / \|\mathbf{X}\|_2, \quad \text{and} \\ \tilde{\mathbf{Y}}_{1,-\mathcal{C}}^T / \|\mathbf{X}\|_2 &= \boldsymbol{\beta}_{-\mathcal{C}} + \mathbf{\Gamma}_{-\mathcal{C}}^{(0)} \boldsymbol{\alpha}^{(0)} + \tilde{\mathbf{E}}_{1,-\mathcal{C}}^T / \|\mathbf{X}\|_2. \end{aligned} \tag{3.4}$$

The number of negative controls $|\mathcal{C}|$ may grow as $p \to \infty$. We impose an additional assumption on the latent factors of the negative controls.

ASSUMPTION 4. $\lim_{p\to\infty} |\mathcal{C}|^{-1}\mathbf{\Gamma}_{\mathcal{C}}^T\mathbf{\Sigma}_{\mathcal{C}}^{-1}\mathbf{\Gamma}_{\mathcal{C}}$ exists and is positive definite.

We consider the following negative control (NC) estimator where $\boldsymbol{\alpha}^{(0)}$ is estimated by generalized least squares:

$$\hat{\boldsymbol{\alpha}}^{\mathrm{NC}} = (\hat{\mathbf{\Gamma}}_{\mathcal{C}}^T\hat{\mathbf{\Sigma}}_{\mathcal{C}}^{-1}\hat{\mathbf{\Gamma}}_{\mathcal{C}})^{-1}\hat{\mathbf{\Gamma}}_{\mathcal{C}}^T\hat{\mathbf{\Sigma}}_{\mathcal{C}}^{-1}\tilde{\mathbf{Y}}_{1,\mathcal{C}}^T/\|\mathbf{X}\|_2 \quad \text{and} \tag{3.5}$$

$$\hat{\boldsymbol{\beta}}_{-\mathcal{C}}^{\mathrm{NC}} = \tilde{\mathbf{Y}}_{1,-\mathcal{C}}^T/\|\mathbf{X}\|_2 - \hat{\mathbf{\Gamma}}_{-\mathcal{C}}\hat{\boldsymbol{\alpha}}^{\mathrm{NC}}. \tag{3.6}$$

This estimator matches the RUV-4 estimator of [25] except that it uses quasi-maximum likelihood estimates of $\mathbf{\Sigma}$ and $\mathbf{\Gamma}$ instead of using PCA, and generalized linear squares instead of ordinary linear squares regression. The details are in Section 5.3.2.

Our goal is to show consistency and asymptotic variance of $\hat{\boldsymbol{\beta}}_{-\mathcal{C}}^{\mathrm{NC}}$. Let $\mathbf{\Sigma}_{\mathcal{C}}$ represents the noise covariance matrix of the variables in $\mathcal{C}$. We then have

THEOREM 3.1. *Under Assumptions 1–4, if $n, p \to \infty$ and $(\log p)^2/n \to 0$, then for any fixed index set $\mathcal{S}$ with finite cardinality and $\mathcal{S} \cap \mathcal{C} = \varnothing$, we have*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathcal{S}}^{\mathrm{NC}} - \boldsymbol{\beta}_{\mathcal{S}}) \xrightarrow{d} \mathrm{N}(\mathbf{0}, (1+\|\boldsymbol{\alpha}\|_2^2)(\mathbf{\Sigma}_{\mathcal{S}} + \mathbf{\Delta}_{\mathcal{S}})), \tag{3.7}$$

where $\mathbf{\Delta}_{\mathcal{S}} = \mathbf{\Gamma}_{\mathcal{S}}(\mathbf{\Gamma}_{\mathcal{C}}^T\mathbf{\Sigma}_{\mathcal{C}}^{-1}\mathbf{\Gamma}_{\mathcal{C}})^{-1}\mathbf{\Gamma}_{\mathcal{S}}^T$.

If in addition, $|\mathcal{C}| \to \infty$, the minimum eigenvalue of $\mathbf{\Gamma}_{\mathcal{C}}^T\mathbf{\Sigma}_{\mathcal{C}}^{-1}\mathbf{\Gamma}_{\mathcal{C}} \to \infty$ by Assumption 4, then the maximum entry of $\mathbf{\Delta}_{\mathcal{S}}$ goes to 0. Therefore, in this case,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathcal{S}}^{\mathrm{NC}} - \boldsymbol{\beta}_{\mathcal{S}}) \xrightarrow{d} \mathrm{N}(\mathbf{0}, (1+\|\boldsymbol{\alpha}\|_2^2)\mathbf{\Sigma}_{\mathcal{S}}). \tag{3.8}$$

The asymptotic variance in (3.8) is the same as the variance of the oracle least squares in (2.2). Comparable oracle efficiency statements can be found in the econometrics literature [7, 63]. This is also the variance used implicitly in RUV-4 as it treats the estimated $\mathbf{Z}$ as given when deriving test statistics for $\boldsymbol{\beta}$. When the number of negative controls is not too large, say $|\mathcal{C}| = 30$, the correction term $\mathbf{\Delta}_S$ is nontrivial and gives more accurate estimate of the variance of $\hat{\boldsymbol{\beta}}^{\mathrm{NC}}$. See Section 6.1 for more simulation results.

3.2.2. *Sparsity scenario.* When the zero indices in $\boldsymbol{\beta}$ are unknown but sparse (so $\mathbf{\Theta} \subseteq \mathbf{\Theta}_2$), the estimation of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ from $\tilde{\mathbf{Y}}_1^T/\|\mathbf{X}\|_2 = \boldsymbol{\beta} + \mathbf{\Gamma}^{(0)}\boldsymbol{\alpha}^{(0)} + \tilde{\mathbf{E}}_1^T/\|\mathbf{X}\|_2$ can be cast as a robust regression by viewing $\tilde{\mathbf{Y}}_1^T$ as observations and $\mathbf{\Gamma}^{(0)}$ as design matrix. The nonzero entries in $\boldsymbol{\beta}$ correspond to outliers in this linear regression.

The problem here has two nontrivial differences compared to classical robust regression. First, we expect some entries of $\boldsymbol{\beta}$ to be nonzero, and our goal is to make inference on the outliers; second, we do not observe the design matrix $\mathbf{\Gamma}^{(0)}$ but only have its estimator $\hat{\mathbf{\Gamma}}$. In fact, if $\boldsymbol{\beta} = \mathbf{0}$ and $\mathbf{\Gamma}^{(0)}$ is observed, the ordinary

least squares estimator of $\boldsymbol{\alpha}^{(0)}$ is unbiased and has variance of order $1/(np)$, because the noise in (3.2) has variance $1/n$ and there are p observations. Our main conclusion is that $\boldsymbol{\alpha}^{(0)}$ can still be estimated very accurately given the two technical difficulties.

Given a robust loss function ρ, we consider the following estimator:

$$\hat{\boldsymbol{\alpha}}^{\mathrm{RR}} = \arg\min \sum_{j=1}^{p} \rho\Big(\frac{\tilde{Y}_{1j}/\|\mathbf{X}\|_2 - \hat{\boldsymbol{\Gamma}}_j^T \boldsymbol{\alpha}}{\hat{\sigma}_j}\Big) \quad \text{and} \tag{3.9}$$

$$\hat{\boldsymbol{\beta}}^{\mathrm{RR}} = \tilde{\mathbf{Y}}_1/\|\mathbf{X}\|_2 - \hat{\boldsymbol{\Gamma}}\hat{\boldsymbol{\alpha}}^{\mathrm{RR}}. \tag{3.10}$$

For a broad class of loss functions ρ, estimating $\boldsymbol{\alpha}$ by (3.9) is equivalent to

$$(\hat{\boldsymbol{\alpha}}^{\mathrm{RR}}, \tilde{\boldsymbol{\beta}}) = \arg\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \sum_{j=1}^{p} \frac{1}{\hat{\sigma}_j^2}(\tilde{Y}_{1j}/\|\mathbf{X}\|_2 - \beta_j - \hat{\boldsymbol{\Gamma}}_j^T \boldsymbol{\alpha})^2 + P_\lambda(\boldsymbol{\beta}), \tag{3.11}$$

where $P_\lambda(\boldsymbol{\beta})$ is a penalty to promote sparsity of $\boldsymbol{\beta}$ [54]. However, $\hat{\boldsymbol{\beta}}^{\mathrm{RR}}$ is not identical to $\tilde{\boldsymbol{\beta}}$, which is a sparse vector that does not have an asymptotic normal distribution. The LEAPP algorithm [59] uses the form (3.11). Replacing it by the robust regression (3.9) and (3.10) allows us to derive significance tests of H_{0j} : $\beta_j = 0$.

We assume a smooth loss ρ for the theoretical analysis.

ASSUMPTION 5. The penalty $\rho : \mathbb{R} \to [0, \infty)$ with $\rho(0) = 0$. The function $\rho(x)$ is nonincreasing when $x \le 0$ and is nondecreasing when $x > 0$. The derivative $\psi = \rho'$ exists and $|\psi| \le D$ for some $D < \infty$. Furthermore, ρ is strongly convex in a neighborhood of 0.

A sufficient condition for the local strong convexity is that $\psi' > 0$ exists in a neighborhood of 0. The next theorem establishes the consistency of $\hat{\boldsymbol{\beta}}^{\mathrm{RR}}$.

THEOREM 3.2. *Under Assumptions 1–3 and 5, if* $n, p \to \infty$, $(\log p)^2/n \to 0$ *and* $\|\boldsymbol{\beta}\|_1/p \to 0$, *then* $\hat{\boldsymbol{\alpha}}^{\mathrm{RR}} \xrightarrow{p} \boldsymbol{\alpha}$. *As a consequence, for any* j, $\hat{\beta}_j^{\mathrm{RR}} \xrightarrow{p} \beta_j$.

To derive the asymptotic distribution, we consider the estimating equation corresponding to (3.9). By taking the derivative of (3.9), $\hat{\boldsymbol{\alpha}}^{\mathrm{RR}}$ satisfies

$$\boldsymbol{\Psi}_{p,\hat{\boldsymbol{\Gamma}},\hat{\boldsymbol{\Sigma}}}(\hat{\boldsymbol{\alpha}}^{\mathrm{RR}}) = \frac{1}{p}\sum_{j=1}^{p} \psi\Big(\frac{\tilde{Y}_{1j}/\|\mathbf{X}\|_2 - \hat{\boldsymbol{\Gamma}}_j^T \hat{\boldsymbol{\alpha}}^{\mathrm{RR}}}{\hat{\sigma}_j}\Big)\hat{\boldsymbol{\Gamma}}_j/\hat{\sigma}_j = \mathbf{0}. \tag{3.12}$$

The next assumption is used to control the higher order term in a Taylor expansion of $\boldsymbol{\Psi}$.

ASSUMPTION 6. The first two derivatives of ψ exist and both $|\psi'(x)| \le D$ and $|\psi''(x)| \le D$ hold at all x for some $D < \infty$.

Examples of loss functions ρ that satisfy Assumptions 5 and 6 include smoothed Huber loss and Tukey's bisquare.

The next theorem gives the asymptotic distribution of $\hat{\boldsymbol{\beta}}^{\mathrm{RR}}$ when the nonzero entries of $\boldsymbol{\beta}$ are sparse enough. The asymptotic variance of $\hat{\boldsymbol{\beta}}^{\mathrm{RR}}$ is, again, the oracle variance in (2.2).

THEOREM 3.3. *Under Assumptions 1–3, 5 and 6, if* $n, p \to \infty$, *with* $(\log p)^2/n \to 0$ *and* $\|\boldsymbol{\beta}\|_1\sqrt{n}/p \to 0$, *then*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathcal{S}}^{\mathrm{RR}} - \boldsymbol{\beta}_{\mathcal{S}}) \xrightarrow{d} \mathrm{N}\big(\mathbf{0}, (1 + \|\boldsymbol{\alpha}\|_2^2)\boldsymbol{\Sigma}_{\mathcal{S}}\big)$$

for any fixed index set S *with finite cardinality.*

If $n/p \to 0$, then a sufficient condition for $\|\boldsymbol{\beta}\|_1\sqrt{n}/p \to 0$ in Theorem 3.3 is $\|\boldsymbol{\beta}\|_1 = O(\sqrt{p})$. If instead $n/p \to c \in (0, \infty)$, then $\|\boldsymbol{\beta}\|_1 = o(\sqrt{p})$ suffices.

3.3. *Hypothesis testing.* In this section, we construct significance tests for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ based on the asymptotic normal distributions in the previous section.

3.3.1. *Test of the primary effects.* We consider the asymptotic test for H_{0j} : $\beta_j = 0$, $j = 1, \ldots, p$ resulting from the asymptotic distributions of $\hat{\beta}_j$ derived in Theorems 3.1 and 3.3:

$$t_j = \frac{\|\mathbf{X}\|_2\hat{\beta}_j}{\hat{\sigma}_j\sqrt{1 + \|\hat{\boldsymbol{\alpha}}\|^2}}, \qquad j = 1, \ldots, p. \tag{3.13}$$

Here, we require $|\mathcal{C}| \to \infty$ for the NC estimator. The null hypothesis H_{0j} is rejected at level-α if $|t_j| > z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$ as usual, where Φ is the cumulative distribution function of the standard normal. Note that here we slightly abuse the notation α to represent the significance level and this should not be confused with the model parameter $\boldsymbol{\alpha}$.

The next theorem shows that the overall type-I error and the family-wise error rate (FWER) can be asymptotically controlled by using the test statistics t_j, $j = 1, \ldots, p$.

THEOREM 3.4. *Let* $\mathcal{N}_p = \{j | \beta_j = 0, j = 1, \ldots, p\}$ *be all the true null hypotheses. Under the assumptions of Theorem 3.1 or Theorem 3.3,* $|\mathcal{C}| \to \infty$ *for the NC scenario, as* $n, p, |\mathcal{N}_p| \to \infty$,

$$\frac{1}{|\mathcal{N}_p|}\sum_{j \in \mathcal{N}_p} I\big(|t_j| > z_{\alpha/2}\big) \xrightarrow{p} \alpha \quad \textit{and} \tag{3.14}$$

$$\limsup \mathrm{P}\Big(\sum_{j \in \mathcal{N}_p} I\big(|t_j| > z_{\alpha/(2p)}\big) \geq 1 \Big) \leq \alpha. \tag{3.15}$$

Although the individual test is asymptotically valid as $t_j \xrightarrow{d} \mathrm{N}(0,1)$, Theorem 3.4 is not a trivial corollary of the asymptotic normal distribution in Theorems 3.1 and 3.3. This is because t_j, $j = 1, \ldots, p$ are not independent for finite samples. The proof of Theorem 3.4 investigates how the dependence of the test statistics diminishes when $n, p \to \infty$. The proof of Theorem 3.4 already requires a careful investigation of the convergence of $\hat{\boldsymbol{\beta}}$ in Theorem 3.3. It is more cumbersome to prove FDR control using our test statistics. In Section 6, we show that FDR is usually well controlled in simulations for the Benjamini–Hochberg procedure when the sample size is large enough.

REMARK 3.2. We find a calibration technique in [59] very useful to improve the type I error and FDR control for finite sample size. Because the asymptotic variance used in (3.13) is the variance of an oracle OLS estimator, when the sample size is not sufficiently large, the variance of $\hat{\beta}^{\mathrm{RR}}$ should be slightly larger than this oracle variance. To correct for this inflation, one can use median absolute deviation (MAD) with customary scaling to match the standard deviation for a Gaussian distribution to estimate the empirical standard error of t_j, $j = 1, \ldots, p$ and divide t_j by the estimated standard error. The performance of this empirical calibration is studied in the simulations in Section 6.1.

3.3.2. *Test of confounding*. We also consider a significance test for $H_{0,\boldsymbol{\alpha}}$: $\boldsymbol{\alpha} = \mathbf{0}$, under which the latent factors are not confounding.

THEOREM 3.5. *Let the assumptions of Theorem 3.1 or Theorem 3.3 and* $|\mathcal{C}| \to \infty$ *for the NC scenario be given. Under the null hypothesis that* $\boldsymbol{\alpha} = \mathbf{0}$, *for* $\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\alpha}}^{\mathrm{NC}}$ *in* (3.5) *or* $\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\alpha}}^{\mathrm{RR}}$ *in* (3.9), *we have*

$$n \cdot \hat{\boldsymbol{\alpha}}^T \hat{\boldsymbol{\alpha}} \xrightarrow{d} \chi_r^2,$$

where χ_r^2 *is the chi-square distribution with* r *degree of freedom*.

Therefore, the null hypothesis $H_{0,\boldsymbol{\alpha}} : \boldsymbol{\alpha} = \mathbf{0}$ is rejected if $n \cdot \hat{\boldsymbol{\alpha}}^T \hat{\boldsymbol{\alpha}} > \chi_{r,\alpha}^2$ where $\chi_{r,\alpha}^2$ is the upper-α quantile of χ_r^2. This test, combined with exploratory factor analysis, can be used as a diagnosis tool for practitioners to check whether the data gathering process has any confounding factors that can bias the multiple hypothesis testing.

4. Extension to multiple regression. In Sections 2 and 3, we assume that there is only one primary variable $\mathbf{X}$ and all the random variables $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ have mean $\mathbf{0}$. In practice, there may be several predictors, or we may want to include an intercept term in the regression model. Here, we develop a multiple regression extension to the original model (2.1a)–(2.1e).

Suppose we observe in total $d = d_0 + d_1$ random predictors that can be separated into two groups:

1. $\mathbf{X}_0$: $n \times d_0$ nuisance covariates that we would like to include in the regression model, and
2. $\mathbf{X}_1$: $n \times d_1$ primary variables whose effects we want to study.

For example, the intercept term can be included in $\mathbf{X}_0$ as a $n \times 1$ vector of 1 (i.e., a random variable with mean 1 and variance 0).

Leek and Storey [39] consider the case $d_0 = 0$ and $d_1 \geq 1$ for SVA and [59] consider the case $d_0 \geq 0$ and $d_1 = 1$ for LEAPP. Here, we study the confounder adjusted multiple regression in full generality, for any $d_0 \geq 0$ and $d_1 \geq 1$. Our model is

$$\mathbf{Y} = \mathbf{X}_0\mathbf{B}_0^T + \mathbf{X}_1\mathbf{B}_1^T + \mathbf{Z}\boldsymbol{\Gamma}^T + \mathbf{E}, \tag{4.1a}$$

$$\begin{pmatrix}\mathbf{X}_{0i}\\ \mathbf{X}_{1i}\end{pmatrix} \text{ are i.i.d. with } \mathrm{E}\left[\begin{pmatrix}\mathbf{X}_{0i}\\ \mathbf{X}_{1i}\end{pmatrix}\quad \begin{pmatrix}\mathbf{X}_{0i}\\ \mathbf{X}_{1i}\end{pmatrix}^T\right] = \boldsymbol{\Sigma}_{\mathbf{X}}, \tag{4.1b}$$

$$\mathbf{Z} \mid (\mathbf{X}_0, \mathbf{X}_1) \sim \mathrm{MN}\big(\mathbf{X}_0\mathbf{A}_0^T + \mathbf{X}_1\mathbf{A}_1^T, \mathbf{I}_n, \mathbf{I}_r\big) \quad \text{and} \tag{4.1c}$$

$$\mathbf{E} \perp\!\!\!\perp (\mathbf{X}_0, \mathbf{X}_1, \mathbf{Z}), \qquad \mathbf{E} \sim \mathrm{MN}(\mathbf{0}, \mathbf{I}_n, \boldsymbol{\Sigma}). \tag{4.1d}$$

The model does not specify means for $\mathbf{X}_{0i}$ and $\mathbf{X}_{1i}$; we do not need them. The parameters in this model are, for $i = 0$ or 1, $\mathbf{B}_i \in \mathbb{R}^{p\times d_i}$, $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times r}$, $\boldsymbol{\Sigma}_{\mathbf{X}} \in \mathbb{R}^{d\times d}$, and $\mathbf{A}_i \in \mathbb{R}^{r\times d_i}$. The parameters $\mathbf{A}$ and $\mathbf{B}$ are the matrix versions of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in model (2.1a)–(2.1e). Additionally, we assume $\boldsymbol{\Sigma}_{\mathbf{X}}$ is invertible. To clarify our purpose, we are primarily interested in estimating and testing for the significance of $\mathbf{B}_1$.

For the multiple regression model (4.1), we again consider the rotation matrix $\mathbf{Q}^T$ that is given by the QR decomposition $\begin{pmatrix}\mathbf{X}_0 & \mathbf{X}_1\end{pmatrix} = \mathbf{QU}$ where $\mathbf{Q} \in \mathbb{R}^{n\times n}$ is an orthogonal matrix and $\mathbf{U}$ is an upper triangular matrix of size $n \times d$. Therefore, we have

$$\mathbf{Q}^T\begin{pmatrix}\mathbf{X}_0 & \mathbf{X}_1\end{pmatrix} = \mathbf{U} = \begin{pmatrix}\mathbf{U}_{00} & \mathbf{U}_{01}\\ \mathbf{0} & \mathbf{U}_{11}\\ \mathbf{0} & \mathbf{0}\end{pmatrix},$$

where $\mathbf{U}_{00}$ is a $d_0 \times d_0$ upper triangular matrix and $\mathbf{U}_{11}$ is a $d_1 \times d_1$ upper triangular matrix. Now let the rotated $\mathbf{Y}$ be

$$\tilde{\mathbf{Y}} = \mathbf{Q}^T\mathbf{Y} = \begin{pmatrix}\tilde{\mathbf{Y}}_0\\ \tilde{\mathbf{Y}}_1\\ \tilde{\mathbf{Y}}_{-1}\end{pmatrix}, \tag{4.2}$$

where $\tilde{\mathbf{Y}}_0$ is $d_0 \times p$, $\tilde{\mathbf{Y}}_1$ is $d_1 \times p$ and $\tilde{\mathbf{Y}}_{-1}$ is $(n-d) \times p$, then we can partition the model into three parts: conditional on both $\mathbf{X}_0$ and $\mathbf{X}_1$ (hence $\mathbf{U}$),

$$\tilde{\mathbf{Y}}_0 = \mathbf{U}_{00}\mathbf{B}_0^T + \mathbf{U}_{01}\mathbf{B}_1^T + \tilde{\mathbf{Z}}_0\mathbf{\Gamma}^T + \tilde{\mathbf{E}}_0, \tag{4.3}$$

$$\tilde{\mathbf{Y}}_1 = \mathbf{U}_{11}\mathbf{B}_1^T + \tilde{\mathbf{Z}}_1\mathbf{\Gamma}^T + \tilde{\mathbf{E}}_1 \sim \mathrm{MN}\big(\mathbf{U}_{11}(\mathbf{B}_1 + \mathbf{\Gamma}\mathbf{A}_1)^T, \mathbf{I}_{d_1}, \mathbf{\Gamma}\mathbf{\Gamma}^T + \mathbf{\Sigma}\big), \tag{4.4}$$

$$\tilde{\mathbf{Y}}_{-1} = \tilde{\mathbf{Z}}_{-1}\mathbf{\Gamma}^T + \tilde{\mathbf{E}}_{-1} \sim \mathrm{MN}\big(\mathbf{0}, \mathbf{I}_{n-d}, \mathbf{\Gamma}\mathbf{\Gamma}^T + \mathbf{\Sigma}\big), \tag{4.5}$$

where $\tilde{\mathbf{Z}} = \mathbf{Q}^T\mathbf{Z}$ and $\tilde{\mathbf{E}} = \mathbf{Q}^T\mathbf{E} \overset{d}{=} \mathbf{E}$. Equation (4.3) corresponds to the nuisance parameters $\mathbf{B}_0$ and is discarded according to the ancillary principle. Equation (4.4) is the multivariate extension to (2.4) that is used to estimate $\mathbf{B}_1$ and equation (4.5) plays the same role as (2.5) to estimate $\mathbf{\Gamma}$ and $\mathbf{\Sigma}$.

We consider the asymptotics when $n, p \to \infty$ and d, r are fixed and known. Since d is fixed, the estimation of $\mathbf{\Gamma}$ is not different from the simple regression case and we can use the maximum likelihood factor analysis described in Section 3.1. Under Assumptions 1–3, the precision results of $\hat{\mathbf{\Gamma}}$ and $\hat{\mathbf{\Sigma}}$ in Lemma 3.1 still hold.

Let $\mathbf{\Sigma}_{\mathbf{X}}^{-1} = \mathbf{\Omega} = \left(\begin{smallmatrix} \mathbf{\Omega}_{00} & \mathbf{\Omega}_{01} \\ \mathbf{\Omega}_{10} & \mathbf{\Omega}_{11} \end{smallmatrix}\right)$. In the proof of Theorems 3.1 and 3.3, we consider a fixed sequence of $\mathbf{X}$ such that $\|\mathbf{X}\|_2/\sqrt{n} \to 1$. Similarly, we have the following lemma in the multiple regression scenario:

LEMMA 4.1. *As* $n \to \infty$, $\frac{1}{n}\mathbf{U}_{11}^T\mathbf{U}_{11} \overset{\text{a.s.}}{\to} \mathbf{\Omega}_{11}^{-1}$.

Similar to (3.2), we can rewrite (4.4) as

$$\tilde{\mathbf{Y}}_1^T\mathbf{U}_{11}^{-T} = \mathbf{B}_1 + \mathbf{\Gamma}\big(\mathbf{A}_1 + \tilde{\mathbf{W}}_1\mathbf{U}_{11}^{-T}\big) + \tilde{\mathbf{E}}_1\mathbf{U}_{11}^{-T},$$

where $\tilde{\mathbf{W}}_1 \sim \mathrm{MN}(\mathbf{0}, \mathbf{I}_{d_1}, \mathbf{I}_p)$ is independent from $\tilde{\mathbf{E}}_1$. As in Section 3.2, we derive statistical properties of the estimate of $\mathbf{B}_1$ for a fixed sequence of $\mathbf{X}$, $\tilde{\mathbf{W}}_1$ and $\mathbf{Z}$, which also hold unconditionally. For simplicity, we assume that the negative controls are a known set of variables $\mathcal{C}$ with $\mathbf{B}_{1,\mathcal{C}} = \mathbf{0}$. We can then estimate each column of $\mathbf{A}_1$ by applying the negative control (NC) or robust regression (RR) we discussed in Sections 3.2.1 and 3.2.2 to the corresponding row of $\tilde{\mathbf{Y}}_1\mathbf{U}_{11}^{-T}$, and then estimate $\mathbf{B}_1$ by

$$\hat{\mathbf{B}}_1 = \tilde{\mathbf{Y}}_1^T\mathbf{U}_{11}^{-T} - \hat{\mathbf{\Gamma}}\hat{\mathbf{A}}_1.$$

Notice that $\tilde{\mathbf{E}}_1\mathbf{U}_{11}^{-T} \sim \mathrm{MN}(\mathbf{0}, \mathbf{\Sigma}, \mathbf{U}_{11}^{-1}\mathbf{U}_{11}^{-T})$. Thus, the "samples" in the robust regression, which are actually the p variables in the original problem are still independent within each column. Though the estimates of each column of $\mathbf{A}_1$ may be correlated, we will show that the correlation will not affect inference on $\mathbf{B}_1$. As a result, we still get asymptotic results similar to Theorem 3.3 for the multiple regression model (4.1).

THEOREM 4.1. *Under Assumptions 1–6, if* $n, p \to \infty$, *with* $(\log p)^2/n \to 0$ *and* $\|\operatorname{vec}(\mathbf{B}_1)\|_1\sqrt{n}/p \to 0$, *then for any fixed index set* $\mathcal{S}$ *with finite cardinality* $|\mathcal{S}|$,

$$\sqrt{n}\big(\hat{\mathbf{B}}^{\mathrm{NC}}_{1,\mathcal{S}} - \mathbf{B}_{1,\mathcal{S}}\big) \xrightarrow{d} \mathrm{MN}\big(\mathbf{0}_{|\mathcal{S}|\times k_1}, \boldsymbol{\Sigma}_{\mathcal{S}} + \boldsymbol{\Delta}_{\mathcal{S}}, \boldsymbol{\Omega}_{11} + \mathbf{A}_1^T\mathbf{A}_1\big) \quad \textit{and} \tag{4.6}$$

$$\sqrt{n}\big(\hat{\mathbf{B}}^{\mathrm{RR}}_{1,\mathcal{S}} - \mathbf{B}_{1,\mathcal{S}}\big) \xrightarrow{d} \mathrm{MN}\big(\mathbf{0}_{|\mathcal{S}|\times k_1}, \boldsymbol{\Sigma}_{\mathcal{S}}, \boldsymbol{\Omega}_{11} + \mathbf{A}_1^T\mathbf{A}_1\big), \tag{4.7}$$

where $\boldsymbol{\Delta}_{\mathcal{S}}$ *is defined in Theorem 3.1.*

As for the asymptotic efficiency of this estimator, we again compare it to the oracle OLS estimator of $\mathbf{B}_1$ which observes confounding variables $\mathbf{Z}$ in (4.1). In the multiple regression model, we claim that $\hat{\mathbf{B}}^{\mathrm{RR}}_1$ still reaches the oracle asymptotic efficiency. In fact, let $\mathbf{B} = (\mathbf{B}_0 \;\; \mathbf{B}_1 \;\; \boldsymbol{\Gamma})$. The oracle OLS estimator of $\mathbf{B}$, $\hat{\mathbf{B}}^{\mathrm{OLS}}$, is unbiased and its vectorization has variance $\mathbf{V}^{-1} \otimes \boldsymbol{\Sigma}/n$ where

$$\mathbf{V} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{X}} & \boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{A}^T \\ \mathbf{A}\boldsymbol{\Sigma}_{\mathbf{X}} & \mathbf{I}_r + \mathbf{A}\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{A}^T \end{pmatrix} \qquad \text{for } \mathbf{A} = (\mathbf{A}_0 \quad \mathbf{A}_1).$$

By the block-wise matrix inversion formula, the top left $d \times d$ block of $\mathbf{V}^{-1}$ is $\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} + \mathbf{A}^T\mathbf{A}$. The variance of $\hat{\mathbf{B}}^{\mathrm{OLS}}_1$ only depends on the bottom right $d_1 \times d_1$ sub-block of this $d \times d$ block, which is simply $\boldsymbol{\Omega}_{11} + \mathbf{A}_1^T\mathbf{A}_1$. Therefore, $\hat{\mathbf{B}}^{\mathrm{OLS}}_1$ is unbiased and its vectorization has variance $(\boldsymbol{\Omega}_{11} + \mathbf{A}_1^T\mathbf{A}_1) \otimes \boldsymbol{\Sigma}/n$, matching the asymptotic variance of $\hat{\mathbf{B}}^{\mathrm{RR}}_1$ in Theorem 4.1.

5. Discussion.

5.1. *Confounding versus unconfounding.* The issue of multiple testing dependence arises because $\mathbf{Z}$ in the true model (1.1) is unobserved. We have focused on the case where $\mathbf{Z}$ is confounded with the primary variable. Some similar results were obtained earlier for the unconfounded case, corresponding to $\boldsymbol{\alpha} = 0$ in our notation. For example, [35] used a factor model to improve the efficiency of significance tests of the regression intercepts. Jin [31], Li and Zhong [40] developed more powerful procedures for testing $\boldsymbol{\beta}$ while still controlling FDR under unconfounded dependence.

In another related work, Fan, Han and Gu [21] imposed a factor structure on the unconfounded test statistics, whereas this paper and the articles discussed later in Section 5.3 assume a factor structure on the raw data. Fan, Han and Gu [21] used an approximate factor model to accurately estimate the false discovery proportion. Their correction procedure also includes a step of robust regression. Nevertheless, it is often difficult to interpret the factor structure of the test statistics. In comparison, the latent variables $\mathbf{Z}$ in our model (2.1a)–(2.1e), whether confounding or not,

can be interpreted as batch effects, laboratory conditions, or other systematic bias. Such problems are widely observed in genetics studies (see, e.g., the review article [37]).

As a final remark, some of the models and methods developed in the context of unconfounded hypothesis testing may be useful for confounded problems as well. For example, the relationship between $\mathbf{Z}$ and $\mathbf{X}$ needs not be linear as in (1.2). In certain applications, it may be more appropriate to use a time-series model [57] or a mixture model [19].

5.2. *Marginal effects versus direct effects*. In Section 1, we switched our interest from the marginal effects $\boldsymbol{\tau}$ in (1.3) to the direct effects $\boldsymbol{\beta}$. We believe that they are usually more scientifically meaningful and interpretable than the marginal effects. For instance, if the treated (control) samples are analyzed by machine A (machine B), and the machine A outputs higher values than B, we certainly do not want to include the effects of this machine to machine variation on the outcome measurements.

When model (2.1a)–(2.1e) is interpreted as a "structural equations model" [11], $\boldsymbol{\beta}$ is indeed the causal effect of $\mathbf{X}$ on $\mathbf{Y}$ [46]. In this paper, we do not make such structural assumptions about the data generating process. Instead, we use (2.1a)–(2.1e) to describe the screening procedure commonly applied in high throughput data analysis. The model (2.1a)–(2.1e) also describes how we think the marginal effects can be confounded, and hence different from the more meaningful direct effects $\boldsymbol{\beta}$. Additionally, the asymptotic setting in this paper is quite different from that in the traditional structural equations model.

5.3. *Comparison with existing confounder adjustment methods*. We discuss in more detail how previous methods of confounder adjustment, namely SVA [38, 39], RUV-4 [25, 26] and LEAPP [59], fit in the framework (2.1a)–(2.1e). See [47] for an alternative approach of bilinear regression with latent factors that is also motivated by high-throughput data analysis.

5.3.1. *SVA*. There are two versions of SVA: the reduced subset SVA (subset-SVA) of [38] and the iteratively reweighted SVA (IRW-SVA) of [39]. Both of them can be interpreted as the two-step statistical procedure in the framework (2.1a)–(2.1e). In the first step, SVA estimates the confounding factors by applying PCA to the residual matrix $(\mathbf{I} - \mathbf{H}_{\mathbf{X}})\mathbf{Y}$ where $\mathbf{H}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is the projection matrix of $\mathbf{X}$. In contrast, we applied factor analysis to the rotated residual matrix $(\mathbf{Q}^T\mathbf{Y})_{-1}$, where $\mathbf{Q}$ comes from the QR decomposition of $\mathbf{X}$ in Section 4. To see why these two approaches lead to the same estimate of $\boldsymbol{\Gamma}$, we introduce the block form of $\mathbf{Q} = \begin{pmatrix}\mathbf{Q}_1 & \mathbf{Q}_2\end{pmatrix}$ where $\mathbf{Q}_1 \in \mathbb{R}^{n\times d}$ and $\mathbf{Q}_2 \in \mathbb{R}^{n\times(n-d)}$. It is easy to show that $(\mathbf{Q}^T\mathbf{Y})_{-1} = \mathbf{Q}_2^T\mathbf{Y}$ and $(\mathbf{I} - \mathbf{H}_{\mathbf{X}})\mathbf{Y} = \mathbf{Q}_2\mathbf{Q}_2^T\mathbf{Y}$. Thus, our rotated matrix $(\mathbf{Q}^T\mathbf{Y})_{-1}$ decorrelates the residual matrix by left-multiplying by $\mathbf{Q}_2$

(because $\mathbf{Q}_2^T\mathbf{Q}_2 = \mathbf{I}_{n-d}$). Because $(\mathbf{Q}_2^T\mathbf{Y})^T\mathbf{Q}_2^T\mathbf{Y} = (\mathbf{Q}_2\mathbf{Q}_2^T\mathbf{Y})^T\mathbf{Q}_2\mathbf{Q}_2^T\mathbf{Y}$, $(\mathbf{Q}^T\mathbf{Y})_{-1}$ and $(\mathbf{I} - \mathbf{H}_\mathbf{X})\mathbf{Y}$ have the same sample covariance matrix, they will yield the same factor loading estimate under PCA and also under MLE. The main advantage of using the rotated matrix is theoretical: the rotated residual matrices have independent rows.

Because SVA does not assume an explicit relationship between the primary variable $\mathbf{X}$ and the confounders $\mathbf{Z}$, it cannot use the regression (3.2) to estimate $\boldsymbol{\alpha}$ (not even defined) and $\boldsymbol{\beta}$. Instead, the two SVA algorithms try to reconstruct the surrogate variables, which are essentially the confounders $\mathbf{Z}$ in our framework. Assuming the true primary effect $\boldsymbol{\beta}$ is sparse, the subset-SVA algorithm finds the outcome variables $\mathbf{Y}$ that have the smallest marginal correlation with $\mathbf{X}$ and uses their principal scores as $\mathbf{Z}$. Then it computes the p-values by F-tests comparing the linear regression models with and without $\mathbf{Z}$. This procedure can easily fail because a small marginal correlation does not imply no real effect of $\mathbf{X}$ due to the confounding factors. For example, most of the marginal effects in the gender study in Figure 1(b) are very small, but after confounding adjustment we find some are indeed significant (see Section 6.2).

The IRW-SVA algorithm modifies subset-SVA by iteratively choosing the subset. At each step, IRW-SVA gives a weight to each outcome variable based on how likely $\beta_j = 0$ the current estimate of surrogate variables. The weights are then used in a weighted PCA algorithm to update the estimated surrogate variables. IRW-SVA may be related to our robust regression estimator in Section 3.2.2 in the sense that an M-estimator is commonly solved by Iteratively Reweighted Least Squares (IRLS) and the weights also represent how likely the data point is an outlier. However, unlike IRLS, the iteratively reweighted PCA algorithm is not even guaranteed to converge. Some previous articles [25, 59] and our experiments in Section 6.1 and the supplementary material [62] show that SVA is outperformed by the NC and RR estimators in most confounded examples.

5.3.2. *RUV.* Gagnon-Bartsch, Jacob and Speed [25] derived the RUV-4 estimator of $\boldsymbol{\beta}$ via a sequence of heuristic calculations. In Section 3.2.1, we derived an analytically more tractable estimator $\hat{\boldsymbol{\beta}}^{\mathrm{NC}}$ which is actually the same as RUV-4, with the only difference being that we use MLE instead of PCA to estimate the factors and GLS instead of OLS in (3.5). To see why $\hat{\boldsymbol{\beta}}^{\mathrm{NC}}$ is essentially the same as $\hat{\boldsymbol{\beta}}^{\mathrm{RUV}-4}$, in the first step of RUV-4 it uses the residual matrix to estimate $\boldsymbol{\Gamma}$ and $\mathbf{Z}$, which yields the same estimate as using the rotated matrix (Section 5.3.1). In the second step, RUV-4 estimates $\boldsymbol{\beta}$ via a regression of $\mathbf{Y}$ on $\mathbf{X}$ and $\hat{\mathbf{Z}} = \mathbf{Q}(\tilde{\mathbf{Z}}_{-1}^T \ \hat{\boldsymbol{\alpha}}^T)^T$. This is equivalent to using ordinary least squares (OLS) to estimate $\boldsymbol{\alpha}$ in (3.4). Based on more heuristic calculations, the authors claim that the RUV-4 estimator has approximately the oracle variance. We rigorously prove this statement in Theorem 3.1 when the number of negative controls is large and give a finite sample correction when the negative controls are few. In Section 6.1, we show this correction is very useful to control the type I error and FDR in simulations.

5.3.3. *LEAPP.* We follow the two-step procedure and robust regression framework in LEAPP [59] in this paper, thus the test statistics t_j^{RR} are very similar to the test statistics in LEAPP. The difference is that LEAPP uses the Θ-IPOD algorithm of [54] for outlier detection, which is robust against outliers at leverage points but is not easy to analyze. Indeed [59] replaced it by the Dantzig selector in its theoretical Appendix. The classical M-estimator, although not robust to leverage points [64], allows us to study the theoretical properties more easily. In practice, LEAPP and RR estimator usually produce very similar results; see Section 6.1 for a numerical comparison.

5.4. *Inference when* $\mathbf{\Sigma}$ *is nondiagonal.* Our analysis is based on the assumption that the noise covariance matrix $\mathbf{\Sigma}$ is diagonal, though in many applications, the researcher might suspect that the outcome variables $\mathbf{Y}$ in model (2.1a)–(2.1e) are still correlated after conditioning on the latent factors. Typical examples include gene regulatory networks [17] and cross-sectional panel data [48], where the variable dependence sometimes cannot be fully explained by the latent factors or may simply require too many of them. Bai and Li [5] extend the theoretical results in [3] to approximate factor models allowing for weakly correlated noise. Approximate factor models have also been discussed in [20].

6. Numerical experiments.

6.1. *Simulations.* We have provided theoretical guarantees of confounder adjusting methods in various settings and the asymptotic regime of $n, p \to \infty$ (e.g., Theorems 3.1–3.4 and 4.1). Now we use numerical simulations to verify these results and further study the finite sample properties of our estimators and tests statistics.

The simulation data are generated from the single primary variable model (2.1a)–(2.1e). More specifically, X_i is a centered binary variable $(X_i + 1)/2 \overset{\text{i.i.d.}}{\sim}$ Bernoulli(0.5), and $\mathbf{Y}_i, \mathbf{Z}_i$ are generated according to (2.1a)–(2.1e).

For the parameters in the model, the noise variances are generated by $\sigma_j^2 \overset{\text{i.i.d.}}{\sim}$ InvGamma(3, 2), $j = 1, \ldots, p$, and so $\mathbb{E}(\sigma_j^2) = \mathrm{Var}(\sigma_j^2) = 1$. We set each $\alpha_k = \|\boldsymbol{\alpha}\|_2/\sqrt{r}$ equally for $k = 1, 2, \ldots, r$ where $\|\boldsymbol{\alpha}\|_2^2$ is set to 1, so the variance of X_i explained by the confounding factors is $R^2 = 50\%$. (Additional results for $R^2 = 5\%$ and 0 are in the supplementary material [62].) The primary effect $\boldsymbol{\beta}$ has independent components β_i taking the values $3\sqrt{1 + \|\boldsymbol{\alpha}\|_2^2}$ and 0 with probability $\pi = 0.05$ and $1 - \pi = 0.95$, respectively, so the nonzero effects are sparse and have effect size 3. This implies that the oracle estimator has power approximately $\mathrm{P}(\mathrm{N}(3, 1) > z_{0.025}) = 0.85$ to detect the signals at a significance level of 0.05. We set the number of latent factors r to be either 2 or 10. For the latent factor loading matrix $\mathbf{\Gamma}$, we take $\mathbf{\Gamma} = \tilde{\mathbf{\Gamma}}\mathbf{D}$ where $\tilde{\mathbf{\Gamma}}$ is a $p \times r$ orthogonal matrix sampled

uniformly from the Stiefel manifold $V_r(\mathbb{R}^p)$, the set of all $p \times r$ orthogonal matrix. Based on Assumption 3, we set the latent factor strength $\mathbf{D} = \sqrt{p} \cdot \operatorname{diag}(d_1, \ldots, d_r)$ where $d_k = 3 - 2(k-1)/(r-1)$ thus d_1 to d_r are distributed evenly inside the interval $[3, 1]$. As the number of factors r can be easily estimated for this strong factor setting (more discussions can be found in [45]), we assume that the number r of factors is known to all of the algorithms in this simulation.

We set $p = 5000$, $n = 100$ or 500 to mimic the data size of many genetic studies. For the negative control scenario, we choose $|\mathcal{C}| = 30$ negative controls at random from the zero positions of $\boldsymbol{\beta}$. We expect that negative control methods would perform better with a larger value of $|\mathcal{C}|$ and worse with a smaller value. The choice $|\mathcal{C}| = 30$ is around the size of the spike-in controls in many microarray experiments [26]. For the loss function in our sparsity scenario, we use Tukey's bisquare which is optimized via IRLS with an ordinary least-square fit as the starting values of the coefficients. Finally, each of the four combinations of n and r is randomly repeated 100 times.

We compare the performance of nine different approaches. There are two baseline methods: the "naive" method estimates $\boldsymbol{\beta}$ by a linear regression of $\mathbf{Y}$ on just the observed primary variable $\mathbf{X}$ and calculates p-values using the classical t-tests, while the "oracle" method regresses $\mathbf{Y}$ on both $\mathbf{X}$ and the confounding variables $\mathbf{Z}$ as described in Section 2.1. There are three methods in the RUV-4/negative controls family: the RUV-4 method [25], our "NC" method which computes test statistics using $\hat{\boldsymbol{\beta}}^{\mathrm{NC}}$ and its variance estimate $(1 + \|\hat{\boldsymbol{\alpha}}\|_2^2)(\hat{\boldsymbol{\Sigma}} + \hat{\boldsymbol{\Delta}})$, and our "NC-ASY" method which uses the same $\hat{\boldsymbol{\beta}}^{\mathrm{NC}}$ but estimates its variance by $(1 + \|\hat{\boldsymbol{\alpha}}\|_2^2)\hat{\boldsymbol{\Sigma}}$. We compare four methods in the SVA/LEAPP/sparsity family: these are "IRW-SVA" [39], "LEAPP" [59], the "LEAPP(RR)" method which is our RR estimator using M-estimation at the robustness stage and computes the test-statistics using (3.13), and the "LEAPP(RR-MAD)" method which uses the median absolute deviation (MAD) of the test statistics in (3.13) to calibrate them (see Section 3.3).

To measure the performance of these methods, we report the type I error (Theorem 3.4), power, false discovery proportion (FDP) and precision of hypotheses with the smallest 100 p-values in the 100 simulations. For both the type I error and power, we set the significance level to be 0.05. For FDP, we use the Benjamini–Hochberg procedure with FDR controlled at 0.2. These metrics are plotted in Figure 2 under different settings of n and r.

First, from Figure 2, we see that the oracle method has exactly the same type I error and FDP as specified, while the naive method and SVA fail drastically. SVA performs performs better than the naive method in terms of the precision of the smallest 100 p-values, but is still much worse than other methods. Next, for the negative control scenario, as we only have $|\mathcal{C}| = 30$ negative controls, ignoring the inflated variance term $\boldsymbol{\Delta}_S$ in Theorem 3.1 will lead to overdispersed test statistics, and that is why the type I error and FDP of both NC-ASY and RUV-4 are much larger than the nominal level. By contrast, the NC method correctly

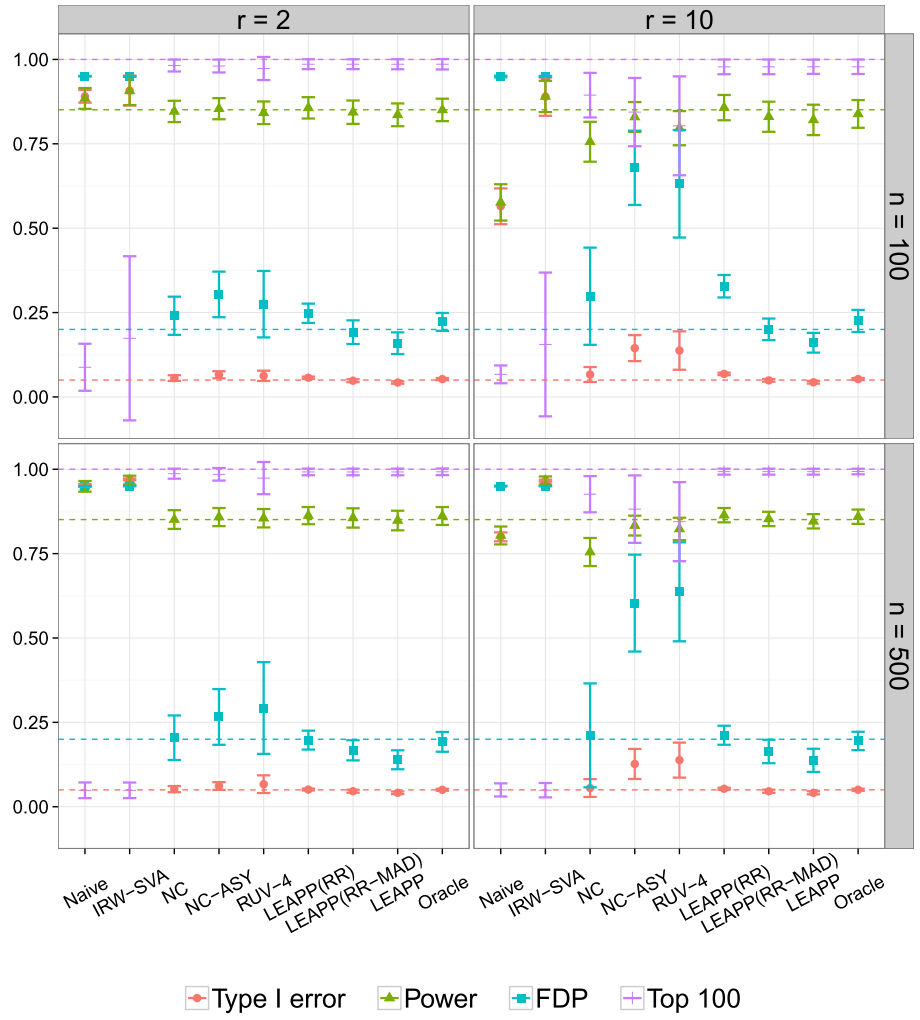


FIG. 2. *Compare the performance of nine different approaches* (*from left to right*): *naive regression ignoring the confounders* (*Naive*), *IRW-SVA*, *negative control with finite sample correction* (*NC*) *in* (3.7), *negative control with asymptotic oracle variance* (*NC-ASY*) *in* (3.8), *RUV*-4, *robust regression* [*LEAPP*(*RR*)], *robust regression with calibration* [*LEAPP*(*RR-MAD*)], *LEAPP*, *oracle regression which observes the confounders* (*Oracle*). *The error bars are one standard deviation over* 100 *repeated simulations*. *The three dashed horizontal lines from bottom to top are the nominal significance level*, *FDR level and oracle power*, *respectively*.

controls type I error and FDP by considering the variance inflation, though as expected it loses some power compared with the oracle. For the sparsity scenario, the "LEAPP(RR)" method performs as the asymptotic theory predicted when $n = 500$, while when $n = 100$ the p-values seem a bit too small. This is not surprising because the asymptotic oracle variance in Theorem 3.3 can be optimistic when the

sample size is not sufficiently large, as we discussed in Remark 3.2. On the other hand, the methods which use empirical calibration for the variance of test statistics, namely the original LEAPP and "LEAPP(RR-MAD)," control both FDP and type I error for data of small sample size in our simulations. The price for the finite sample calibration is that it tends to be slightly conservative, resulting in a loss of power to some extent.

In conclusion, the simulation results are consistent with our theoretical guarantees when p is as large as 5000 and n is as large as 500. When n is small, the variance of the test statistics will be larger than the asymptotic variance for the sparsity scenario and we can use empirical calibrations (such as MAD) to adjust for the difference.

6.2. *Real data examples*. In this section, we return to the three motivating real data examples in Section 1. The main goal here is to demonstrate a practical procedure for confounder adjustment and show that our asymptotic results are reasonably accurate in real data. In an open-source R package `cate` (available on CRAN), we also provide the necessary tools to carry out the procedure.

6.2.1. *The datasets*. First, we briefly describe the three datasets. The first dataset [55] tries to identify candidate genes associated with the extent of emphysema and can be downloaded from the GEO database (Series GSE22148). We preprocessed the data using the standard Robust Multi-array Average (RMA) approach [30]. The primary variable of interest is the severity (moderate or severe) of the Chronic Obstructive Pulmonary Disease (COPD). The dataset also include age, gender, batch and date of the 143 sampled patients which are served as nuisance covariates.

The second and third datasets are taken from [25] where they used them to compare RUV methods with other methods such as SVA and LEAPP. The original scientific studies are [61] and [10], respectively. The primary variable of interest is gender in both datasets, though the original objective in [10] is to identify genes associated with Alzheimer's disease. Gagnon-Bartsch, Jacob and Speed [25] switch the primary variable to gender in order to have a gold standard: the differentially expressed genes should mostly come from or relate to the X or Y chromosome. We follow their suggestion and use this standard to study the performance of our RR estimator. In addition, as the first COPD dataset also contains gender information of the samples, we apply this suggestion and use gender as the primary variable for the COPD data as a supplementary dataset.

Finally, we want to mention that the second dataset has repeated samples from the same individuals while the individual information is lost. We suspect that the individual information are then strong latent factors which caused the atypical concentration of the histograms in Figure 1(b) and Figure 1(d). This suggests necessity of a latent factor model for this dataset.

6.2.2. *Confounder adjustment*. Recall that without the confounder adjustment, the distribution of the regression t-statistics in these datasets can be skewed, noncentered, underdispersed or overdispersed as shown in Figure 1. The adjustment method used here is the maximum likelihood factor analysis described in Section 3.1 followed by the robust regression (RR) method with Tukey's bisquare loss described in Section 3.2.2. Since the true number of confounders is unknown, we increase r from 1 to $n/2$ and study the empirical performance. We report the results without empirical calibration for illustrative purposes, though in practice we suggest using calibration for better control of type I errors and FDP.

In Table 2 and Figure 3, we present the results after confounder adjustment for the three datasets. We report two groups of summary statistics in Table 2: the first group is several summary statistics of all the z-statistics computed using (3.13), including the mean, median, standard deviation, median absolute deviation (scaled for consistency of normal distribution), skewness and the medcouple. The medcouple [12] is a robust measure of skewness. After subtracting the median observation, some positive and some negative values remain. For any pair of values $x_1 \geq 0$ and $x_2 \leq 0$ with $x_1 + |x_2| > 0$, one can compute $(x_1 - |x_2|)/(x_1 + |x_2|)$. The medcouple is the median of all those ratios. The second group of statistics has performance metrics to evaluate the effectiveness of the confounder adjustment. See the caption of Table 2 for more detail.

In all three datasets, the z-statistics become more centered at 0 and less skewed as we include a few confounders in the model. Though the standard deviation (SD) suggests overdispersed variance, the overdispersion will go away if we add MAD calibration as SD and MAD have similar values. The similarity between SD and MAD values also indicates that the majority of statistics after confounder adjustment are approximately normally distributed. Note that the medcouple values shrink towards zero after adjustment, suggesting that skewness then only arises from small fraction of the genes, which is in accordance with our assumptions that the primary effects should be sparse.

In practice, some latent factors may be too weak to meet Assumption 3 (i.e. $d_j \ll \sqrt{p}$), making it difficult to choose an appropriate r. A practical way to pick the number of confounders r with presence of heteroscedastic noise we investigate here is the bi-cross-validation (BCV) method of [45], which uses randomly held-out submatrices to estimate the mean squared error of reconstructing factor loading matrix. It is shown in [45] that BCV outperforms many existing methods in recovering the latent signal matrix and the number of factors r, especially in high-dimensional datasets ($n, p \to \infty$). In Figure 3, we demonstrate the performance of BCV on these three datasets. The r selected by BCV is respectively 33, 25 and 11 [Figure 3(a), (c) and (e)], and they all result in the presumed shape of z-statistics distribution [Figure 3(b), (d) and (f)]. For the second and the third datasets where we have a gold standard, the r selected by BCV has near optimal performance in selecting genes on the X/Y chromosome [columns 3 and 4 in Table 2(b) and (c)].

TABLE 2

Summary of the adjusted z-statistics. The first group is summary statistics of the z-statistics before the empirical calibration. The second group is some performance metrics after the empirical calibration, including total number of significant genes of p-value less than 0.01 *in Remark* 3.2 (*#sig.*), *number of the genes on X/Y chromosome that have p-value less than* 0.01 (*X/Y*), *the number among the* 100 *most significant genes that are on the X/Y chromosome* (*top* 100) *and the p-value of the confounding test in Section* 3.3.2. *The bold row corresponds to the r selected by BCV* (*Figure* 3)

(a) Dataset 1 ($n = 143$, $p = 54{,}675$). Primary variable: severity of COPD

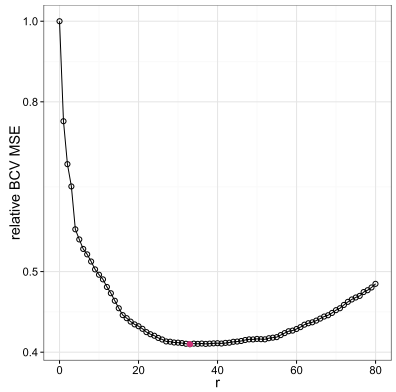
r	mean	median	sd	mad	skewness	medc.	#sig.	*p*-value
0	−0.16	0.024	2.65	2.57	−0.104	−0.091	164	NA
1	−0.45	−0.39	2.85	2.52	−0.25	0.00074	1162	0.0057
2	0.012	−0.039	1.35	1.33	0.139	0.042	542	<1e–10
3	0.014	−0.05	1.43	1.41	0.169	0.048	552	<1e–10
5	−0.029	−0.11	1.52	1.48	0.236	0.057	647	<1e–10
7	−0.1	−0.14	1.42	1.35	0.109	0.027	837	<1e–10
10	−0.06	−0.085	1.13	1.12	0.103	0.022	506	<1e–10
20	−0.083	−0.095	1.2	1.19	0.0604	0.0095	479	<1e–10
33	**−0.099**	**−0.11**	**1.33**	**1.3**	**0.0727**	**0.0056**	**579**	**<1e–10**
40	−0.1	−0.12	1.43	1.4	0.0775	0.0072	585	<1e–10
50	−0.16	−0.17	1.58	1.53	0.0528	0.0032	678	<1e–10

(b) Dataset 2 ($n = 84$, $p = 12{,}600$). Primary variable: gender

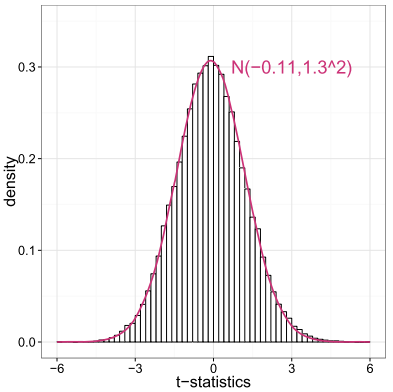
r	mean	median	sd	mad	skewness	medc.	#sig.	X/Y	top 100	*p*-value
0	0.11	0.043	0.36	0.237	2.99	0.2	1036	58	11	NA
1	−0.44	−0.47	1.06	1.04	0.688	0.035	108	20	20	0.74
2	−0.14	−0.15	1.15	1.13	0.601	0.015	113	21	21	0.31
3	0.013	0.012	1.13	1.08	0.795	−0.01	168	34	28	0.03
5	0.044	0.019	1.18	1.08	0.878	0.017	238	32	27	0.0083
7	0.03	0.012	1.26	1.15	0.784	0.0062	269	35	25	0.006
10	0.023	0.00066	1.36	1.24	0.661	0.011	270	38	27	0.019
15	0.049	0.022	1.46	1.31	0.584	0.012	296	36	29	0.00082
20	0.029	−0.0009	1.53	1.36	0.502	0.019	314	36	28	7.2e–07
25	**0.048**	**0.012**	**1.68**	**1.48**	**0.452**	**0.026**	**354**	**37**	**27**	**1.1e–06**
30	0.026	0.012	1.82	1.61	0.436	0.0068	337	40	27	8.7e–08
40	0.061	0.046	2.07	1.79	0.642	0.0028	363	41	27	7.7e–10

(c) Dataset 3 ($n = 31$, $p = 22{,}283$). Primary variable: gender

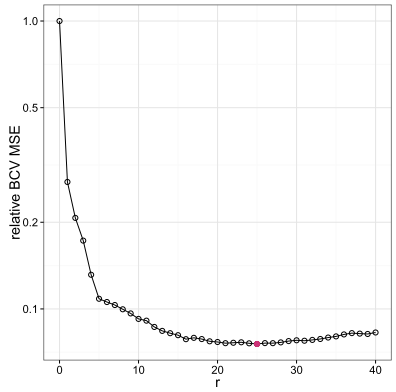
r	mean	median	sd	mad	skewness	medc.	#sig.	X/Y	top 100	*p*-value
0	−1.8	−1.8	0.599	0.513	−3.46	0.082	418	39	20	NA
1	−0.55	−0.56	1.09	1.01	−1.53	0.01	261	29	23	0.00024
2	−0.2	−0.22	1.2	1.11	−0.99	0.014	320	38	22	0.00014
3	−0.096	−0.12	1.27	1.18	−0.844	0.017	311	42	25	0.00014
5	−0.33	−0.32	1.31	1.22	−1.29	−0.011	305	35	23	2.1e–07
7	−0.37	−0.36	1.46	1.36	−0.855	−0.0099	300	38	23	4.0e–07
11	**−0.13**	**−0.12**	**1.51**	**1.36**	**−0.601**	**−0.0051**	**432**	**48**	**31**	**1.8e–09**
15	−0.12	−0.13	1.83	1.62	−0.341	0.013	492	54	25	2.3e–08
20	−0.13	−0.14	2.61	2.23	−0.327	0.0045	613	50	26	4.0e–06



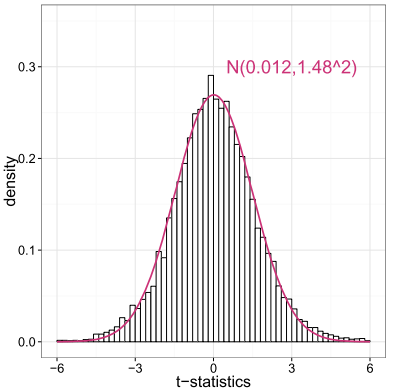
(a) Dataset 1: BCV selects $r = 33$.



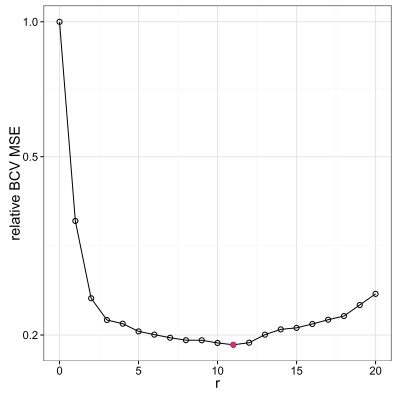
(b) Dataset 1: histogram.



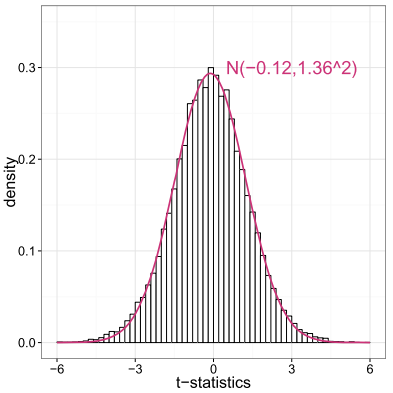
(c) Dataset 2: BCV selects $r = 25$.



(d) Dataset 2: histogram.



(e) Dataset 3: BCV selects $r = 11$.



(f) Dataset 3: histogram.

FIG. 3. *Histograms of z-statistics after confounder adjustment* (*without calibration*) *using the number of confounders r selected by bi-cross-validation.*

Another method we applied is proposed by [43] based on the empirical distribution of eigenvalues. This method estimates r as 2, 9 and 3, respectively, for the three datasets. Table 3 of [25] has the "top 100" values for RUV-4 on the second and third dataset. They reported 26 for LEAPP, 28 for RUV-4, and 27 for SVA in the second dataset, and 27 for LEAPP, 31 for RUV-4, and 26 for SVA in the third dataset. Notice that the precision of the top 100 significant genes is relatively stable when r is above certain number. Intuitively, the factor analysis is applied to the residuals of $\mathbf{Y}$ on $\mathbf{X}$ and the overestimated factors also have very small eigenvalues, thus they usually do not change $\hat{\boldsymbol{\beta}}$ a lot. See also [25] for more discussion on the robustness of the negative control estimator to overestimating r.

Lastly, we want to point out that both the small sample size of the datasets and presence of weak factors can result in overdispersed variance of the test statistics. The BCV plots indicate presence of many weak factors in the first two datasets. In the third dataset, the sample size n is only 31, so the adjustment result is not ideal. Nevertheless, the empirical performance (e.g., number of X/Y genes in top 100) suggests it is still beneficial to adjust for the confounders.

Acknowledgments. The authors thank Bhaswar Bhattacharya, Murat Erdogdu, Jian Li, Weijie Su and Yunting Sun for helpful discussion.

J. Wang and Q. Zhao contributed equally to this paper. This paper was completed when Jingshu Wang and Qingyuan Zhao were Ph.D. candidates at Stanford University.

SUPPLEMENTARY MATERIAL

Supplement to "Confounder adjustment in multiple hypothesis testing" (DOI: 10.1214/16-AOS1511SUPP; .pdf). We provide detailed proof for the theoretical results in this paper and some additional numerical results.

REFERENCES

[1] ALTER, O., BROWN, P. O. and BOTSTEIN, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. USA* **97** 10101–10106.

[2] ANDERSON, T. W. and RUBIN, H. (1956). Statistical inference in factor analysis. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1954–1955, *Vol. V* 111–150. Univ. California Press, Berkeley and Los Angeles. MR0084943

[3] BAI, J. and LI, K. (2012). Statistical analysis of factor models of high dimension. *Ann. Statist.* **40** 436–465. MR3014313

[4] BAI, J. and LI, K. (2014). Theory and methods of panel data models with interactive effects. *Ann. Statist.* **42** 142–170. MR3178459

[5] BAI, J. and LI, K. (2016). Maximum likelihood estimation and inference for approximate factor models of high dimension. *Rev. Econ. Stat.* **98** 298–309.

[6] BAI, J. and NG, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70** 191–221. MR1926259

[7] BAI, J. and NG, S. (2006). Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica* **74** 1133–1150. MR2238213
[8] BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392
[9] BENJAMINI, Y. and YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29** 1165–1188. MR1869245
[10] BLALOCK, E. M., GEDDES, J. W., CHEN, K. C., PORTER, N. M., MARKESBERY, W. R. and LANDFIELD, P. W. (2004). Incipient Alzheimer's disease: Microarray correlation analyses reveal major transcriptional and tumor suppressor responses. *Proc. Natl. Acad. Sci. USA* **101** 2173–2178.
[11] BOLLEN, K. A. (1989). *Structural Equations with Latent Variables*. Wiley, New York. MR0996025
[12] BRYS, G., HUBERT, M. and STRUYF, A. (2004). A robust measure of skewness. *J. Comput. Graph. Statist.* **13** 996–1017. MR2109062
[13] CHANDRASEKARAN, V., PARRILO, P. A. and WILLSKY, A. S. (2012). Latent variable graphical model selection via convex optimization. *Ann. Statist.* **40** 1935–1967. MR3059067
[14] CLARKE, S. and HALL, P. (2009). Robustness of multiple testing procedures against dependence. *Ann. Statist.* **37** 332–358. MR2488354
[15] CRAIG, A., CLOAREC, O., HOLMES, E., NICHOLSON, J. K. and LINDON, J. C. (2006). Scaling and normalization effects in NMR spectroscopic metabonomic data sets. *Anal. Chem.* **78** 2262–2267.
[16] DESAI, K. H. and STOREY, J. D. (2012). Cross-dimensional inference of dependent high-dimensional data. *J. Amer. Statist. Assoc.* **107** 135–151. MR2949347
[17] DE LA FUENTE, A., BING, N., HOESCHELE, I. and MENDES, P. (2004). Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics* **20** 3565–3574.
[18] EFRON, B. (2007). Correlation and large-scale simultaneous significance testing. *J. Amer. Statist. Assoc.* **102** 93–103. MR2293302
[19] EFRON, B. (2010). Correlated z-values and the accuracy of large-scale statistical estimates. *J. Amer. Statist. Assoc.* **105** 1042–1055. MR2752597
[20] FAN, J. and HAN, X. (2013). Estimation of false discovery proportion with unknown dependence. Available at arXiv:1305.7007.
[21] FAN, J., HAN, X. and GU, W. (2012). Estimating false discovery proportion under arbitrary covariance dependence. *J. Amer. Statist. Assoc.* **107** 1019–1035. MR3010887
[22] FARE, T. L., COFFEY, E. M., DAI, H., HE, Y. D., KESSLER, D. A., KILIAN, K. A., KOCH, J. E., LEPROUST, E., MARTON, M. J., MEYER, M. R. et al. (2003). Effects of atmospheric ozone on microarray data quality. *Anal. Chem.* **75** 4672–4675.
[23] FISHER, R. A. (1935). *The Design of Experiments*. Oliver & Boyd, Edinburgh.
[24] FRIGUET, C., KLOAREG, M. and CAUSEUR, D. (2009). A factor model approach to multiple testing under dependence. *J. Amer. Statist. Assoc.* **104** 1406–1415. MR2750571
[25] GAGNON-BARTSCH, J., JACOB, L. and SPEED, T. P. (2013). Removing unwanted variation from high dimensional data with negative controls. Technical Report 820, Dept. Statistics, Univ. California, Berkeley, Berkeley, CA.
[26] GAGNON-BARTSCH, J. A. and SPEED, T. P. (2012). Using control genes to correct for unwanted variation in microarray data. *Biostatistics* **13** 539–552.
[27] GASCH, A. P., SPELLMAN, P. T., KAO, C. M., CARMEL-HAREL, O., EISEN, M. B., STORZ, G., BOTSTEIN, D. and BROWN, P. O. (2000). Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* **11** 4241–4257.
[28] GREENLAND, S., ROBINS, J. M. and PEARL, J. (1999). Confounding and collapsibility in causal inference. *Statist. Sci.* **14** 29–46.

[29] GRZEBYK, M., WILD, P. and CHOUANIÈRE, D. (2004). On identification of multi-factor models with correlated residuals. *Biometrika* **91** 141–151. MR2050465

[30] IRIZARRY, R. A., HOBBS, B., COLLIN, F., BEAZER-BARCLAY, Y. D., ANTONELLIS, K. J., SCHERF, U., SPEED, T. P. et al. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4** 249–264.

[31] JIN, J. (2012). Comment: "Estimating false discovery proportion under arbitrary covariance dependence." [MR3010887] *J. Amer. Statist. Assoc.* **107** 1042–1045. MR3010891

[32] KISH, L. (1959). Some statistical problems in research design. *Am. Sociol. Rev.* **24** 328–338.

[33] KORN, E. L., TROENDLE, J. F., MCSHANE, L. M. and SIMON, R. (2004). Controlling the number of false discoveries: Application to high-dimensional genomic data. *J. Statist. Plann. Inference* **124** 379–398. MR2080371

[34] KUROKI, M. and PEARL, J. (2014). Measurement bias and effect restoration in causal inference. *Biometrika* **101** 423–437. MR3215357

[35] LAN, W. and DU, L. (2014). A factor-adjusted multiple testing procedure with application to mutual fund selection. Available at arXiv:1407.5515.

[36] LAZAR, C., MEGANCK, S., TAMINAU, J., STEENHOFF, D., COLETTA, A., MOLTER, C., WEISS-SOLÍS, D. Y., DUQUE, R., BERSINI, H. and NOWÉ, A. (2013). Batch effect removal methods for microarray gene expression data integration: A survey. *Brief. Bioinform.* **14** 469–490.

[37] LEEK, J. T., SCHARPF, R. B., BRAVO, H. C., SIMCHA, D., LANGMEAD, B., JOHNSON, W. E., GEMAN, D., BAGGERLY, K. and IRIZARRY, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **11** 733–739.

[38] LEEK, J. T. and STOREY, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* **3** 1724–1735.

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

[40] LI, J. and ZHONG, P.-S. (2016). A rate optimal procedure for recovering sparse differences between high-dimensional means under dependence. *Ann. Statist.* To appear.

[41] LIN, D. W., COLEMAN, I. M., HAWLEY, S., HUANG, C. Y., DUMPIT, R., GIFFORD, D., KEZELE, P., HUNG, H., KNUDSEN, B. S., KRISTAL, A. R. et al. (2006). Influence of surgical manipulation on prostate gene expression: Implications for molecular correlates of treatment effects and disease prognosis. *J. Clin. Oncol.* **24** 3763–3770.

[42] MARONNA, R. A., MARTIN, R. D. and YOHAI, V. J. (2006). *Robust Statistics*: *Theory and Methods*. Wiley, Chichester. MR2238141

[43] ONATSKI, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *Rev. Econ. Stat.* **92** 1004–1016.

[44] OWEN, A. B. (2005). Variance of the number of false discoveries. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67** 411–426. MR2155346

[45] OWEN, A. B. and WANG, J. (2016). Bi-cross-validation for factor analysis. *Statist. Sci.* **31** 119–139. MR3458596

[46] PEARL, J. (2009). *Causality*: *Models*, *Reasoning*, *and Inference*, 2nd ed. Cambridge Univ. Press, Cambridge. MR2548166

[47] PERRY, P. O. and PILLAI, N. S. (2013). Degrees of freedom for combining regression with factor analysis. Preprint. Available at arXiv:1310.7269.

[48] PESARAN, M. H. (2004). General diagnostic tests for cross section dependence in panels. Cambridge Working Papers in Economics No. 0435.

[49] PRICE, A. L., PATTERSON, N. J., PLENGE, R. M., WEINBLATT, M. E., SHADICK, N. A. and REICH, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38** 904–909.

[50] RANSOHOFF, D. F. (2005). Bias as a threat to the validity of cancer molecular-marker research. *Nat. Rev. Cancer* **5** 142–149.

[51] RHODES, D. R. and CHINNAIYAN, A. M. (2005). Integrative analysis of the cancer transcriptome. *Nat. Genet.* **37** S31–S37.

[52] SCHWARTZMAN, A. (2010). Comment: "Correlated z-values and the accuracy of large-scale statistical estimates." [MR2752597] *J. Amer. Statist. Assoc.* **105** 1059–1063. MR2752600

[53] SCHWARTZMAN, A., DOUGHERTY, R. F. and TAYLOR, J. E. (2008). False discovery rate analysis of brain diffusion direction maps. *Ann. Appl. Stat.* **2** 153–175. MR2415598

[54] SHE, Y. and OWEN, A. B. (2011). Outlier detection using nonconvex penalized regression. *J. Amer. Statist. Assoc.* **106** 626–639. MR2847975

[55] SINGH, D., FOX, S. M., TAL-SINGER, R., PLUMB, J., BATES, S., BROAD, P., RILEY, J. H. and CELLI, B. (2011). Induced sputum genes associated with spirometric and radiological disease severity in COPD ex-smokers. *Thorax* **66** 489–495.

[56] STOREY, J. D., TAYLOR, J. E. and SIEGMUND, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **66** 187–205. MR2035766

[57] SUN, W. and CAI, T. T. (2009). Large-scale multiple testing under dependence. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **71** 393–424. MR2649603

[58] SUN, Y. (2011). On latent systemic effects in multiple hypotheses. Ph.D. thesis, Stanford University.

[59] SUN, Y., ZHANG, N. R. and OWEN, A. B. (2012). Multiple hypothesis testing adjusted for latent variables, with an application to the AGEMAP gene expression data. *Ann. Appl. Stat.* **6** 1664–1688. MR3058679

[60] TUSHER, V. G., TIBSHIRANI, R. and CHU, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* **98** 5116–5121.

[61] VAWTER, M. P., EVANS, S., CHOUDARY, P., TOMITA, H., MEADOR-WOODRUFF, J., MOLNAR, M., LI, J., LOPEZ, J. F., MYERS, R., COX, D. et al. (2004). Gender-specific gene expression in post-mortem human brain: Localization to sex chromosomes. *Neuropsychopharmacology* **29** 373–384.

[62] WANG, J., ZHAO, Q., HASTIE, T. and OWEN, A. B. (2017). Supplement to "Confounder adjustment in multiple hypothesis testing." DOI:10.1214/16-AOS1511SUPP.

[63] WANG, S., CUI, G. and LI, K. (2015). Factor-augmented regression models with structural change. *Econom. Lett.* **130** 124–127. MR3336182

[64] YOHAI, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *Ann. Statist.* **15** 642–656. MR0888431

J. WANG
Q. ZHAO
DEPARTMENT OF STATISTICS
THE WHARTON SCHOOL
UNIVERSITY OF PENNSYLVANIA
400 HUNTSMAN HALL, 3730 WALNUT ST
PHILADELPHIA, PENNSYLVANIA 19104
USA
E-MAIL: jingshuw@wharton.upenn.edu
qyzhao@wharton.upenn.edu

T. HASTIE
A. B. OWEN
DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
390 SERRA MALL
STANFORD, CALIFORNIA 94305
USA
E-MAIL: hastie@stanford.edu
owen@stanford.edu