Statistica Sinica Preprint No: SS-2018-0345	
Title	Unifying and Generalizing Methods for Removing Unwanted Variation Based on Negative Controls
Manuscript ID	SS-2018-0345
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202018.0345
Complete List of Authors	David Gerard and Matthew Stephens
Corresponding Author	David Gerard
E-mail	gerard.1787@gmail.com
Notice: Accepted version subject to English editing.	

Statistica Sinica

Unifying and Generalizing Methods for Removing Unwanted Variation Based on Negative Controls

[1]David Gerard and [2]Matthew Stephens

[1]*Department of Mathematics and Statistics, American University, Washington DC, USA*

[2]*Departments of Human Genetics and Statistics, University of Chicago, Chicago IL, USA*

Abstract: Unwanted variation, including hidden confounding, is a well-known problem in many fields, particularly large-scale gene expression studies. Recent proposals to use control genes, genes assumed to be unassociated with the covariates of interest, have led to new methods to deal with this problem. Going by the moniker **R**emoving **U**nwanted **V**ariation (RUV), there are many versions, e.g. RUV1, RUV2, RUV4, RUVinv, RUVrinv, RUVfun. In this paper, we introduce a general framework, RUV*, that both unites and generalizes these approaches. This unifying framework helps clarify connections between existing methods. In particular we provide conditions under which RUV2 and RUV4 are equivalent. The RUV* framework also preserves an advantage of RUV approaches, their modularity, which facilitates the development of novel methods based on existing matrix imputation algorithms. We illustrate this by implementing RUVB, a version of RUV* based on Bayesian factor analysis. In realistic simulations based on real data we found that RUVB is competitive with existing methods in terms of both power and calibration, although we also highlight the challenges of

providing consistently reliable calibration among data sets.

Key words and phrases: batch effect, correlated test, gene expression, hidden confounding, negative control, RNA-seq, unobserved confounding, unwanted variation

1. Introduction Many experiments and observational studies in genetics are overwhelmed with unwanted sources of variation. Examples include: processing date (Akey et al., 2007), the lab that collected a sample (Irizarry et al., 2005), the batch in which a sample was processed (Leek et al., 2010), and subject attributes such as environmental factors (Gibson, 2008) and ancestry (Price et al., 2006). These factors, if ignored, can result in disastrously wrong conclusions (Gilad and Mizrahi-Man, 2015). They can induce dependencies between samples, and inflate test statistics, making it difficult to control false discovery rates (Efron, 2004, 2008, 2010).

Many of the sources of variation mentioned above are likely to be observed, in which case standard methods exist to control for them (Johnson et al., 2007). However, every study likely also contains unobserved sources of unwanted variation, and these can cause equally profound problems (Leek and Storey, 2007), even in the ideal case of a randomized experiment. To illustrate this we took 20 samples from an RNA-seq dataset (GTEx Consortium, 2015) and randomly assigned them into two groups of 10 samples.

Since group assignment is entirely independent of the expression levels of each gene, the group labels are theoretically unassociated with all genes and any observed "signal" must be artefactual. Figure 1 shows histograms of the p-values from two-sample t-tests for three different randomizations. In each case the distribution of the p-values differs greatly from the theoretical uniform distribution. Thus, even in this ideal scenario where group labels were randomly assigned, problems can arise. One way to understand this is to note that the same randomization is being applied to all genes. Consequently, if many genes are affected by an unobserved factor, and this factor happens by chance to be correlated with the randomization, then the p-value distributions will be non-uniform. In this sense the problems here can be viewed as being due to correlation among the p values; see Efron (2010) for extensive discussion. (The issue of whether the problems in any given study are caused by correlation, confounding, or something different is both interesting and subtle; see discussion in Efron (2010); Schwartzman (2010) for example. For this reason we adopt the "unwanted variation" terminology from Gagnon-Bartsch and Speed (2012), rather than alternative terminologies such as "hidden confounding".)

In recent years many methods have been introduced to try to solve problems due to unwanted variation. Perhaps the simplest approach is to

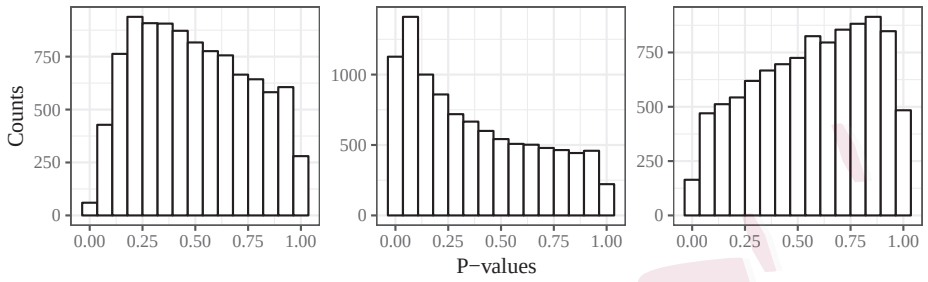

Figure 1: Histograms of p-values from two-sample t-tests when group labels are randomly assigned to samples. Each panel is from a different random seed. The p-value distributions all clearly deviate from uniform.

estimate sources of unwanted variation using principal components analysis (Price et al., 2006), and then to control for these factors by using them as covariates in subsequent analyses. Indeed, in genome-wide association studies this simple method is widely used. However, in gene expression studies it suffers from the problem that the principal components will typically also contain the signal of interest, so controlling for them risks removing that signal. To address this Leek and Storey (2007, 2008) introduced **S**urrogate **V**ariable **A**nalysis (SVA), which uses an iterative algorithm to attempt to estimate latent factors that do not include the signal of interest (see also Lucas et al. (2006)). To account for unwanted variation, SVA assumes a factor-augmented regression model (Section 2.1), which has a long history (Fisher and Mackenzie, 1923; Cochran, 1943; Williams, 1952; Tukey, 1962;

Gollob, 1968; Mandel, 1969, 1971; Efron and Morris, 1972; Freeman, 1973; Gabriel, 1978, and others). Since SVA, a large number of different approaches have emerged along similar lines, including Behzadi et al. (2007); Kang et al. (2008); Carvalho et al. (2008); Kang et al. (2008); Stegle et al. (2008); Friguet et al. (2009); Kang et al. (2010); Listgarten et al. (2010); Stegle et al. (2010); Wu and Aryee (2010); Gagnon-Bartsch and Speed (2012); Fusi et al. (2012); Stegle et al. (2012); Sun et al. (2012); Gagnon-Bartsch et al. (2013); Mostafavi et al. (2013); Perry and Pillai (2013); Yang et al. (2013); Chen and Zhou (2017); Lee et al. (2017); Wang et al. (2017); Gerard and Stephens (2018); McKennan and Nicolae (2018a,b), among others.

As noted above, a key difficulty in adjusting for unwanted variation in expression studies is distinguishing between the effect of a treatment and the effect of factors that are correlated with a treatment. Available methods deal with this problem in different ways. Here we focus on methods that use "negative controls" to help achieve this goal. In the context of a gene expression study, a negative control is a gene whose expression is assumed to be unassociated with all covariates (and treatments) of interest. Under this assumption, negative controls can be used to separate sources of unwanted variation from the treatment effects. The idea of using negative controls in this way appears in Lucas et al. (2006), and has been recently popularized

by Gagnon-Bartsch and Speed (2012) and Gagnon-Bartsch et al. (2013) in a series of methods and software going by the moniker **R**emoving **U**nwanted **V**ariation (RUV). There are many methods, including RUV2 (for RUV 2-step), RUV4, RUVinv (a special case of RUV4), RUVrinv, RUVfun, and RUV1.

Understanding the relative merits and properties of the different RUV methods, which are all aimed at solving essentially the same problem, is a non-trivial task. The main contribution of this paper is to outline a general framework, RUV*, that encompasses all versions of RUV (Section 4). RUV* represents the problem as a general matrix imputation procedure, both providing a unifying conceptual framework, and opening up new approaches based on the large literature in matrix imputation. Our RUV* framework also provides a simple and modular way to account for uncertainty in the estimated sources of unwanted variation, which is an issue ignored by most methods. On the way to this general framework we make detailed connections between RUV2 and RUV4, exploiting the formulation in Wang et al. (2017).

On notation: throughout we denote matrices using bold capital letters ($\boldsymbol{A}$), except for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, which are also matrices. Bold lowercase letters are vectors ($\boldsymbol{a}$), and non-bold lowercase letters are scalars (a). Where there is

no chance for confusion, we use non-bold lowercase to denote scalar elements of vectors or matrices. For example, a_{ij} is the (i,j)th element of $\boldsymbol{A}$ and a_i is the ith element of $\boldsymbol{a}$. The notation $\boldsymbol{A}_{n\times m}$ denotes that the matrix $\boldsymbol{A}$ is an n by m matrix. The matrix transpose is denoted $\boldsymbol{A}^\intercal$ and the matrix inverse is denoted $\boldsymbol{A}^{-1}$. Sets are generally denoted with calligraphic letters ($\mathcal{A}$), and the complement of a set is denoted with a bar ($\bar{\mathcal{A}}$).

2. RUV4 and RUV2

2.1 Review of the two-step rotation method

Most existing approaches to this problem (Leek and Storey, 2007, 2008; Gagnon-Bartsch and Speed, 2012; Sun et al., 2012; Gagnon-Bartsch et al., 2013; Wang et al., 2017) are based in some way on using Factor Analysis (FA) to capture unwanted variation. Specifically, they assume:

$$\boldsymbol{Y}_{n\times p} = \boldsymbol{X}_{n\times k}\boldsymbol{\beta}_{k\times p} + \boldsymbol{Z}_{n\times q}\boldsymbol{\alpha}_{q\times p} + \boldsymbol{E}_{n\times p}, \tag{2.1}$$

where, in the context of a gene-expression study, y_{ij} is the normalized expression level of the jth gene on the ith sample, $\boldsymbol{X}$ contains the observed covariates, $\boldsymbol{\beta}$ contains the coefficients of $\boldsymbol{X}$, $\boldsymbol{Z}$ is a matrix of unobserved factors (sources of unwanted variation), $\boldsymbol{\alpha}$ contains the coefficients of $\boldsymbol{Z}$, and $\boldsymbol{E}$ contains independent (Gaussian) errors with means 0 and column-

specific variances $\text{var}(e_{ij}) = \sigma_j^2$. In this model, the only known quantities are $\boldsymbol{Y}$ and $\boldsymbol{X}$.

To fit (2.1), it is common to apply a two-step approach (e.g. Gagnon-Bartsch et al. (2013); Sun et al. (2012); Wang et al. (2017)). The first step regresses out $\boldsymbol{X}$ and then, using the residuals of this regression, estimates $\boldsymbol{\alpha}$ and the σ_j's. The second step then assumes that $\boldsymbol{\alpha}$ and the σ_j's are known and estimates $\boldsymbol{\beta}$ and $\boldsymbol{Z}$. Wang et al. (2017) helpfully frame this two-step approach as a rotation followed by estimation in two independent models. We now review this approach.

First, we let $\boldsymbol{X} = \boldsymbol{QR}$ denote the QR decomposition of $\boldsymbol{X}$, where $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ is an orthogonal matrix ($\boldsymbol{Q}^\intercal\boldsymbol{Q} = \boldsymbol{Q}\boldsymbol{Q}^\intercal = \boldsymbol{I}_n$) and $\boldsymbol{R}_{n\times k} = (\boldsymbol{R}_1^\intercal, \boldsymbol{0})^\intercal$, where $\boldsymbol{R}_1 \in \mathbb{R}^{k\times k}$ is an upper-triangular matrix. Multiplying (2.1) on the left by $\boldsymbol{Q}^\intercal$ yields

$$\boldsymbol{Q}^\intercal\boldsymbol{Y} = \boldsymbol{R}\boldsymbol{\beta} + \boldsymbol{Q}^\intercal\boldsymbol{Z}\boldsymbol{\alpha} + \boldsymbol{Q}^\intercal\boldsymbol{E}. \tag{2.2}$$

Suppose that $k = k_1 + k_2$, where the first k_1 covariates of $\boldsymbol{X}$ are not of direct interest, but are included because of various modeling decisions (e.g. an intercept term, or covariates that need to be controlled for). The last k_2 columns of $\boldsymbol{X}$ are the variables of interest whose putative associations with $\boldsymbol{Y}$ the researcher wishes to test. Let $\boldsymbol{Y}_1 \in \mathbb{R}^{k_1\times p}$ be the first k_1 rows of $\boldsymbol{Q}^\intercal\boldsymbol{Y}$, $\boldsymbol{Y}_2 \in \mathbb{R}^{k_2\times p}$ be the next k_2 rows of $\boldsymbol{Q}^\intercal\boldsymbol{Y}$, and $\boldsymbol{Y}_3 \in \mathbb{R}^{(n-k)\times p}$ be

the last $n-k$ rows of $\boldsymbol{Q}^\intercal \boldsymbol{Y}$. Conformably partition $\boldsymbol{Q}^\intercal \boldsymbol{Z}$ into $\boldsymbol{Z}_1$, $\boldsymbol{Z}_2$, and $\boldsymbol{Z}_3$, and $\boldsymbol{Q}^\intercal \boldsymbol{E}$ into $\boldsymbol{E}_1$, $\boldsymbol{E}_2$, and $\boldsymbol{E}_3$. Let

$$\boldsymbol{R}_1 = \begin{pmatrix} \boldsymbol{R}_{11} & \boldsymbol{R}_{12} \\ 0 & \boldsymbol{R}_{22} \end{pmatrix}.$$

Finally, partition $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\intercal, \boldsymbol{\beta}_2^\intercal)^\intercal$ so that $\boldsymbol{\beta}_1 \in \mathbb{R}^{k_1 \times p}$ contains the coefficients for the first k_1 covariates and $\boldsymbol{\beta}_2 \in \mathbb{R}^{k_2 \times p}$ contains the coefficients for the last k_2 covariates. Then (2.2) may be written as three models

$$\boldsymbol{Y}_1 = \boldsymbol{R}_{11}\boldsymbol{\beta}_1 + \boldsymbol{R}_{12}\boldsymbol{\beta}_2 + \boldsymbol{Z}_1\boldsymbol{\alpha} + \boldsymbol{E}_1, \tag{2.3}$$

$$\boldsymbol{Y}_2 = \boldsymbol{R}_{22}\boldsymbol{\beta}_2 + \boldsymbol{Z}_2\boldsymbol{\alpha} + \boldsymbol{E}_2, \tag{2.4}$$

$$\boldsymbol{Y}_3 = \boldsymbol{Z}_3\boldsymbol{\alpha} + \boldsymbol{E}_3. \tag{2.5}$$

Importantly, the error terms in (2.3), (2.4), and (2.5) are mutually independent. This follows from the easily-proved fact that $\boldsymbol{E}$ is equal in distribution to $\boldsymbol{Q}^\intercal \boldsymbol{E}$. The two-step estimation procedure mentioned above becomes: first, estimate $\boldsymbol{\alpha}$ and the σ_j's using (2.5); second, estimate $\boldsymbol{\beta}_2$ and $\boldsymbol{Z}_2$ given $\boldsymbol{\alpha}$ and the σ_j's using (2.4). Equation (2.3) contains the nuisance parameters $\boldsymbol{\beta}_1$ and is ignored.

2.2 RUV4

One approach to distinguishing between unwanted variation and effects of interest is to use "control genes" (Lucas et al., 2006; Gagnon-Bartsch and

Speed, 2012). A control gene is a gene that is assumed *a priori* to be unassociated with the covariate(s) of interest. More formally, the set of control genes, $\mathcal{C} \subseteq \{1, \ldots, p\}$, has the property that

$$\beta_{ij} = 0 \text{ for all } i = k_1 + 1, \ldots, k, \text{ and } j \in \mathcal{C},$$

and is a subset of the truly null genes. Examples of control genes used in practice are spike-in controls (Jiang et al., 2011) used to adjust for technical factors (such as sample batch) and housekeeping genes (Eisenberg and Levanon, 2013) used to adjust for both technical and biological factors (such as subject ancestry).

RUV4 (Gagnon-Bartsch et al., 2013) uses control genes to estimate $\boldsymbol{\beta}_2$ in the presence of unwanted variation. Let $\boldsymbol{Y}_{2\mathcal{C}} \in \mathbb{R}^{k_2 \times m}$ denote the submatrix of $\boldsymbol{Y}_2$ with columns that correspond to the m control genes. Similarly subset the relevant columns to obtain $\boldsymbol{\beta}_{2\mathcal{C}} \in \mathbb{R}^{k_2 \times m}$, $\boldsymbol{\alpha}_{\mathcal{C}} \in \mathbb{R}^{q \times m}$, and $\boldsymbol{E}_{2\mathcal{C}} \in \mathbb{R}^{k_2 \times m}$. The steps for RUV4, including a variation from Wang et al. (2017), are presented in Procedure 1. (For simplicity we focus on point estimates of effects here, deferring assessment of standard errors to Section S6 of the Supplementary Material.)

The key idea in Procedure 1 is that for the control genes model (2.4)

Procedure 1 RUV4

1: Estimate $\boldsymbol{\alpha}$ and $\boldsymbol{\Sigma}$ using FA (Definition 1) on $\boldsymbol{Y}_3$ in (2.5). Call these estimates $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\Sigma}}$.

2: Estimate $\boldsymbol{Z}_2$ using control genes (equation (2.8)). Let $\hat{\boldsymbol{\Sigma}}_{\mathcal{C}} = \text{diag}(\hat{\sigma}^2_{j_1}, \ldots, \hat{\sigma}^2_{j_m})$ for $j_i \in \mathcal{C}$ for all $i = 1, \ldots, m$.

RUV4 in Gagnon-Bartsch et al. (2013) estimates $\boldsymbol{Z}_2$ by ordinary least squares (OLS)

$$\hat{\boldsymbol{Z}}_2 = \boldsymbol{Y}_{2\mathcal{C}} \hat{\boldsymbol{\alpha}}_{\mathcal{C}}^\intercal (\hat{\boldsymbol{\alpha}}_{\mathcal{C}} \hat{\boldsymbol{\alpha}}_{\mathcal{C}}^\intercal)^{-1}. \tag{2.6}$$

Alternatively, Wang et al. (2017) implement a variation on RUV4 (which we call CATE, and is implemented in the R package `cate`) that estimates $\boldsymbol{Z}_2$ by generalized least squares (GLS)

$$\hat{\boldsymbol{Z}}_2 = \boldsymbol{Y}_{2\mathcal{C}} \hat{\boldsymbol{\Sigma}}_{\mathcal{C}}^{-1} \hat{\boldsymbol{\alpha}}_{\mathcal{C}}^\intercal (\hat{\boldsymbol{\alpha}}_{\mathcal{C}} \hat{\boldsymbol{\Sigma}}_{\mathcal{C}}^{-1} \hat{\boldsymbol{\alpha}}_{\mathcal{C}}^\intercal)^{-1}. \tag{2.7}$$

3: Estimate $\boldsymbol{\beta}_2$ using (2.4) by

$$\hat{\boldsymbol{\beta}}_2 = \boldsymbol{R}_{22}^{-1} (\boldsymbol{Y}_2 - \hat{\boldsymbol{Z}}_2 \hat{\boldsymbol{\alpha}}).$$

becomes

$$\begin{aligned} \boldsymbol{Y}_{2\mathcal{C}} &= \boldsymbol{R}_{22}\boldsymbol{\beta}_{2\mathcal{C}} + \boldsymbol{Z}_2\hat{\boldsymbol{\alpha}}_{\mathcal{C}} + \boldsymbol{E}_{2\mathcal{C}}, \\ &= \boldsymbol{Z}_2\hat{\boldsymbol{\alpha}}_{\mathcal{C}} + \boldsymbol{E}_{2\mathcal{C}}, \qquad (2.8) \\ e_{2\mathcal{C}ij} &\overset{ind}{\sim} N(0, \hat{\sigma}_j^2). \qquad (2.9) \end{aligned}$$

The equality in (2.8) follows from the property of control genes that $\boldsymbol{\beta}_{2\mathcal{C}} = \mathbf{0}$. Step 2 of Procedure 1 uses (2.8) to estimate $\boldsymbol{Z}_2$.

Step 1 of Procedure 1 requires a FA of $\boldsymbol{Y}_3$. We formally define a FA as follows.

Definition 1. A Factor Analysis (*FA*), $\mathcal{F}$, of rank $q \leq \min(n, p)$ on $\boldsymbol{Y} \in \mathbb{R}^{n\times p}$ is a set of three functions $\mathcal{F} = \{\hat{\boldsymbol{\Sigma}}(\boldsymbol{Y}), \hat{\boldsymbol{Z}}(\boldsymbol{Y}), \hat{\boldsymbol{\alpha}}(\boldsymbol{Y})\}$ such that $\hat{\boldsymbol{\Sigma}}(\boldsymbol{Y}) \in \mathbb{R}^{p\times p}$ is diagonal with positive diagonal entries, $\hat{\boldsymbol{Z}}(\boldsymbol{Y}) \in \mathbb{R}^{n\times q}$ has rank q, and $\hat{\boldsymbol{\alpha}}(\boldsymbol{Y}) \in \mathbb{R}^{q\times p}$ has rank q.

RUV4 allows the analyst to use any FA they desire. Thus, RUV4 is not a single method, but a collection of methods indexed by the FA used. When we want to be explicit about this indexing, we write RUV4($\mathcal{F}$).

2.3 RUV2

Procedure 2 summarizes the RUV2 method introduced in Gagnon-Bartsch and Speed (2012). It involves two steps: first estimate the factors causing

unwanted variation from the control genes, and then include these factors as covariates in the regression models for the non-control genes. Gagnon-Bartsch et al. (2013) extend this procedure to deal with nuisance covariates by adding a preliminary step that rotates $\boldsymbol{Y}$ and $\boldsymbol{X}$ onto the orthogonal complement of the space spanned by the nuisance covariates (equation (64) in Gagnon-Bartsch et al., 2013).

Procedure 2 RUV2 (without nuisance covariates; Gagnon-Bartsch and Speed (2012))

1: From (2.1), estimate $\boldsymbol{Z}$ by FA on $\boldsymbol{Y}_{\mathcal{C}}$. Call this estimate $\hat{\boldsymbol{Z}}$.

2: Estimate $\boldsymbol{\beta}$ by regressing $\boldsymbol{Y}$ on $(\boldsymbol{X}, \hat{\boldsymbol{Z}})$. That is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\intercal \boldsymbol{S} \boldsymbol{X})^{-1} \boldsymbol{X}^\intercal \boldsymbol{S} \boldsymbol{Y},$$

where $\boldsymbol{S} = \boldsymbol{I}_n - \hat{\boldsymbol{Z}}(\hat{\boldsymbol{Z}}^\intercal \hat{\boldsymbol{Z}})^{-1} \hat{\boldsymbol{Z}}^\intercal$.

Like RUV4, RUV2 is a class of methods indexed by the FA used, which we here denote RUV2$_{old}(\mathcal{F})$. In Procedure 3 we present a method, RUV2$_{new}(\mathcal{F})$, that we then prove is equivalent to RUV2$_{old}$ (Theorem 1; proved in Section S2 of the Supplementary Material).

Theorem 1. *For a given orthogonal matrix* $\boldsymbol{Q} \in \mathbb{R}^{n \times n}$ *and an arbitrary*

Procedure 3 RUV2 in rotated model framework of Section 2.1

1: Estimate $\boldsymbol{Z}_2$ and $\boldsymbol{Z}_3$ by FA on $\binom{\boldsymbol{Y}_{2C}}{\boldsymbol{Y}_{3C}}$. Call these estimates $\hat{\boldsymbol{Z}}_2$ and $\hat{\boldsymbol{Z}}_3$.

2: Estimate $\boldsymbol{\alpha}$ and $\boldsymbol{\Sigma}$ by regressing $\boldsymbol{Y}_3$ on $\hat{\boldsymbol{Z}}_3$. That is

$$\hat{\boldsymbol{\alpha}} = (\hat{\boldsymbol{Z}}_3^\intercal \hat{\boldsymbol{Z}}_3^{-1})\hat{\boldsymbol{Z}}_3^\intercal \boldsymbol{Y}_3 \text{ and} \tag{2.10}$$

$$\hat{\boldsymbol{\Sigma}} = \text{diag}[(\boldsymbol{Y}_3 - \hat{\boldsymbol{Z}}_3\hat{\boldsymbol{\alpha}})^\intercal(\boldsymbol{Y}_3 - \hat{\boldsymbol{Z}}_3\hat{\boldsymbol{\alpha}})]/(n-k-q). \tag{2.11}$$

3: Estimate $\boldsymbol{\beta}_2$ with

$$\hat{\boldsymbol{\beta}}_2 = \boldsymbol{R}_{22}^{-1}(\boldsymbol{Y}_2 - \hat{\boldsymbol{Z}}_2\hat{\boldsymbol{\alpha}}). \tag{2.12}$$

non-singular matrix $\boldsymbol{A}(\boldsymbol{Y})$ that (possibly) depends on $\boldsymbol{Y}$, suppose

$$\mathcal{F}_1(\boldsymbol{Y}) = \{\hat{\boldsymbol{\Sigma}}(\boldsymbol{Y}), \hat{\boldsymbol{Z}}(\boldsymbol{Y}), \hat{\boldsymbol{\alpha}}(\boldsymbol{Y})\}, \text{ and} \tag{2.13}$$

$$\mathcal{F}_2(\boldsymbol{Y}) = \{\hat{\boldsymbol{\Sigma}}(\boldsymbol{Q}^\intercal\boldsymbol{Y}), \boldsymbol{Q}\hat{\boldsymbol{Z}}(\boldsymbol{Q}^\intercal\boldsymbol{Y})\boldsymbol{A}(\boldsymbol{Y}), \boldsymbol{A}^{-1}(\boldsymbol{Y})\hat{\boldsymbol{\alpha}}(\boldsymbol{Q}^\intercal\boldsymbol{Y})\}. \tag{2.14}$$

Then

$$RUV2_{old}(\mathcal{F}_2) = RUV2_{new}(\mathcal{F}_1).$$

That is, Procedure 2 using FA (2.13) *is equivalent to Procedure 3 using FA* (2.14).

The equivalence of RUV2_{old} and RUV2_{new} in Theorem 1 involves using different factor analyses in each procedure. One can ask under what

conditions the two procedures would be equivalent if given the *same* FA. Corollary 1 states that it suffices for the FA to be *left orthogonally equivariant* (see Section S3 of the Supplementary Material for proof).

Definition 2. A FA of rank q on $\boldsymbol{Y} \in \mathbb{R}^{n\times p}$ is *left orthogonally equivariant* if

$$\{\hat{\boldsymbol{\Sigma}}(\boldsymbol{Q}^\intercal\boldsymbol{Y}), \hat{\boldsymbol{Z}}(\boldsymbol{Q}^\intercal\boldsymbol{Y})\boldsymbol{A}(\boldsymbol{Y}), \boldsymbol{A}(\boldsymbol{Y})^{-1}\hat{\boldsymbol{\alpha}}(\boldsymbol{Q}^\intercal\boldsymbol{Y})\} = \{\hat{\boldsymbol{\Sigma}}(\boldsymbol{Y}), \boldsymbol{Q}^\intercal\hat{\boldsymbol{Z}}(\boldsymbol{Y}), \hat{\boldsymbol{\alpha}}(\boldsymbol{Y})\},$$

for all fixed orthogonal $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ and an arbitrary non-singular $\boldsymbol{A}(\boldsymbol{Y}) \in \mathbb{R}^{q\times q}$ that (possibly) depends on $\boldsymbol{Y}$.

Corollary 1. *Suppose $\mathcal{F}$ is a left orthogonally equivariant FA. Then*

$$RUV2_{old}(\mathcal{F}) = RUV2_{new}(\mathcal{F}).$$

A well-known FA that is left orthogonally equivariant is the truncated singular value decomposition (formally defined in Section S1 of the Supplementary Material), and this is the only option in the R package `ruv` (Gagnon-Bartsch, 2015).

From now on we use RUV2 to refer to Procedure 3 and not Procedure 2, even if the FA is *not* orthogonally equivariant. (By Theorem 1, this corresponds to Procedure 2 with some other FA.)

3. RUV3

Gagnon-Bartsch et al. (2013) provide a lengthy discussion comparing RUV2 with RUV4 (their section 3.4). However, they provide no mathematical equivalencies. We now introduce RUV3, a procedure that is a version of both RUV2 and RUV4. We show that it is the only such procedure that is both RUV2 and RUV4.

3.1 The RUV3 procedure

The main goal in all methods is to estimate $\boldsymbol{\beta}_{2\bar{\mathcal{C}}}$, the coefficients corresponding to the non-control genes. This involves incorporating information from four models, which can be written in matrix form:

$$\begin{pmatrix} \boldsymbol{Y}_{2\mathcal{C}} & \boldsymbol{Y}_{2\bar{\mathcal{C}}} \\ \boldsymbol{Y}_{3\mathcal{C}} & \boldsymbol{Y}_{3\bar{\mathcal{C}}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{Z}_2\boldsymbol{\alpha}_{\mathcal{C}} + \boldsymbol{E}_{2\mathcal{C}} & \boldsymbol{R}_{22}\boldsymbol{\beta}_{2\bar{\mathcal{C}}} + \boldsymbol{Z}_2\boldsymbol{\alpha}_{\bar{\mathcal{C}}} + \boldsymbol{E}_{2\bar{\mathcal{C}}} \\ \boldsymbol{Z}_3\boldsymbol{\alpha}_{\mathcal{C}} + \boldsymbol{E}_{3\mathcal{C}} & \boldsymbol{Z}_3\boldsymbol{\alpha}_{\bar{\mathcal{C}}} + \boldsymbol{E}_{3\bar{\mathcal{C}}} \end{pmatrix}. \quad (3.1)$$

The major difference between RUV2 and RUV4 is how the estimation procedures interact in (3.1); see Figure 2 for illustration. RUV2 performs FA on $(\boldsymbol{Y}_{2\mathcal{C}}^\intercal, \boldsymbol{Y}_{3\mathcal{C}}^\intercal)^\intercal$, then regresses $\boldsymbol{Y}_{3\bar{\mathcal{C}}}$ on the estimated factor loadings. RUV4 performs FA on $(\boldsymbol{Y}_{3\mathcal{C}}, \boldsymbol{Y}_{3\bar{\mathcal{C}}})$, then regresses $\boldsymbol{Y}_{2\mathcal{C}}$ on the estimated factors. The main goal in both, however, is to estimate $\boldsymbol{Z}_2\boldsymbol{\alpha}_{\bar{\mathcal{C}}}$ given $\boldsymbol{Y}_{2\mathcal{C}}$, $\boldsymbol{Y}_{3\mathcal{C}}$, and $\boldsymbol{Y}_{3\bar{\mathcal{C}}}$.

Estimating $\boldsymbol{Z}_2\boldsymbol{\alpha}_{\bar{\mathcal{C}}}$ given $\boldsymbol{Y}_{2\mathcal{C}}$, $\boldsymbol{Y}_{3\mathcal{C}}$, and $\boldsymbol{Y}_{3\bar{\mathcal{C}}}$ is, in essence, a matrix

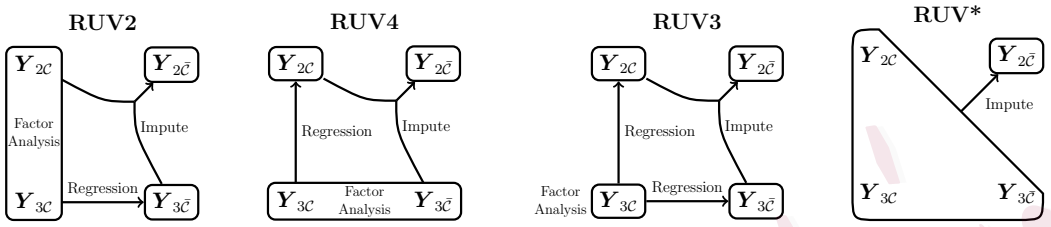

Figure 2: Pictorial representation of the differences between RUV2, RUV4, RUV3, and RUV*.

imputation problem. In the context of matrix imputation (and not removing unwanted variation), Owen and Wang (2016), generalizing methods of Owen and Perry (2009), suggest that after applying a FA to $\boldsymbol{Y}_{3\mathcal{C}}$, one use the estimates $\hat{\boldsymbol{Z}}_2$ and $\hat{\boldsymbol{\alpha}}_{\bar{\mathcal{C}}}$ from (3.2) and (3.3), respectively, and then set $\widehat{\boldsymbol{Z}_2\boldsymbol{\alpha}_{\bar{\mathcal{C}}}} = \hat{\boldsymbol{Z}}_2\hat{\boldsymbol{\alpha}}_{\bar{\mathcal{C}}}$. This corresponds to a FA followed by two regressions followed by an imputation step. Following the theme of this paper, we would add an additional step and estimate $\boldsymbol{\beta}_{2\bar{\mathcal{C}}}$ with (3.5).

This estimation procedure (Procedure 4) unifies RUV2 and RUV4, and so we call it RUV3. The unification is formalized in the following theorem (see Section S4 of the Supplementary Material for proof).

Theorem 2. *A procedure is both a version of RUV4 (Procedure 1) and RUV2 (Procedure 3) if, and only if, it is also a version of RUV3 (Procedure 4).*

Procedure 4 RUV3

1: Perform FA on $\boldsymbol{Y}_{3\mathcal{C}}$ to obtain estimates of $\boldsymbol{Z}_3$, $\boldsymbol{\alpha}_{\mathcal{C}}$ and $\boldsymbol{\Sigma}_{\mathcal{C}}$.

2: Regress $\boldsymbol{Y}_{2\mathcal{C}}$ on $\hat{\boldsymbol{\alpha}}_{\mathcal{C}}$ to obtain an estimate of $\boldsymbol{Z}_2$ and regress $\boldsymbol{Y}_{3\bar{\mathcal{C}}}$ on $\hat{\boldsymbol{Z}}_3$ to obtain estimates of $\boldsymbol{\alpha}_{\bar{\mathcal{C}}}$ and $\boldsymbol{\Sigma}_{\bar{\mathcal{C}}}$. That is

$$\hat{\boldsymbol{Z}}_2 = \boldsymbol{Y}_{2\mathcal{C}}\hat{\boldsymbol{\Sigma}}_{\mathcal{C}}^{-1}\hat{\boldsymbol{\alpha}}_{\mathcal{C}}^{\intercal}(\hat{\boldsymbol{\alpha}}_{\mathcal{C}}\hat{\boldsymbol{\Sigma}}_{\mathcal{C}}^{-1}\hat{\boldsymbol{\alpha}}_{\mathcal{C}}^{\intercal})^{-1}, \tag{3.2}$$

$$\hat{\boldsymbol{\alpha}}_{\bar{\mathcal{C}}} = (\hat{\boldsymbol{Z}}_3^{\intercal}\hat{\boldsymbol{Z}}_3)^{-1}\hat{\boldsymbol{Z}}_3^{\intercal}\boldsymbol{Y}_{3\bar{\mathcal{C}}}, \tag{3.3}$$

$$\hat{\boldsymbol{\Sigma}}_{\bar{\mathcal{C}}} = \text{diag}\left[(\boldsymbol{Y}_{3\bar{\mathcal{C}}} - \hat{\boldsymbol{Z}}_3\hat{\boldsymbol{\alpha}}_{\bar{\mathcal{C}}})^{\intercal}(\boldsymbol{Y}_{3\bar{\mathcal{C}}} - \hat{\boldsymbol{Z}}_3\hat{\boldsymbol{\alpha}}_{\bar{\mathcal{C}}})\right]/(n-k-q). \tag{3.4}$$

3: Estimate $\boldsymbol{\beta}_2$ by

$$\hat{\boldsymbol{\beta}}_2 = \boldsymbol{R}_{22}^{-1}(\boldsymbol{Y}_{2\bar{\mathcal{C}}} - \hat{\boldsymbol{Z}}_2\hat{\boldsymbol{\alpha}}_{\bar{\mathcal{C}}}). \tag{3.5}$$

4. A more general framework: RUV*

A key insight that arises from unifying RUV2 and RUV4 (and RUV3) into a single framework is that they share a common goal: estimation of $\boldsymbol{Z}_2\boldsymbol{\alpha}_{\bar{\mathcal{C}}}$, which represents the combined effects of all sources of unwanted variation on $\boldsymbol{Y}_{2\bar{\mathcal{C}}}$. This insight suggests a more general approach: any matrix imputation procedure could be used to estimate $\boldsymbol{Z}_2\boldsymbol{\alpha}_{\bar{\mathcal{C}}}$; RUV2, RUV3, and RUV4 are just three versions that rely heavily on linear associations between submatrices. Indeed, we need not even assume a factor model for the form of the unwanted variation. And we can further incorporate uncertainty in the estimates. In this section we develop these ideas to provide a more general framework for removing unwanted variation, which we call RUV*.

4.1 More general approaches to matrix imputation

To allow for more general approaches to matrix imputation we generalize (3.1) to

$$\begin{pmatrix} \boldsymbol{Y}_{2\mathcal{C}} & \boldsymbol{Y}_{2\bar{\mathcal{C}}} \\ \boldsymbol{Y}_{3\mathcal{C}} & \boldsymbol{Y}_{3\bar{\mathcal{C}}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Omega}(\boldsymbol{\phi})_{2\mathcal{C}} & \boldsymbol{\Omega}(\boldsymbol{\phi})_{2\bar{\mathcal{C}}} \\ \boldsymbol{\Omega}(\boldsymbol{\phi})_{3\mathcal{C}} & \boldsymbol{\Omega}(\boldsymbol{\phi})_{3\bar{\mathcal{C}}} \end{pmatrix} + \begin{pmatrix} \boldsymbol{0} & \boldsymbol{R}_{22}\boldsymbol{\beta}_2 \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} + \boldsymbol{E}, \quad (4.1)$$

where $\boldsymbol{\Omega}$ is the unwanted variation parameterized by some $\boldsymbol{\phi}$. When the unwanted variation is represented by a factor model, we have that $\boldsymbol{\phi} = \{\boldsymbol{Z}, \boldsymbol{\alpha}\}$ and $\boldsymbol{\Omega}(\boldsymbol{\phi}) = \boldsymbol{Z}\boldsymbol{\alpha}$.

The simplest version of RUV* fits this model in two steps:

1. Use any appropriate procedure to estimate $\boldsymbol{\Omega}_{2\bar{\mathcal{C}}}(\boldsymbol{\phi})$ given $\{\boldsymbol{Y}_{2\mathcal{C}}, \boldsymbol{Y}_{3\mathcal{C}}, \boldsymbol{Y}_{3\bar{\mathcal{C}}}\}$;
2. Estimate $\boldsymbol{\beta}_2$ by

$$\boldsymbol{R}_{22}^{-1}(\boldsymbol{Y}_{2\bar{\mathcal{C}}} - \boldsymbol{\Omega}_{2\bar{\mathcal{C}}}(\hat{\boldsymbol{\phi}})).$$

This idea is represented in the far right panel of Figure 2, and its relationship with other RUV approaches are illustrated in Supplementary Figure S1. Rather than restrict factors to be estimated via linear regression, RUV* allows any imputation procedure to be used to estimate $\boldsymbol{\Omega}_{2\bar{\mathcal{C}}}(\boldsymbol{\phi})$. This opens up a large literature on matrix imputation for use in removing unwanted variation with control genes (Hoff, 2007; Allen and Tibshirani, 2010; Candes and Plan, 2010; Stekhoven and Bühlmann, 2012; van Buuren, 2012; Josse et al., 2016, for example). (Note that RUV* is more general than RUVfun from Gagnon-Bartsch et al. (2013); Section S5 of the Supplementary Material.)

4.2 Incorporating uncertainty in estimated unwanted variation

Like previous RUV methods, the second step of RUV* treats the estimate of $\boldsymbol{\Omega}_{2\bar{\mathcal{C}}}(\boldsymbol{\phi})$ from the first step as if it were "known". Here we generalize this, using Bayesian ideas to propagate uncertainty.

Although the use of Bayesian methods in this context is not new (Stegle

et al., 2008, 2010; Fusi et al., 2012; Stegle et al., 2012), our development here shares one of the great advantages of the RUV methods: *modularity*. That is, RUV methods separate the analysis into smaller self-contained steps: the FA step and the regression step. Modularity is widely used in many fields: mathematicians modularize results using theorems, lemmas and corollaries; computer scientists modularize code using functions and classes. Modularity has many benefits, including: (1) it is easier to conceptualize an approach if it is broken into small simple steps, (2) it is easier to discover and correct mistakes, and (3) it is easier to improve an approach by improving specific steps. These advantages also apply to statistical analysis and methods development. For example, in RUV if one wishes to use a new method for FA then this does not require a whole new approach; one simply replaces the truncated SVD with the new FA.

To describe this generalized RUV* we introduce a latent variable $\tilde{\boldsymbol{Y}}_{2\bar{\mathcal{C}}}$ and write (4.1) as

$$\begin{pmatrix} \boldsymbol{Y}_{2\mathcal{C}} & \tilde{\boldsymbol{Y}}_{2\bar{\mathcal{C}}} \\ \boldsymbol{Y}_{3\mathcal{C}} & \boldsymbol{Y}_{3\bar{\mathcal{C}}} \end{pmatrix} = \boldsymbol{\Omega}(\boldsymbol{\phi}) + \boldsymbol{E}, \tag{4.2}$$

$$\boldsymbol{Y}_{2\bar{\mathcal{C}}} = \boldsymbol{R}_{22}\boldsymbol{\beta}_2 + \tilde{\boldsymbol{Y}}_{2\bar{\mathcal{C}}}. \tag{4.3}$$

Now consider the following two-step procedure:

1. Use any appropriate procedure to obtain a conditional distribution

$h(\tilde{\boldsymbol{Y}}_{2\bar{\mathcal{C}}}) = p(\tilde{\boldsymbol{Y}}_{2\bar{\mathcal{C}}}|\mathcal{Y}_m)$, where $\mathcal{Y}_m = \{\boldsymbol{Y}_{2\mathcal{C}}, \boldsymbol{Y}_{3\mathcal{C}}, \boldsymbol{Y}_{3\bar{\mathcal{C}}}\}$.

2. Perform inference for $\boldsymbol{\beta}_2$ using the likelihood

$$\begin{aligned} L(\boldsymbol{\beta}_2) &= p(\boldsymbol{Y}_{2\bar{\mathcal{C}}}, \mathcal{Y}_m|\boldsymbol{\beta}_2) \\ &= p(\mathcal{Y}_m) \int p(\boldsymbol{Y}_{2\bar{\mathcal{C}}}|\tilde{\boldsymbol{Y}}_{2\bar{\mathcal{C}}}, \boldsymbol{\beta}_2) p(\tilde{\boldsymbol{Y}}_{2\bar{\mathcal{C}}}|\mathcal{Y}_m)\, d\tilde{\boldsymbol{Y}}_{2\bar{\mathcal{C}}} \\ &\propto \int \delta(\boldsymbol{Y}_{2\bar{\mathcal{C}}} - \tilde{\boldsymbol{Y}}_{2\bar{\mathcal{C}}} - \boldsymbol{R}_{22}\boldsymbol{\beta}_2) p(\tilde{\boldsymbol{Y}}_{2\bar{\mathcal{C}}}|\mathcal{Y}_m)\, d\tilde{\boldsymbol{Y}}_{2\bar{\mathcal{C}}} \\ &= h(\boldsymbol{Y}_{2\bar{\mathcal{C}}} - \boldsymbol{R}_{22}\boldsymbol{\beta}_2) \end{aligned}$$

where $\delta(\cdot)$ indicates the Dirac delta function.

Of course, in step 2 one could do classical inference for $\boldsymbol{\beta}_2$, or place a prior on $\boldsymbol{\beta}_2$ and perform Bayesian inference.

This procedure requires an analytic form for the conditional distribution h. An alternative is to assume that we can sample from this conditional distribution, which yields a convenient sample-based (or "multiple imputation") RUV* algorithm.

1. Use any appropriate procedure to obtain samples $\tilde{\boldsymbol{Y}}^{(1)}_{2\bar{\mathcal{C}}}, \ldots, \tilde{\boldsymbol{Y}}^{(t)}_{2\bar{\mathcal{C}}}$ from a conditional distribution $p(\tilde{\boldsymbol{Y}}_{2\bar{\mathcal{C}}}|\mathcal{Y}_m)$.
2. Approximate the likelihood for $L(\boldsymbol{\beta}_2)$ by using the fact that $\hat{\boldsymbol{\beta}}^{(i)}_2 = \boldsymbol{R}^{-1}_{22}(\boldsymbol{Y}_{2\bar{\mathcal{C}}} - \tilde{\boldsymbol{Y}}^{(i)}_{2\bar{\mathcal{C}}})$ are sampled from a distribution proportional to $L(\boldsymbol{\beta}_2)$. (This distribution is guaranteed to be proper; Section S11 of the Supplementary Material.)

For example, in step 2 we can approximate the likelihood for each element of $\boldsymbol{\beta}_2$ by a normal likelihood

$$L(\beta_{2j}) \approx N(\beta_{2j}; \hat{\beta}_{2j}, \hat{s}_j^2), \tag{4.4}$$

where $\hat{\beta}_{2j}$ and $\hat{s}_j$ are respectively the mean and standard deviation of $\hat{\boldsymbol{\beta}}_2^{(i)}$. Alternatively, a t likelihood can be used. Either approach provides an estimate and standard error for each element of $\boldsymbol{\beta}_2$ that accounts for uncertainty in the estimated unwanted variation. (In contrast, the various methods used by other RUV approaches do not account for this uncertainty; Section S6 of the Supplementary Material.) Here we use these values to rank the "significance" of genes by the value of $\hat{\beta}_{2j}/\hat{s}_j$. They could also be used as inputs to the empirical Bayes method in Stephens (2017) to obtain measurements of significance related to false discovery rates.

Other approaches to inference in Step 2 are also possible. For example, given a specific prior on $\boldsymbol{\beta}_2$, Bayesian inference for $\boldsymbol{\beta}_2$ could be performed by re-weighting these samples according to this prior distribution (Section S8 of the Supplementary Material). This re-weighting yields an arbitrarily accurate approximation to the posterior distribution $p(\boldsymbol{\beta}_2|\mathcal{Y}_m, \boldsymbol{Y}_{2\bar{\mathcal{C}}})$ (Section S9 of the Supplementary Material). Posterior summaries using this re-weighting scheme are easy to derive (Section S12 of the Supplementary Material).

To illustrate the potential for RUV* to produce new methods for removing unwanted variation we implemented a version of RUV*, using a Markov chain Monte Carlo scheme to fit a simple Bayesian Factor analysis model, and hence perform the sampling-based imputation in Step 1 of RUV*. See Section S10 of the Supplementary Material for details. We refer to this method as RUVB.

5. Empirical evaluations

We now compare methods using simulations based on real data (GTEx Consortium, 2015). The simulation procedure is described in detail in Section S13 of the Supplementary Material. In brief, we use random subsets of real expression data to create "null data" that contains real (but unknown) "unwanted variation", and then modify these null data to add known signal. We varied the sample size ($n = 6, 10, 20, 40$), number of genes ($p = 1000$), number of control genes ($m = 10, 100$), and the proportion of null genes ($\pi_0 = 0.5, 0.9, 1$).

Being based on real data, these simulations involve realistic levels of unwanted variation. However, they also represent a "best-case" scenario in which treatment labels were randomized with respect to the factors causing this unwanted variation (see Section S16 of the Supplementary Material for

studying the effects of correlated confounding). They also represent a best case scenario in that the control genes given to each method are simulated to be genuinely null (See Section S17 of the Supplementary Material for studying the effects of misspecifying the negative controls). Even in this best-case scenario unwanted variation is a major issue, and, as we shall see, obtaining well calibrated inferences is challenging.

5.1 Summary of methods compared

We compared standard ordinary least squares regression (OLS) against five other approaches: RUV2, RUV3, RUV4, CATE (the GLS variant of RUV4), and RUVB. In preceding sections we have focused on how these methods obtain point estimates for $\boldsymbol{\beta}_2$. However in practice one also needs to find standard errors for these estimates. Just as there are many approaches to producing point estimates, there are also many approaches to producing standard errors. Key techniques used include "MAD variance calibration" (Wang et al., 2017), "control gene variance calibration" (Gagnon-Bartsch et al., 2013) and Empirical Bayes variance moderation (EBVM) (Smyth, 2004); see Section S6 of the Supplementary Material for more details. Our experience is that the choice of these techniques can greatly affect results, particularly calibration of interval estimates. We therefore experimented

with several approaches to standard error estimation for each method, and summarize results by presenting the best-performing version of each method. See Section S15 of the Supplementary Material for more extensive discussion.

For RUVB, we considered two approaches to producing mean and variance estimates: (i) Using sample-based posterior summaries (Section S12 of the Supplementary Material), and (ii) Using the normal approximation to the likelihood in Equation (4.4).

5.2 Comparisons: sensitivity vs specificity

We compare the power of methods to distinguish null and non-null genes by computing the area under the receiver operating characteristic curve (AUC) for each method as the significance threshold is varied.

The clearest result here is that all the methods consistently outperform standard OLS (Supplementary Figure S3). This emphasizes the benefits of removing unwanted variation in improving power to detect real effects. For small sample size comparisons (e.g. 3 vs 3) the gains are smaller, though still apparent, presumably because reliably estimating the unwanted variation is harder for small samples.

A second clear pattern is that the use of EBVM when estimating stan-

dard errors consistently improved AUC performance: the best-performing method in each family used EBVM. As might be expected, these benefits of EBVM are greatest for smaller sample sizes (Supplementary Figure S3).

Compared with these two clear patterns, differences among the best-performing methods in each family are more subtle. Figure 3a compares the AUC of the best method in each family with that of RUVB, which performed best overall in this comparison. (Results are shown for $\pi_0 = 0.5$; results for $\pi_0 = 0.9$ are similar). We highlight four main results:

1. RUVB has the best mean AUC among all methods we explored;
2. RUV4/CATE methods perform less well (relative to RUVB) when there are few control genes and the sample size is large;
3. In contrast, RUV2 methods perform less well (relative to RUVB) when the sample size is small and there are few control genes;
4. RUV3 performs somewhat stably (relative to RUVB) across the sample sizes.

The mean AUCs for RUVB are in Supplementary Figure S2.

5.3 Comparisons: calibration

We also assessed the calibration of methods by examining the empirical coverage of their nominal 95% confidence intervals for each effect (based on

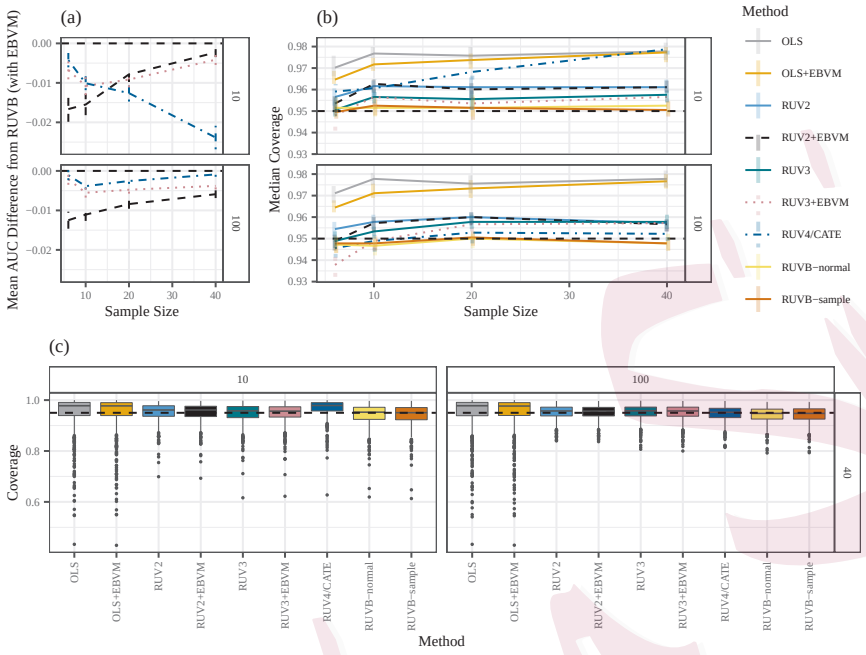

Figure 3: (a) Comparison of AUC achieved by best-performing method in each family vs RUVB. Each point shows the observed mean difference in AUC, with vertical lines indicating 95% confidence intervals for the mean. Results are shown for $\pi_0 = 0.5$ with 10 control genes (upper facet) or 100 control genes (lower facet). All results are below zero (the dashed horizontal line), indicating superior performance of RUVB. (b) Median coverage for the best performing methods' 95% confidence intervals when $\pi_0 = 0.5$. The vertical lines are bootstrap 95% confidence intervals for the median coverage, made transparent and slightly horizontally dodged to increase clarity. The horizontal dashed line is at 0.95. (c) Boxplots of coverage for the best performing methods' 95% confidence intervals when $\pi_0 = 0.5$ and $n = 40$. For both (b) and (c) the left and right facets show results for 10 and 100 control genes respectively.

standard theory for the relevant t distribution in each case).

We begin by examining "typical" coverage for each method in each scenario by computing the median (across datasets) of the empirical coverage. We found that, without variance calibration, all method families except RUV4/CATE could achieve satisfactory typical coverage (somewhere between 0.94 and 0.97) across all scenarios (Figure 3b) shows results for $\pi_0 = 0.5$; other values yielded similar results, not shown). The best performing RUV4/CATE method was often overly conservative in scenarios with few control genes, especially with larger sample sizes.

Although these median coverage results are encouraging, in practice having small variation in coverage among datasets is also important. That is, we would like methods to have near-95% coverage in most data-sets, and not only on average. Here the results (Figure 3c; Supplementary Figure S4) are less encouraging: coverage of methods with good typical coverage (median coverage close to 95%) varied considerably among datasets. This said, variability does improve for larger sample sizes and more control genes, and in this case all methods improve noticeably on OLS (Figure 3c, right facet). A particular concern is that, across all these methods, for many datasets, empirical coverage can be much lower than the nominal goal of 95%. Such datasets might be expected to lead to problems with over-identification

of significant null genes ("false positives"), and under-estimation of false discovery rates.

To summarize variability in coverage — as well as any tendency to be conservative or anti-conservative — we calculated the proportion of datasets where the actual coverage deviated substantially from 95%, which we defined as being either less than 90% or greater than 97.5%. Figure 4 shows the mean proportions for each method (where the mean was taken over the methods that use each type of variance calibration technique).The key findings are:

1. RUVB (the normal and sample-based versions) has "balanced" errors in coverage: its empirical coverage is as likely to be too high as too low.

2. MAD calibration tends to produce highly conservative coverage — that is, its coverage is very often much larger than the claimed 95%, and seldom much lower. This will tend to reduce false positive significant results, but also substantially reduce power to detect real effects. The exception is that when all genes are null ($\pi_0 = 1$), MAD calibration works well for larger sample sizes. These results are likely explained partly by non-null genes biasing upwards the variance calibration parameter, an issue also noted in Sun et al. (2012).

3. Control-gene calibration is often anti-conservative when there are few control genes. However, it can work well when the sample size is large and there are many control genes. Interestingly, with few control genes the anti-conservative behavior gets worse as sample size increases.

5.4 Additional Simulations

As mentioned earlier, the simulation results in Sections 5.2 and 5.3 are from a best-case scenario where treatment labels are randomized for each individual. To study the effects of correlated confounding, we extended our simulation approach to allow for treatment labels to be correlated with latent factors (Section S14 of the Supplementary Material). Our results, presented in Section S16 of the Supplementary Material, indicate that RUVB and RUV3 remain competitive in the presence of correlated confounders.

More insidious is the result of misspecifying the negative controls. We study, in Section S17 of the Supplementary Material, the effects of misspecifying the negative controls. Our results indicate that RUVB and RUV2 are very sensitive to the negative controls assumption, while RUV3 and RUV4 are relatively robust to the negative controls assumption (an anonymous reviewer suggested that this might be a result of the regression steps in

(2.7) and (3.2)). Thus, the use of RUVB (as well as RUV2) should only occur when one has high quality negative controls.

Software

The methods developed in this paper are implemented in the R package `vicar` available at `https://github.com/dcgerard/vicar`. Code and instructions for reproducing the empirical evaluations in Section 5 are available at `https://github.com/dcgerard/ruvb_sims`.

Supplementary Material

The online supplementary material contains proofs, additional theoretical and simulation details, and additional simulation results.

Acknowledgements

Some of the original code for the simulated dataset generation was based on implementations by Mengyin Lu, to whom we are indebted. This work was supported by NIH grant HG002585 and by a grant from the Gordon and Betty Moore Foundation (Grant GBMF #4559).

References

Akey, J. M., S. Biswas, J. T. Leek, and J. D. Storey (2007). On the design and analysis of gene expression studies in human populations. *Nature genetics 39*(7), 807–809.

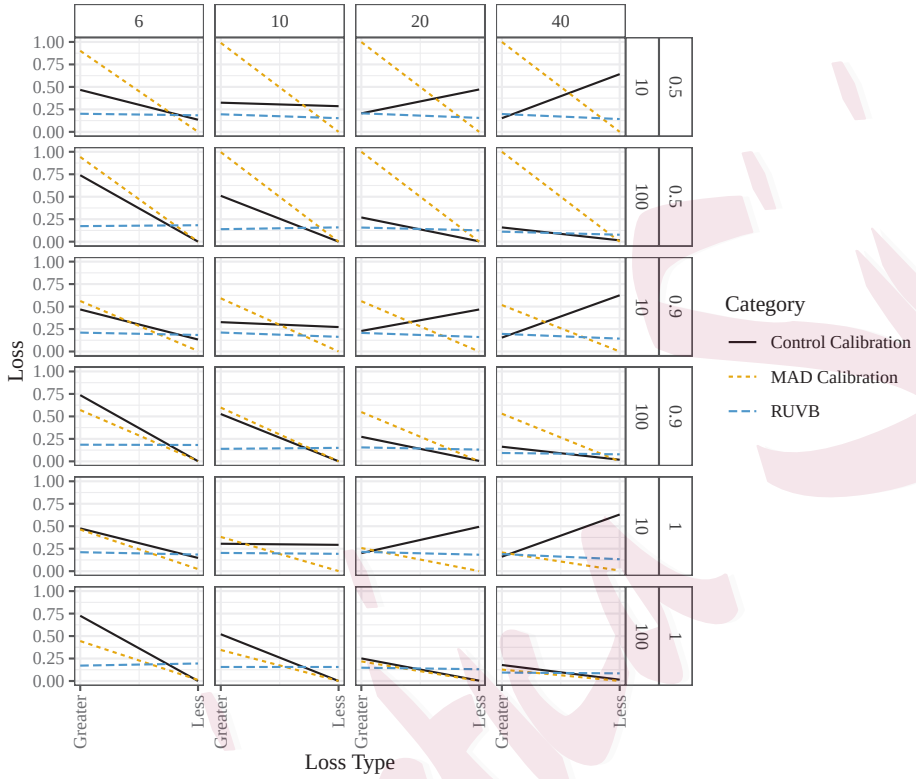

Figure 4: Mean proportion of times the coverage was either greater than 0.975 (Greater) or less than 0.9 (Less). The column facets distinguish between sample sizes while the row facets distinguish between the number of control genes and the proportion of genes that are null. The means were taken over the variance calibration method: MAD calibrated (S6.3), control-gene calibrated (S6.1), or the sample-based or normal-based RUVB approach.

Allen, G. I. and R. Tibshirani (2010). Transposable regularized covariance models with an application to missing data imputation. *The Annals of Applied Statistics 4*(2), 764–790.

Behzadi, Y., K. Restom, J. Liau, and T. T. Liu (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *Neuroimage 37*(1), 90–101.

Candes, E. J. and Y. Plan (2010). Matrix completion with noise. *Proceedings of the IEEE 98*(6), 925–936.

Carvalho, C. M., J. Chang, J. E. Lucas, J. R. Nevins, Q. Wang, and M. West (2008). High-dimensional sparse factor modeling: Applications in gene expression genomics. *Journal of the American Statistical Association 103*(484), 1438–1456. PMID: 21218139.

Chen, M. and X. Zhou (2017). Controlling for confounding effects in single cell RNA sequencing studies using both control and target genes. *Scientific reports 7*(1), 13587.

Cochran, W. G. (1943). The comparison of different scales of measurement for experimental results. *Ann. Math. Statist. 14*(3), 205–216.

Efron, B. (2004). Large-scale simultaneous hypothesis testing. *Journal of the American Statistical Association 99*(465), 96–104.

Efron, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statistical science 23*(1), 1–22.

Efron, B. (2010). Correlated z-values and the accuracy of large-scale statistical estimates. *Journal of the American Statistical Association 105*(491), 1042–1055.

Efron, B. and C. Morris (1972). Empirical Bayes on vector observations: An extension of Stein's method. *Biometrika 59*(2), 335.

Eisenberg, E. and E. Y. Levanon (2013). Human housekeeping genes, revisited. *Trends in Genetics 29*(10), 569–574.

Fisher, R. A. and W. A. Mackenzie (1923). Studies in crop variation. ii. the manurial response of different potato varieties. *The Journal of Agricultural Science 13*(3), 311–320.

Freeman, G. H. (1973). Statistical methods for the analysis of genotype-environment interactions. *Heredity 31*(3), 339–354.

Friguet, C., M. Kloareg, and D. Causeur (2009). A factor model approach to multiple testing under dependence. *Journal of the American Statistical Association 104*(488), 1406–1415.

Fusi, N., O. Stegle, and N. D. Lawrence (2012). Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies. *PLoS Comput Biol 8*(1), e1002330.

Gabriel, K. R. (1978). Least squares approximation of matrices by additive and multiplicative models. *Journal of the Royal Statistical Society. Series B (Methodological) 40*(2), 186–196.

Gagnon-Bartsch, J. (2015). *ruv: Detect and Remove Unwanted Variation using Negative Controls*. R package version 0.9.6.

Gagnon-Bartsch, J., L. Jacob, and T. Speed (2013). Removing unwanted variation from high dimensional data with negative controls. Technical report, Technical Report 820, Depart-

ment of Statistics, University of California, Berkeley.

Gagnon-Bartsch, J. A. and T. P. Speed (2012). Using control genes to correct for unwanted variation in microarray data. *Biostatistics 13*(3), 539–552.

Gerard, D. and M. Stephens (2018). Empirical Bayes shrinkage and false discovery rate estimation, allowing for unwanted variation. *Biostatistics*, kxy029.

Gibson, G. (2008). The environmental contribution to gene expression profiles. *Nature Reviews Genetics 9*(8), 575–581.

Gilad, Y. and O. Mizrahi-Man (2015). A reanalysis of mouse encode comparative gene expression data. *F1000Research 4*.

Gollob, H. F. (1968). A statistical model which combines features of factor analytic and analysis of variance techniques. *Psychometrika 33*(1), 73–115.

GTEx Consortium (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science 348*(6235), 648–660.

Hoff, P. D. (2007). Model averaging and dimension selection for the singular value decomposition. *J. Amer. Statist. Assoc. 102*(478), 674–685.

Irizarry, R. A., D. Warren, F. Spencer, I. F. Kim, S. Biswal, B. C. Frank, E. Gabrielson, J. G. N. Garcia, J. Geoghegan, G. Germino, C. Griffin, S. C. Hilmer, E. Hoffman, A. E. Jedlicka, E. Kawasaki, F. Martínez-Murillo, L. Morsberger, H. Lee, D. Petersen, J. Quackenbush, A. Scott, M. Wilson, Y. Yang, S. Q. Ye, and W. Yu (2005). Multiple-laboratory comparison

of microarray platforms. *Nature methods 2*(5), 345–350.

Jiang, L., F. Schlesinger, C. A. Davis, Y. Zhang, R. Li, M. Salit, T. R. Gingeras, and B. Oliver (2011). Synthetic spike-in standards for RNA-seq experiments. *Genome research 21*(9), 1543–1551.

Johnson, W. E., C. Li, and A. Rabinovic (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics 8*(1), 118–127.

Josse, J., S. Sardy, and S. Wager (2016). denoiseR: A package for low rank matrix estimation. *arXiv preprint arXiv:1602.01206*.

Kang, H. M., J. H. Sul, S. K. Service, N. A. Zaitlen, S.-y. Kong, N. B. Freimer, C. Sabatti, and E. Eskin (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature genetics 42*(4), 348–354.

Kang, H. M., C. Ye, and E. Eskin (2008). Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics 180*(4), 1909–1925.

Kang, H. M., N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman, M. J. Daly, and E. Eskin (2008). Efficient control of population structure in model organism association mapping. *Genetics 178*(3), 1709–1723.

Lee, S., W. Sun, F. A. Wright, and F. Zou (2017). An improved and explicit surrogate variable analysis procedure by coefficient adjustment. *Biometrika 104*(2), 303–316.

Leek, J. T., R. B. Scharpf, H. C. Bravo, D. Simcha, B. Langmead, W. E. Johnson, D. Geman, K. Baggerly, and R. A. Irizarry (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics 11*(10), 733–739.

Leek, J. T. and J. D. Storey (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics 3*(9), 1724–1735.

Leek, J. T. and J. D. Storey (2008). A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences 105*(48), 18718–18723.

Listgarten, J., C. Kadie, E. E. Schadt, and D. Heckerman (2010). Correction for hidden confounders in the genetic analysis of gene expression. *Proceedings of the National Academy of Sciences 107*(38), 16465–16470.

Lucas, J., C. Carvalho, Q. Wang, A. Bild, J. Nevins, and M. West (2006). Sparse statistical modelling in gene expression genomics. In K.-A. Do, P. Müller, and M. Vannucci (Eds.), *Bayesian inference for gene expression and proteomics*, pp. 155–176. Cambridge University Press.

Mandel, J. (1969). The partitioning of interaction in analysis of variance. *Journal of Research of the National Bureau of Standards-B. Mathematical Sciences 73B*(4), 309–328.

Mandel, J. (1971). A new analysis of variance model for non-additive data. *Technometrics 13*(1), 1–18.

McKennan, C. and D. Nicolae (2018a). Accounting for unobserved covariates with varying degrees of estimability in high dimensional biological data. *ArXiv e-prints*.

McKennan, C. and D. Nicolae (2018b). Estimating and accounting for unobserved covariates in high dimensional correlated data. *ArXiv e-prints*.

Mostafavi, S., A. Battle, X. Zhu, A. E. Urban, D. Levinson, S. B. Montgomery, and D. Koller (2013). Normalizing RNA-sequencing data by modeling hidden covariates with prior knowledge. *PLOS ONE 8*(7), 1–10.

Owen, A. B. and P. O. Perry (2009). Bi-cross-validation of the SVD and the nonnegative matrix factorization. *Ann. Appl. Stat. 3*(2), 564–594.

Owen, A. B. and J. Wang (2016). Bi-cross-validation for factor analysis. *Statist. Sci. 31*(1), 119–139.

Perry, P. O. and N. S. Pillai (2013). Degrees of freedom for combining regression with factor analysis. *arXiv preprint arXiv:1310.7269*.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics 38*(8), 904–909.

Schwartzman, A. (2010). Comment. *Journal of the American Statistical Association 105*(491), 1059–1063.

Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology 3*(1).

Stegle, O., A. Kannan, R. Durbin, and J. Winn (2008). Accounting for non-genetic factors improves the power of eQTL studies. In *Research in Computational Molecular Biology*, pp. 411–422. Springer.

Stegle, O., L. Parts, R. Durbin, and J. Winn (2010). A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput Biol 6*(5), e1000770.

Stegle, O., L. Parts, M. Piipari, J. Winn, and R. Durbin (2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature protocols 7*(3), 500–507.

Stekhoven, D. J. and P. Bühlmann (2012). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics 28*(1), 112–118.

Stephens, M. (2017). False discovery rates: a new deal. *Biostatistics 18*(2), 275–294.

Sun, Y., N. R. Zhang, and A. B. Owen (2012). Multiple hypothesis testing adjusted for latent variables, with an application to the AGEMAP gene expression data. *Ann. Appl. Stat. 6*(4), 1664–1688.

Tukey, J. W. (1962). The future of data analysis. *Ann. Math. Statist. 33*(1), 1–67.

van Buuren, S. (2012). *Flexible Imputation of Missing Data*. Chapman and Hall/CRC.

Wang, J., Q. Zhao, T. Hastie, and A. B. Owen (2017). Confounder adjustment in multiple hypothesis testing. *Ann. Statist. 45*(5), 1863–1894.

Williams, E. J. (1952). The interpretation of interactions in factorial experiments. *Biometrika 39*(1/2), 65–81.

Wu, Z. and M. J. Aryee (2010). Subset quantile normalization using negative control features. *Journal of Computational Biology 17*(10), 1385–1395.

Yang, C., L. Wang, S. Zhang, and H. Zhao (2013). Accounting for non-genetic factors by low-rank representation and sparse regression for eQTL mapping. *Bioinformatics 29*(8), 1026–1034.

Department of Mathematics and Statistics, American University, Washington DC, USA

E-mail: (dgerard@american.edu)

Departments of Human Genetics and Statistics, University of Chicago, Chicago IL, USA

E-mail: (mstephens@uchicago.edu)