Biometrika (2019), **106**, 4, *pp.* 823–840
Printed in Great Britain

doi: 10.1093/biomet/asz037
Advance Access publication 16 September 2019

Accounting for unobserved covariates with varying degrees of estimability in high-dimensional biological data

By CHRIS McKENNAN and DAN NICOLAE

Department of Statistics, University of Chicago, 5747 S. Ellis Avenue, Chicago,
Illinois 60637, U.S.A.

cgm29@galton.uchicago.edu nicolae@galton.uchicago.edu

Summary

An important phenomenon in high-throughput biological data is the presence of unobserved covariates that can have a significant impact on the measured response. When these covariates are also correlated with the covariate of interest, ignoring or improperly estimating them can lead to inaccurate estimates of and spurious inference on the corresponding coefficients of interest in a multivariate linear model. We first prove that existing methods to account for these unobserved covariates often inflate Type I error for the null hypothesis that a given coefficient of interest is zero. We then provide alternative estimators for the coefficients of interest that correct the inflation, and prove that our estimators are asymptotically equivalent to the ordinary least squares estimators obtained when every covariate is observed. Lastly, we use previously published DNA methylation data to show that our method can more accurately estimate the direct effect of asthma on DNA methylation levels compared to existing methods, the latter of which likely fail to recover and account for latent cell type heterogeneity.

Some key words: Batch effect; Cell type heterogeneity; Confounding; High-dimensional factor analysis; Unobserved covariates; Unwanted variation.

1. Introduction

High-throughput genetic, DNA methylation, metabolomic and proteomic data are often influenced by unobserved covariates that are difficult or impossible to record (Johnson et al., 2007; Leek et al., 2010; Houseman et al., 2012). Suppose we observe data $Y \in \mathbb{R}^{p\times n}$, where the number of genomic units, p, is on the order of or larger than the sample size, n. For example, in most DNA methylation data, the number of studied methylation sites, p, is between 10^4 and 10^6 and $n = O\left(10^2\right)$. Assume the true model for Y is

$$Y_{p\times n} = B_{p\times d}X_{n\times d}^{\mathrm{T}} + L_{p\times K}C_{n\times K}^{\mathrm{T}} + \mathcal{E}_{p\times n},$$
$$\mathcal{E}_{p\times n} \sim MN_{p\times n}\left(0, \Sigma_{p\times p}, I_n\right), \quad \Sigma = \mathrm{diag}\left(\sigma_1^2, \ldots, \sigma_p^2\right), \tag{1}$$

where $B = \left(\beta_1 \cdots \beta_p\right)^{\mathrm{T}} \in \mathbb{R}^{p\times d}$, $X \in \mathbb{R}^{n\times d}$ contains the covariates of interest and $C \in \mathbb{R}^{n\times K}$ contains the K unobserved covariates. Our goal is to estimate and perform inference on the coefficients of interest, $\beta_1, \ldots, \beta_p \in \mathbb{R}^d$.

Under model (1), the naive ordinary least squares estimate of B,

$$Y_1 = YX\left(X^{\mathrm{T}}X\right)^{-1} = B + L\{\left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}C\}^{\mathrm{T}} + \mathcal{E}X\left(X^{\mathrm{T}}X\right)^{-1} = B + L\Omega^{\mathrm{T}} + \mathcal{E}_1,$$

is biased by $L\Omega^{\mathrm{T}}$, where $\Omega = \left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}C$ is the ordinary least squares coefficient estimate for the regression of C on to X. The bias induced by L and Ω is often consequential in biological data. For example, in DNA methylation studies where disease status is the covariate of interest, DNA methylation Y depends on the latent cellular heterogeneity of the n samples (Jaffe & Irizarry, 2014), and cellular heterogeneity often depends on disease status X (Fahy, 2002; Stein et al., 2016). Ignoring unobserved covariates C when analysing these types of data can therefore drastically affect the interpretation of results.

There have been a number of methods proposed to estimate and correct for the latent factors C in model (1) (Leek & Storey, 2008; Gagnon-Bartsch & Speed, 2012; Sun et al., 2012; Gagnon-Bartsch et al., 2013; Houseman et al., 2014; Lee et al., 2017). While these methods perform well on selected datasets, they either do not have the requisite theory to justify downstream inference on $\beta_1, \ldots, \beta_p$ (Leek & Storey, 2008; Sun et al., 2012; Houseman et al., 2014; Lee et al., 2017) or they require the practitioner to have prior knowledge regarding which coefficients $\beta_1, \ldots, \beta_p$ are zero (Gagnon-Bartsch & Speed, 2012; Gagnon-Bartsch et al., 2013).

Recently, Fan & Han (2017) and Wang et al. (2017) proposed methods that first compute $\hat{L}$, an estimate of L, from $YP_X^{\perp} = L\left(P_X^{\perp}C\right)^{\mathrm{T}} + \mathcal{E}P_X^{\perp}$, where $P_X^{\perp} \in \mathbb{R}^{n\times n}$ is the orthogonal projection matrix on to the orthogonal complement of X. They then estimate Ω by regressing Y_1 on to $\hat{L}$, and finally estimate $\beta_1, \ldots, \beta_p$ by subtracting the estimated bias $\hat{L}\hat{\Omega}^{\mathrm{T}}$ from Y_1. The advantage of this estimation paradigm is obvious: it decouples the estimation of L and $\beta_1, \ldots, \beta_p$ without requiring the practitioner to have prior knowledge regarding which coefficients $\beta_1, \ldots, \beta_p$ are zero. These articles are quite remarkable because, when their assumptions hold, the authors prove that they can perform inference on $\beta_1, \ldots, \beta_p$ that is as accurate as when C is known. However, it has been observed that these methods tend to inflate test statistics and cause anticonservative inference in both simulated and real data (van Iterson et al., 2017).

One source of the discrepancy between theory and practice is that the aforementioned articles assume that all K of the nonzero eigenvalues of $\mathcal{I} = P_X^{\perp}C\left(p^{-1}L^{\mathrm{T}}L\right)C^{\mathrm{T}}P_X^{\perp}$ are on the order of the number of samples, n, and are overtly larger than the average residual variance $p^{-1}\left(\sigma_1^2 + \cdots + \sigma_p^2\right)$. If these assumptions were valid, there would be an unambiguous gap between the Kth and $[K+1]$th eigenvalues of $p^{-1}P_X^{\perp}Y^{\mathrm{T}}YP_X^{\perp}$. However, this rarely occurs in practice (Cangelosi & Goriely, 2007; Owen & Wang, 2016; Wang et al., 2017). When these eigenvalue assumptions are violated, we show that previous methods' techniques to estimate Ω from the regression of Y_1 onto $\hat{L}$ are sensitive to the error in the estimated design matrix $\hat{L}$, which causes inaccurate estimates of $\beta_1, \ldots, \beta_p$. In practice, some of the nonzero eigenvalues of $\mathcal{I}$ will not be large if the sample size is not sufficiently large, if some of the K latent covariates do not influence the response of every genomic unit, or if some of the latent covariates are correlated with the covariate of interest X, since this will dampen $P_X^{\perp}C$. The latter is common in DNA methylation data because unobserved cellular heterogeneity is often correlated with X (Jaffe & Irizarry, 2014).

The purpose of this article is to fill the described gap in the literature by studying the unobserved covariate problem when some or all of the K nonzero eigenvalues of $\mathcal{I}$ are not exceedingly large. We prove that when the eigenvalues fall below a certain threshold, then for fixed $g \in \{1, \ldots, p\}$, previous methods have a propensity to inflate Type I error when testing the null hypothesis $H_{0,g} : \beta_g = 0$, and even tend to falsely reject $H_{0,g}$ when using the conservative Bonferroni correction. We then provide alternative estimators for $\beta_1, \ldots, \beta_p$ and prove that when B is suitably sparse, our estimators are asymptotically equivalent to the ordinary least squares estimators obtained using the design matrix $(X\ C)$, regardless of the size of the eigenvalues of $\mathcal{I}$. We lastly use simulated data and real DNA methylation data from Nicodemus-Johnson et al. (2016) to show that latent covariates with ostensibly small effects can be detrimental to inference if not properly accounted for, and that our method can better account for latent covariates than the leading competitors.

2. The model, our estimation procedure and intuition

2.1. *Notation*

For any integer $n \geqslant 1$, we define $[n] = \{1, \ldots, n\}$. For any matrix $G \in \mathbb{R}^{n\times m}$, we define $P_G \in \mathbb{R}^{n\times n}$ and $P_G^{\perp} \in \mathbb{R}^{n\times n}$ to be the orthogonal projection matrices that project vectors on to the image of G and the orthogonal complement of G, respectively, and $G_{r*} \in \mathbb{R}^m$, $G_{*s} \in \mathbb{R}^n$ and $G_{rs} \in \mathbb{R}$ to be the rth row, sth column and (r, s) element of G. Lastly, we define $1_n, 0_n \in \mathbb{R}^n$ to be the vectors of all ones and all zeros and use the notation $W \overset{\mathcal{D}}{=} Z$ if the random variables, or matrices, W and Z have the same distribution.

2.2. *A model for the data*

Let $Y \in \mathbb{R}^{p\times n}$ be the observed data, where Y_{gi} is an observation at genomic unit $g \in [p]$ in sample $i \in [n]$. Let $X \in \mathbb{R}^{n\times d}$ be an observed, full rank matrix containing the covariates of interest and define $B = \left(\beta_1 \cdots \beta_p\right)^{\mathrm{T}} \in \mathbb{R}^{p\times d}$ to be their corresponding coefficients across all p genomic units. We also define an additional covariate matrix $C \in \mathbb{R}^{n\times K}$ and let $L \in \mathbb{R}^{p\times K}$ be its corresponding coefficient. We assume that C is unobserved, but K is known. Evidently, K is rarely known in true data applications. While we acknowledge that estimating K is a challenging problem, there is a large body of work devoted to estimating it (Leek & Storey, 2008; Onatski, 2010; Gagnon-Bartsch & Speed, 2012; Owen & Wang, 2016; McKennan & Nicolae, 2018). We discuss how different values of K affect our downstream estimates in § 4. We assume (1) is the true model for Y, and we define

$$\rho = p^{-1}\left(\sigma_1^2 + \cdots + \sigma_p^2\right). \tag{2}$$

We also define

$$\Omega = \left(X^{\mathrm{T}}X\right)^{-1} X^{\mathrm{T}}C \in \mathbb{R}^{d\times K}, \qquad C_2 = P_X^{\perp}C \in \mathbb{R}^{n\times K} \tag{3}$$

to be the ordinary least squares coefficient estimates and residuals from the regression of C on to X, respectively. We have not assumed an explicit relationship between C and X, because one can always decompose C as

$$C = P_X C + P_X^{\perp} C = X\Omega + C_2.$$

A more general model for Y would be $Y = BX^{\mathrm{T}} + MZ^{\mathrm{T}} + LC^{\mathrm{T}} + \mathcal{E}$, where $Z \in \mathbb{R}^{n\times r}$ contains observed nuisance covariates, like the intercept or technical covariates, whose coefficients $M_{1*}, \ldots, M_{p*}$ are not of interest. We can get back to model (1) by multiplying Y on the right by a matrix whose columns form an orthonormal basis for the null space of Z^{T}. Therefore, we work exclusively with model (1) and assume any observed nuisance factors have already been rotated out, as they would be in ordinary least squares.

2.3. *Estimating B when C is unobserved*

We break Y into two independent pieces using a technique proposed in Sun et al. (2012):

$$Y_1 = YX\left(X^{\mathrm{T}}X\right)^{-1} = B + L\Omega^{\mathrm{T}} + \mathcal{E}_1, \tag{4}$$

$$Y_2 = YP_X^{\perp} = LC_2^{\mathrm{T}} + \mathcal{E}_2, \tag{5}$$

where $\mathcal{E}_1 = \mathcal{E}X\left(X^{\mathrm{T}}X\right)^{-1}$ and $\mathcal{E}_2 = \mathcal{E}P_X^{\perp}$ are independent because $\mathcal{E} \sim MN_{p\times n}\left(0, \Sigma, I_n\right)$ and $X^{\mathrm{T}}P_X^{\perp} = 0$. The matrix Y_1 is the ordinary least squares estimate of B that ignores C, and the rows of $\mathcal{E}_2$ lie on an $(n-d)$-dimensional linear subspace of $\mathbb{R}^n$. We now describe how to use Y_1 and Y_2 to derive the ordinary least squares estimates of $\beta_1, \ldots, \beta_p$ when C is observed. This will provide a template for estimating $\beta_1, \ldots, \beta_p$ when C is unobserved.

Algorithm 1 (Ordinary least squares when C is observed). Let $Y_1 \in \mathbb{R}^{p\times d}$, $Y_2 \in \mathbb{R}^{p\times n}$, $X \in \mathbb{R}^{n\times d}$ and $C \in \mathbb{R}^{n\times K}$ be given. Our goal is to use ordinary least squares to estimate and perform inference on $\beta_1, \ldots, \beta_p$, the rows of $B \in \mathbb{R}^{p\times d}$.

(a) Set $C_2 = P_X^{\perp}C$. Use Y_2 to estimate $\Sigma = \mathrm{diag}\left(\sigma_1^2, \ldots, \sigma_p^2\right)$ and L as

$$\left(\hat{\sigma}_g^{\mathrm{OLS}}\right)^2 = (n-d-K)^{-1}Y_{2_{g*}}^{\mathrm{T}}P_{C_2}^{\perp}Y_{2_{g*}} \quad (g = 1, \ldots, p),$$
$$\hat{L}^{\mathrm{OLS}} = Y_2C_2\left(C_2^{\mathrm{T}}C_2\right)^{-1},$$

where $Y_{2_{g*}} \in \mathbb{R}^n$ is the gth row of Y_2.

(b) Set $\Omega = \left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}C$.

(c) Define the ordinary least squares estimate of β_g to be

$$\hat{\beta}_g^{\mathrm{OLS}} = Y_{1_{g*}} - \Omega\hat{L}_{g*}^{\mathrm{OLS}} \quad (g = 1, \ldots, p), \tag{6}$$

where $Y_{1_{g*}} \in \mathbb{R}^d$ and $\hat{L}_{g*}^{\mathrm{OLS}} \in \mathbb{R}^K$ are the gth rows of Y_1 and $\hat{L}^{\mathrm{OLS}}$, respectively.

It is straightforward to derive the asymptotic properties of the estimators defined in Algorithm 1. In Step (a), $\hat{\sigma}_g^{\mathrm{OLS}} = \sigma_g + o_{\mathrm{p}}(1)$ as $n \to \infty$ and $\hat{L}^{\mathrm{OLS}} \sim MN_{p\times K}\left\{L, \Sigma, \left(C_2^{\mathrm{T}}C_2\right)^{-1}\right\}$. Since $\mathcal{E}_2$ is independent of $\mathcal{E}_1$, both of these estimates are independent of Y_1. This implies that the asymptotic distribution of $\hat{\beta}_g^{\mathrm{OLS}}$ is

$$\left(\hat{\sigma}_g^{\mathrm{OLS}}\right)^{-1}\left\{\left(X^{\mathrm{T}}X\right)^{-1} + \Omega\left(C_2^{\mathrm{T}}C_2\right)^{-1}\Omega^{\mathrm{T}}\right\}^{-1/2}\left(\hat{\beta}_g^{\mathrm{OLS}} - \beta_g\right) \overset{\mathcal{D}}{=} Z + o_{\mathrm{p}}(1) \quad (g = 1, \ldots, p)$$

as $n \to \infty$, where $Z \sim N_d\left(0, I_d\right)$.

A property of the ordinary least squares estimate $\hat{\beta}_g^{\mathrm{OLS}}$ is

$$\hat{\beta}_g^{\mathrm{OLS}} = Y_{1_{g*}} - \Omega\hat{L}_{g*}^{\mathrm{OLS}} = \left(X^{\mathrm{T}}P_C^{\perp}X\right)^{-1}X^{\mathrm{T}}P_C^{\perp}Y_{g*} \quad (g = 1, \ldots, p).$$

That is, $\hat{\beta}_g^{\mathrm{OLS}}$ depends only on the column space C, meaning we may replace C with $C\Psi$ as input in Algorithm 1 for any invertible matrix $\Psi \in \mathbb{R}^{K\times K}$. In particular, we may choose Ψ so that $n^{-1}C_2^{\mathrm{T}}C_2 = n^{-1}C^{\mathrm{T}}P_X^{\perp}C = I_K$. This parametrization of C, and therefore C_2, is convenient because it suggests that a reasonable estimate of C_2 when C is unobserved is a scalar multiple of the first K right singular vectors of Y_2. Using this intuition, we now present our method to estimate and perform inference on $\beta_1, \ldots, \beta_p$ when C is unobserved. This is described in Algorithm 2, which mimics the three steps of Algorithm 1.

Algorithm 2 (Estimation and inference when C is unobserved). Let $Y_1 \in \mathbb{R}^{p\times d}$, $Y_2 \in \mathbb{R}^{p\times n}$, $X \in \mathbb{R}^{n\times d}$ and $K \geqslant 1$ be given. Our goal is to estimate and perform inference on $\beta_1, \ldots, \beta_p$, the rows of $B \in \mathbb{R}^{p\times d}$.

(a) Let $Y_2 = U\,\mathrm{diag}\,(\tau_1, \ldots, \tau_n)\,V^{\mathrm{T}}$ be the singular value decomposition of Y_2 where $\tau_1 \geqslant \cdots \geqslant \tau_n \geqslant 0$ and $U^{\mathrm{T}}U = V^{\mathrm{T}}V = I_n$. Define $\hat{C}_2 = n^{1/2}\,(V_{*1} \cdots V_{*K})$, where V_{*k} is the kth column of $V \in \mathbb{R}^{n\times n}$. Estimate $\Sigma = \mathrm{diag}\left(\sigma_1^2, \ldots, \sigma_p^2\right)$ and L as

$$\hat{\sigma}_g^2 = (n-d-K)^{-1}\, Y_{2_{g*}}^{\mathrm{T}} P_{\hat{C}_2}^{\perp} Y_{2_{g*}} \quad (g = 1, \ldots, p), \tag{7}$$

$$\hat{L} = Y_2\hat{C}_2\left(\hat{C}_2^{\mathrm{T}}\hat{C}_2\right)^{-1}. \tag{8}$$

(b) Define $\hat{\rho} = p^{-1}\left(\hat{\sigma}_1^2 + \cdots + \hat{\sigma}_p^2\right)$ and

$$\hat{\lambda}_k = np^{-1}\hat{L}_{*k}^{\mathrm{T}}\hat{L}_{*k} \quad (k = 1, \ldots, K), \tag{9}$$

where $\hat{L}_{*k}$ is the kth column of $\hat{L} \in \mathbb{R}^{p\times K}$. Estimate Ω as

$$\hat{\Omega} = Y_1^{\mathrm{T}}\hat{L}\left(\hat{L}^{\mathrm{T}}\hat{L}\right)^{-1}\mathrm{diag}\left\{\hat{\lambda}_1/\left(\hat{\lambda}_1 - \hat{\rho}\right), \ldots, \hat{\lambda}_K/\left(\hat{\lambda}_K - \hat{\rho}\right)\right\}. \tag{10}$$

(c) Estimate β_g as

$$\hat{\beta}_g = Y_{1_{g*}} - \hat{\Omega}\hat{L}_{g*} \quad (g = 1, \ldots, p). \tag{11}$$

Just like the estimates $\left(\hat{\sigma}_g^{\mathrm{OLS}}\right)^2$ and $\hat{L}^{\mathrm{OLS}}$, $\hat{\sigma}_g^2$ and $\hat{L}$ defined in (7) and (8) are independent of Y_1. To perform inference on β_g, we assume

$$\hat{\sigma}_g^{-1}\left\{\left(X^{\mathrm{T}}X\right)^{-1} + \hat{\Omega}\left(\hat{C}_2^{\mathrm{T}}\hat{C}_2\right)^{-1}\hat{\Omega}^{\mathrm{T}}\right\}^{-1/2}\left(\hat{\beta}_g - \beta_g\right) \sim N_d\left(0, I_d\right) \quad (g = 1, \ldots, p).$$

2.4. *Intuition regarding Step (b) of Algorithm 2*

The estimates of σ_g^2 $(g = 1, \ldots, p)$ and L in Step (a) of Algorithm 2 are similar to those used in Sun et al. (2012), Gagnon-Bartsch et al. (2013), Lee et al. (2017) and Wang et al. (2017). However, the estimate of Ω in Step (b) is different from those used in previous methods. Recall from (4) that $Y_1 = B + L\Omega^{\mathrm{T}} + \mathcal{E}_1$. If B is sufficiently sparse, Sun et al. (2012), Gagnon-Bartsch et al. (2013), Lee et al. (2017) and Wang et al. (2017) propose using variations of the following estimator to recover Ω:

$$\hat{\Omega}^{\text{shrunk}} = Y_1^{\mathrm{T}}\hat{L}\left(\hat{L}^{\mathrm{T}}\hat{L}\right)^{-1}. \tag{12}$$

That is, they ignore the uncertainty in $\hat{L}$ when regressing Y_1 on to $\hat{L}$. To see why this is imprudent, let $\hat{R} = \hat{L} - L$ be the residual and suppose for the sake of argument that $L \approx 0$. Then the regression coefficients from the regression $Y_1 \sim \hat{L}$ should be very close to 0, since $\hat{L} \approx \hat{R}$ is independent of Y_1. In other words, existing estimates of Ω are shrunk towards 0. We quantify the shrinkage exactly in § 3.3 and use that result to derive an inflation term, $\hat{\Gamma} = \text{diag}\left\{\hat{\lambda}_1/(\hat{\lambda}_1 - \hat{\rho}), \ldots, \hat{\lambda}_K/(\hat{\lambda}_K - \hat{\rho})\right\}$. We then use $\hat{\Gamma}$ to inflate the shrunken estimate $\hat{\Omega}^{\text{shrunk}}$, which allows us to better estimate $\beta_1, \ldots, \beta_p$ in Step (c) of Algorithm 2.

The importance of the inflation term $\hat{\Gamma}$ in (10) is related to how informative the data are for C. The estimate $\hat{\lambda}_k$ $(k = 1, \ldots, K)$ defined in (9) is the kth largest eigenvalue of $p^{-1}Y_2^{\mathrm{T}}Y_2$, and can therefore be viewed as an estimate of λ_k, the kth largest eigenvalue of $\mathcal{I} = p^{-1}E(Y_2)^{\mathrm{T}}E(Y_2) = P_X^{\perp}C\left(p^{-1}L^{\mathrm{T}}L\right)C^{\mathrm{T}}P_X^{\perp}$. The eigenvalue λ_k is also the kth largest eigenvalue of $\left(np^{-1}L^{\mathrm{T}}L\right)\left(n^{-1}C^{\mathrm{T}}P_X^{\perp}C\right)$. When λ_k is sufficiently large for all $k = 1, \ldots, K$, we say that the data are strongly informative for the latent factors C. Under this regime, $\hat{\lambda}_1, \ldots, \hat{\lambda}_K$ will tend to dominate $\hat{\rho}$, an estimate of the constant ρ defined in (2), meaning $\hat{\Gamma}$ will be negligible. In this case it suffices to use $\hat{\Omega}^{\text{shrunk}}$ or other previously proposed estimates of Ω in place of $\hat{\Omega}$ in (11). On the other hand, we say the data are only moderately informative for C if one or more of $\lambda_1, \ldots, \lambda_K$ is not large. This can occur if the sample size n is not large enough, if some of the columns of C do not affect the expression or methylation of all p genomic units, or if X is correlated with the columns of C, since this will dampen $P_X^{\perp}C$. In these cases, $\hat{\Gamma}_{11}, \ldots, \hat{\Gamma}_{KK}$ will be moderate to large. In fact, we prove in § 3.3 and show with simulation and a real data example in § 4 that existing methods that ignore the shrinkage in their estimates of Ω are not amenable to inference. We define the informativeness of the data for C precisely in Definition 1 in § 3.3.

3. Theoretical results

3.1. *Assumptions*

In all of our assumptions and theoretical results, we assume model (1) holds, Y_1 and Y_2 are as defined in (4) and (5), and $B = \left(\beta_1 \cdots \beta_p\right)^{\mathrm{T}} \in \mathbb{R}^{p \times d}$.

Assumption 1. (a) Let X be an observed, nonrandom matrix such that $\lim_{n\to\infty} n^{-1}X^{\mathrm{T}}X = \Sigma_X \succ 0$.

(b) Let $LC^{\mathrm{T}} \in \mathbb{R}^{p\times n}$ be an unobserved, nonrandom matrix with K nonzero singular values, where $K \geqslant 1$ is a known constant.

(c) For some constant $c_1 > 1$, $\sigma_g^2 \in \left[c_1^{-1}, c_1\right]$ for all $g = 1, \ldots, p$.

Under (a) and (b), $E(Y_2) = L\left(P_X^{\perp}C\right)^{\mathrm{T}} = LC_2^{\mathrm{T}}$, $E(Y_1) = B + L\Omega^{\mathrm{T}}$ and $\text{var}\left(Y_{g1}\right) = \sigma_g^2$ $(g = 1, \ldots, p)$ are identifiable. The choice to treat LC^{T} as nonrandom is to illustrate that ignoring this term tends to bias estimates of B. However, all of our results in § § 3.2–3.4 can be extended to the case when LC^{T} is a random variable using results from the Supplementary Material. Item (c) is a standard assumption in the high-dimensional factor analysis literature (Bai & Li, 2012; Wang et al., 2017). We next place assumptions on LC_2^{T}.

Assumption 2. Let $\mathcal{I} = C_2 \left(p^{-1}L^{\mathrm{T}}L\right) C_2^{\mathrm{T}} \in \mathbb{R}^{n\times n}$ and $c_2 > 1$ be a constant and let:

(a) $L_{g*}^{\mathrm{T}} \left(n^{-1}C_2^{\mathrm{T}}C_2\right) L_{g*} \leqslant c_2^2$ for all $g = 1, \ldots, p$;

(b) $\mathcal{I}$ has K nonzero eigenvalues $\lambda_1 > \cdots > \lambda_K > 0$ such that $\lambda_k \in \left[c_2^{-1}, c_2 n\right]$ and $(\lambda_k - \lambda_{k+1})/\lambda_k \geqslant c_2^{-1}$ for all $k = 1, \ldots, K$, where $\lambda_{K+1} = 0$;

(c) p be a nondecreasing function of n such that $n/p \leqslant c_2$ and $n^{3/2}/(p\lambda_K) \to 0$ as $n \to \infty$.

The quantity $L_{g*}^{\mathrm{T}} \left(n^{-1}C_2^{\mathrm{T}}C_2\right) L_{g*}$ is identifiable because LC_2^{T} is identifiable, and (a) is equivalent to $\|L_{g*}\|_2 \leqslant c_2$ for all $g = 1, \ldots, p$ if $n^{-1}C_2^{\mathrm{T}}C_2 = I_K$. We comment on this further after we state Proposition 1 below. The assumptions on $\lambda_1, \ldots, \lambda_K$ in (b) are weaker than those considered in previous work that provide inferential guarantees, which focused on the case when $\lambda_1 \asymp \lambda_K \asymp n$ (Bai & Li, 2012; Fan & Han, 2017; Wang et al., 2017). Lee et al. (2017) do allow $\lambda_K = o(n)$, provided $\lambda_1 \asymp \lambda_K$ and $\lambda_K \to \infty$ as $n \to \infty$. However, they only prove the consistency of their estimates of $\beta_1, \ldots, \beta_p$. In fact, we show in § 3.3 that inference with their method, as well as other existing methods, is fallacious if $\lambda_K \in \left[c^{-1}, cn^{1/2}\right]$ for some $c > 1$. The assumptions on n, p in (c) are the same as those used by Wang et al. (2017), who only consider the case $\lambda_1 \asymp \lambda_K \asymp n$. We next place assumptions on the parameters of $E(Y_1) = B + L\Omega^{\mathrm{T}}$.

Assumption 3. Let $c_3 > 0$ be a constant.

(a) Let $p^{-1}\left\{I(B_{1r} \neq 0) + \cdots + I\left(B_{pr} \neq 0\right)\right\} = o\left(n^{-3/2}\lambda_K\right)$ for all $r = 1, \ldots, d$ as $n, p \to \infty$.

(b) Let $|B_{gr}| \leqslant c_3$ for all $g = 1, \ldots, p$ and $r = 1, \ldots, d$.

(c) Let $C \in \mathbb{R}^{n\times K}$ be any matrix such that $E(Y) = BX^{\mathrm{T}} + LC^{\mathrm{T}}$ for some $L \in \mathbb{R}^{p\times K}$. Then for $\Omega = \left(X^{\mathrm{T}}X\right)^{-1} X^{\mathrm{T}}C$ and $C_2 = P_X^{\perp}C$, $\|\Omega \left(n^{-1}C_2^{\mathrm{T}}C_2\right)^{-1} \Omega^{\mathrm{T}}\|_2 \leqslant c_3$.

Item (a) is the same sparsity as assumed in Wang et al. (2017). Item (c) is justifiable because we prove that B and $\Omega\left(n^{-1}C_2^{\mathrm{T}}C_2\right)^{-1}\Omega^{\mathrm{T}}$ are identifiable under Assumptions 1, 2 and 3(a) in Proposition 1 below, and Proposition S1 in § S2.1 of the Supplementary Material.

In DNA methylation data with $p \approx 3 \times 10^5$–8×10^5, $n \approx 10^2$ and in the previously unexplored regime $\lambda_K \in \left[c^{-1}, cn^{1/2}\right]$, Assumption 3(a) restricts the number of genomic units with nonzero coefficient of interest to be on the order of hundreds to thousands, which is common in many studies (Liu et al., 2018; Morales et al., 2016; Yang et al., 2017; Zhang et al., 2018). We also show through simulations that we can egregiously violate Assumption 3(a) and still perform accurate inference on $\beta_1, \ldots, \beta_p$. We now state a proposition regarding the identifiability of L and C.

Proposition 1. *Let* $\mathcal{G} = \{\mathrm{diag}(a_1, \ldots, a_K) : a_1, \ldots, a_K \in \{-1, 1\}\}$, *suppose Assumptions* 1 *and* 2 *hold and define the parameter space*

$$\begin{aligned}\Theta_{(0)} = \Big\{ &(L, C) \in \mathbb{R}^{p\times K} \times \mathbb{R}^{n\times K} : E(Y_2) = LC_2^{\mathrm{T}}, n^{-1}C_2^{\mathrm{T}}C_2 = I_K, \\ &np^{-1}L^{\mathrm{T}}L = \mathrm{diag}(\lambda_1, \ldots, \lambda_K) \textit{ for } C_2 = P_X^{\perp}C\Big\}.\end{aligned} \tag{13}$$

Then $\Theta_{(0)}$ *is nonempty and if* $\left\{L^{(a)}, C^{(a)}\right\}, \left\{L^{(b)}, C^{(b)}\right\} \in \Theta_{(0)}$, *then* $L^{(b)} = L^{(a)}\Pi$ *and* $C_2^{(b)} = C_2^{(a)}\Pi$ *for some* $\Pi \in \mathcal{G}$. *If Assumptions* 1, 2 *and* 3(a) *hold, then there exists a constant* $c_4 > 0$ *such that* B *is identifiable and*

$$\Theta_{(1)} = \Theta_{(0)} \cap \left\{(L, C) \in \mathbb{R}^{p\times K} \times \mathbb{R}^{n\times K} : E(Y_1) = B + L\Omega^{\mathrm{T}} \textit{ for } \Omega = \left(X^{\mathrm{T}}X\right)^{-1} X^{\mathrm{T}}C\right\} \tag{14}$$

is nonempty for all $n \geqslant c_4$. Further, if $\{L^{(a)}, C^{(a)}\}, \{L^{(b)}, C^{(b)}\} \in \Theta_{(1)}$, then $L^{(b)} = L^{(a)}\Pi$ and $C^{(b)} = C^{(a)}\Pi$ for some $\Pi \in \mathcal{G}$ for all $n \geqslant c_4$.

The condition that $(L, C) \in \Theta_{(0)}$ is a classic constraint to identify the components of factor models (Bai & Li, 2012). If $(L, C) \in \Theta_{(0)}$, Assumption 2(a) becomes $\|L_{g*}\|_2 \leqslant c_2$ for all $g = 1, \ldots, p$, and if $(L, C) \in \Theta_{(1)}$, $\Omega \left(n^{-1}C_2^{\mathrm{T}}C_2\right)^{-1} \Omega^{\mathrm{T}} = \Omega\Omega^{\mathrm{T}}$. While we prove it is unnecessary to assume a particular parametrization of L and C to estimate and perform inference on $\beta_1, \ldots, \beta_p$ using Algorithm 2, we use the parameter spaces $\Theta_{(0)}$ and $\Theta_{(1)}$ in the statements of theoretical results regarding the accuracy of estimates of L and Ω, respectively, in § § 3.2–3.4.

3.2. *Asymptotic properties of the estimates from Step (a) of Algorithm 2*

We start by illustrating the asymptotic properties of $\hat{\sigma}_g^2$ $(g = 1, \ldots, p)$ and $\hat{L}$ defined in (7) and (8).

Lemma 1. *Suppose Assumptions 1 and 2 hold and $n \to \infty$. Then, for ρ defined in (2),*

$$\hat{\sigma}_g^2 = \sigma_g^2 + o_{\mathrm{p}}(1) \quad (g = 1, \ldots, p), \tag{15}$$

$$\hat{\rho} = p^{-1}\left(\hat{\sigma}_1^2 + \cdots + \hat{\sigma}_p^2\right) = \rho + o_{\mathrm{p}}\left(n^{-1/2}\right). \tag{16}$$

Lemma 2. *Suppose Assumptions 1 and 2 hold and $n \to \infty$. Then, for $\hat{\lambda}_1, \ldots, \hat{\lambda}_K$ defined in (9),*

$$\hat{\lambda}_k/\lambda_k = 1 + \rho/\lambda_k + o_{\mathrm{p}}\left(n^{-1/2}\right) \quad (k = 1, \ldots, K). \tag{17}$$

Let $\Theta_{(0)}$ and $\hat{C}_2$ be as defined in (13) and Step (a) of Algorithm 2, respectively. If we also assume that $(L, C) \in \Theta_{(0)}$ and the K diagonal elements of $\hat{C}_2^{\mathrm{T}}C_2$ are nonnegative, then, for $W \sim N_K(0, I_K)$,

$$n^{1/2}\hat{\sigma}_g^{-1}\left(\hat{L}_{g*} - L_{g*}\right) \overset{\mathcal{D}}{=} W + o_{\mathrm{p}}(1) \quad (g = 1, \ldots, p). \tag{18}$$

Remark 1. The identifiability constraints, that $(L, C) \in \Theta_{(0)}$ and $\hat{C}_2^{\mathrm{T}}C_2$ has nonnegative diagonal elements, are equivalent to the IC3 constraint used in Bai & Li (2012) to identify the components of factor models.

Remark 2. When C is observed and $(L, C) \in \Theta_{(0)}$, (17) and (18) hold for the ordinary least squares estimator $\hat{L}^{\mathrm{OLS}}$ defined in Step (a) of Algorithm 1.

Lemmas 1 and 2 show that $\hat{L}_{g*}$ and $\hat{\sigma}_g^2$ have the same asymptotic properties as $\hat{L}_{g*}^{\mathrm{OLS}}$ and $\left(\hat{\sigma}_g^{\mathrm{OLS}}\right)^2$, the ordinary least squares estimates of L_{g*} and σ_g^2 defined in Algorithm 1. However, (17) states that the estimates of λ_k are biased by ρ, which we show below is the primary reason why previously proposed methods often return inflated test statistics.

3.3. *Previous estimates of Ω in Step (b) of Algorithm 2 inflate test statistics*

Existing methods that use the estimation paradigm outlined in Algorithm 2 ignore the uncertainty in $\hat{L}$, and use variations of $\hat{\Omega}^{\mathrm{shrunk}}$ to estimate Ω. We show in Proposition 2 and Corollary 1 below that these methods tend to underestimate Ω, which can lead to spurious inference on $\beta_1, \ldots, \beta_p$.

Proposition 2. *Suppose Assumptions 1, 2 and 3 hold with $\beta_1 = \cdots = \beta_p = 0$, $n \to \infty$ and $(L, C) \in \Theta_{(1)}$, where $\Theta_{(1)}$ was defined in (14). In addition, suppose the diagonal elements of $\hat{C}_2^{\mathrm{T}} C_2$ are nonnegative and $\lambda_1/\lambda_K \leqslant c_5$ for some constant $c_5 > 1$. If we estimate Ω as $\hat{\Omega}^{\text{shrunk}}$ defined in (12), then*

$$n^{1/2} \| \hat{\Omega}^{\text{shrunk}} - \Omega \operatorname{diag}\{\lambda_1/(\rho+\lambda_1), \ldots, \lambda_K/(\rho+\lambda_K)\} \|_2 = o_{\mathrm{p}}(1). \tag{19}$$

Corollary 1. *Fix some $g \in [p]$ and let $c_6 > 0$ be a small constant. In addition to the assumptions of Proposition 2, suppose $d = 1$ and the following hold:*

(i) *We replace $\hat{\Omega}$ with $\hat{\Omega}^{\text{shrunk}}$ in (11) and estimate $\beta_g = 0 \in \mathbb{R}$ as $\hat{\beta}_g^{\text{shrunk}} = Y_{1_{g*}} - \hat{\Omega}^{\text{shrunk}} \hat{L}_{g*}$.*

(ii) *There exists some constant $\epsilon > 0$ such that, $|\sum_{k=1}^{K} \Omega_k L_{gk} \{(\lambda_K + \rho) / (\lambda_k + \rho)\}| \geqslant \epsilon$, where Ω_k is the kth element of $\Omega \in \mathbb{R}^{1 \times K}$.*

Define $z_g = \hat{\sigma}_g^{-1} \left(\|X\|_2^{-2} + n^{-1} \|\hat{\Omega}^{\text{shrunk}}\|_2^2 \right)^{-1/2} \hat{\beta}_g^{\text{shrunk}}$ to be the gth z-score and let $\alpha \in (0, 1)$ be any significance level. Then for $q_{1-\alpha/2}$, the $1 - \alpha/2$ quantile of the standard normal distribution, there exists a constant $\delta > 0$ such that, as $n \to \infty$,

$$\begin{cases} \operatorname{pr}\left(|z_g| > q_{1-\alpha/2}\right) = \alpha + o(1) & \text{if } \lambda_K^{-1} n^{1/2} \to 0, \\ \operatorname{pr}\left(|z_g| > q_{1-\alpha/2}\right) \geqslant \alpha + \delta + o(1) & \text{if } \lambda_K^{-1} n^{1/2} \geqslant c_6, \\ \operatorname{pr}\left(|z_g| > q_{1-\alpha/2}\right) = 1 + o(1) & \text{if } \lambda_K^{-1} n^{1/2} \to \infty, \\ \operatorname{pr}\left\{|z_g| > q_{1-(p^{-1}\alpha)/2}\right\} = 1 + o(1) & \text{if } \lambda_K^{-1} n^{1/2 - c_6} \to \infty \text{ and} \\ & \quad n^{-r} p \to 0 \text{ for some constant } r > 0, \end{cases}$$

where $p^{-1}\alpha$ is the Bonferroni threshold at a level α.

Remark 3. Gagnon-Bartsch et al. (2013) used $\hat{\Omega}^{\text{shrunk}}$ to estimate Ω, but Lee et al. (2017) and Wang et al. (2017) used slightly different estimators. We prove analogous versions of Proposition 2 and Corollary 1 for the estimators used by Lee et al. (2017) and Wang et al. (2017) in the Supplementary Material.

Remark 4. The assumption that $\lambda_1/\lambda_K \leqslant c_5$ made in Proposition 2 and Corollary 1 requires the eigenvalues be on the same order of magnitude. It is a standard assumption made by previous authors who use versions of Algorithm 2 to estimate $\beta_1, \ldots, \beta_p$ (Lee et al., 2017; Wang et al., 2017). In Remark 6, after the statement of Theorem 2, we discuss how to extend it to allow λ_1/λ_K to diverge.

When Condition (ii) in the statement of Corollary 1 does not hold, it implies that the bias ΩL_{g*} in $Y_{1_{g*}}$ is minor, or the largest components of Ω load on to the columns of L corresponding to the largest eigenvalues λ_k, which are the components least affected by the shrinkage in Proposition 2. The shrinkage in $\hat{\Omega}^{\text{shrunk}}$ will have less of an impact on inference in these cases. If $K = 1$, Condition (ii) can be replaced with $|\Omega L_{g*}| \geqslant \epsilon$ for some constant $\epsilon > 0$.

The results of Proposition 2 and Corollary 1 show that ignoring the uncertainty in $\hat{L}$ when estimating Ω can lead to inflated test statistics and Type I errors if $\lambda_K^{-1} n^{1/2}$ is not small enough, even if one uses the conservative Bonferroni threshold. We therefore define the informativeness of the data for C in terms of the magnitude of λ_K in relation to $n^{1/2}$.

Definition 1 (Informativeness of the data for C). *The data Y are strongly informative for C if $\lambda_K^{-1}n^{1/2} \to 0$ as $n \to \infty$, and moderately informative for C if there exists a constant $c_7 > 1$ such that $\lambda_K \in \left[c_7^{-1}, n^{1/2}c_7\right]$ for all n.*

Corollary 1 shows that existing methods risk performing anticonservative inference when the data are only moderately informative for C. We next show that our shrinkage-corrected estimate of Ω in (10) begets estimates of $\beta_1, \ldots, \beta_p$ that are asymptotically equivalent to the corresponding ordinary least squares estimates obtained when C is known, even when the data are only moderately informative for C.

3.4. *Estimates of $\beta_1, \ldots, \beta_p$ from Algorithms 1 and 2 are asymptotically equivalent*

We first prove that our shrinkage-corrected estimate of Ω, $\hat{\Omega}$, corrects the aforementioned shrinkage present in existing methods' estimates of Ω.

Lemma 3. *Suppose Assumptions 1, 2 and 3 hold and $(L, C) \in \Theta_{(1)}$. Further, assume the diagonal entries of $\hat{C}_2^{\mathrm{T}}C_2$ are nonnegative and $\lambda_1/\lambda_K \leqslant c_5$, where $c_5 > 1$ was defined in the statement of Proposition 2. If $\hat{\Omega}$ is defined as in (10) and $n \to \infty$, then*

$$n^{1/2}\|\hat{\Omega} - \Omega\|_2 = o_{\mathrm{p}}(1). \tag{20}$$

We use this result to prove that inference with $\hat{\beta}_g$ $(g = 1, \ldots, p)$ is asymptotically equivalent to the ordinary least squares estimator obtained when C is known.

Theorem 1. *Let $g \in [p]$ and suppose Assumptions 1, 2 and 3 hold with $\lambda_1/\lambda_K \leqslant c_5$ and $n \to \infty$. Then inference with $\hat{\beta}_g$ is asymptotically equivalent to inference with $\hat{\beta}_g^{\mathrm{OLS}}$ in the following sense:*

$$n^{1/2}\|\hat{\beta}_g - \hat{\beta}_g^{\mathrm{OLS}}\|_2 = o_{\mathrm{p}}(1), \tag{21}$$

$$\hat{\sigma}_g^{-1}\left\{\left(X^{\mathrm{T}}X\right)^{-1} + n^{-1}\hat{\Omega}\hat{\Omega}^{\mathrm{T}}\right\}^{-1/2}\left(\hat{\beta}_g - \beta_g\right) \overset{\mathcal{D}}{=} Z + o_{\mathrm{p}}(1). \tag{22}$$

The estimates $\hat{\beta}_g^{\mathrm{OLS}}$, $\hat{\Omega}$ and $\hat{\beta}_g$ are defined in (6), (10) and (11), and $Z \sim N_d\,(0, I_d)$.

In some real experimental data, the largest eigenvalue λ_1 may be substantially larger than the smallest eigenvalue λ_K. We therefore extend Theorem 1 to relax the assumption that the λ_k are the same order of magnitude in the following theorem.

Theorem 2. *Let $g \in [p]$, suppose Assumptions 1, 2 and 3 hold and assume $n \to \infty$. Define $m_k \in \mathbb{R}^p$ to be the kth left singular vector of LC_2^{T} $(k = 1, \ldots, K)$. If $(\lambda_r\lambda_s)^{1/2}\,|m_r^{\mathrm{T}}\Sigma m_s| \leqslant c_8\lambda_{\max(r,s)}$ for some constant $c_8 > 0$ for all r, $s \in [K]$, then (21) and (22) hold.*

Remark 5. Under Assumptions 1 and 2, $(\lambda_r\lambda_s)^{1/2}\,|m_r^{\mathrm{T}}\Sigma m_s|$ is identifiable for all $r, s \in [K]$. If Assumptions 1 and 2 hold and $(L, C) \in \Theta_{(0)}$, $(\lambda_r\lambda_s)^{1/2}\,|m_r^{\mathrm{T}}\Sigma m_s| = |np^{-1}L_{*r}^{\mathrm{T}}\Sigma L_{*s}|$ for all $r, s \in [K]$.

Remark 6. Proposition 2 and Corollary 1 can be extended to accommodate data where λ_1/λ_K diverges by replacing the condition that $\lambda_1/\lambda_K \leqslant c_5$ with $(\lambda_r\lambda_s)^{1/2}\,|m_r^{\mathrm{T}}\Sigma m_s| \leqslant c_8\lambda_{\max(r,s)}$ for all $r, s \in [K]$.

The condition on $m_r^{\mathrm{T}} \Sigma m_s$ $(r, s = 1, \ldots, K)$ is quite general, as it can be shown to hold in probability when $L_{g*} \sim F_\ell$ and $\sigma_g^2 \sim F_{\sigma^2}$ $(g = 1, \ldots, p)$ for any distributions F_ℓ and F_{σ^2} with compact support, such that $np^{-1} L^{\mathrm{T}} L$ has eigenvalues bounded away from 0 with high probability. We refer the reader to the Supplementary Material for more detail.

3.5. *Inference on the relationship between C and X*

One may be interested in understanding the origin of C. For example, if components of $\hat{\Omega}$ were large, it would be informative to know if this were due to random experimental variation, or if some of the columns of C truly depended on X. To incorporate this type of inference, we state the following theorem that allows C, and therefore Ω, to be treated as a random variable.

THEOREM 3. *Let $c_9 > 1$ be a constant. In addition to Assumptions 11, 11 and 3(b), suppose the following hold:*

(i) $\|X\|_\infty \leqslant c_9$ *and* $L \in \mathbb{R}^{p \times K}$ *is a nonrandom matrix such that* $np^{-1} L^{\mathrm{T}} L = \mathrm{diag}\,(\lambda_1, \ldots, \lambda_K)$, *where* $K \geqslant 1$ *is known;*

(ii) *let* $\lambda_{K+1} = 0$. *Then* $\lambda_k \in \left[c_9^{-1}, c_9 n\right]$ *and* $(\lambda_k - \lambda_{k+1})/\lambda_k \geqslant c_9^{-1}$ *for all* $k \in [K]$, $\|L_{g*}\|_2 \leqslant c_9$ *for all* $g \in [p]$ *and* $|np^{-1} L_{*r}^{\mathrm{T}} \Sigma L_{*s}^{\mathrm{T}}| \leqslant c_9 \lambda_{\max(r,s)}$ *for all* $r, s \in [K]$;

(iii) p *is a nondecreasing function of* n *such that* $n/p \leqslant c_9$, $n^{3/2}/(\lambda_K p) \to 0$ *as* $n \to \infty$ *and* $p^{-1}\left\{I(B_{1r} \mp 0) + \cdots + I\left(B_{pr} \mp 0\right)\right\} = o\left(n^{-3/2} \lambda_K\right)$ *for all* $r \in [d]$;

(iv) $C = XA + \Xi \in \mathbb{R}^{n \times K}$, *where* $A \in \mathbb{R}^{d \times K}$ *is nonrandom and* $\Xi \in \mathbb{R}^{n \times K}$ *has independent and identically distributed rows* $\Xi_{1*}, \ldots, \Xi_{n*} \in \mathbb{R}^K$ *that are independent of* $\mathcal{E}$ *such that* $E(\Xi_{i*}) = 0$, $E\left(\Xi_{i*} \Xi_{i*}^{\mathrm{T}}\right) = I_K$ *and* $E\left(\Xi_{ik}^4\right) < \infty$ *for all* $i \in [n]$ *and* $k \in [K]$.

Let $\mathcal{W}_d(I_d, K)$ *be the standard Wishart distribution in* d *dimensions with* K *degrees of freedom. If the null hypothesis* $A = 0$ *is true and* $n \to \infty$, *then*

$$\left(X^{\mathrm{T}} X\right)^{1/2} \hat{\Omega} \hat{\Omega}^{\mathrm{T}} \left(X^{\mathrm{T}} X\right)^{1/2} \overset{\mathcal{D}}{=} V + o_{\mathrm{p}}(1),$$

where $\hat{\Omega}$ *is defined in* (10) *and* $V \sim \mathcal{W}_d(I_d, K)$. *If* $d = 1$, $V \sim \chi_K^2$.

Remark 7. Under the definition of C in (iv), $\Omega = A + \left(X^{\mathrm{T}} X\right)^{-1} X^{\mathrm{T}} \Xi$ and $E(\Omega) = A$.

4. SIMULATIONS AND DATA ANALYSIS

4.1. *Simulation study*

In this section we use simulations to compare the performance of our shrinkage-corrected method defined by Algorithm 2 with that of methods proposed in Leek & Storey (2008), Gagnon-Bartsch & Speed (2012), Gagnon-Bartsch et al. (2013), Lee et al. (2017) and Wang et al. (2017), as well as the ordinary least squares estimator when C is known and when it is ignored. We do not include results from Fan & Han (2017) or Houseman et al. (2014), because these methods perform similarly to those proposed in Lee et al. (2017) and Wang et al. (2017). In all of our simulations, we set $n = 100$, $p = 10^5$ and $K = 10$ to mimic DNA methylation data where p ranges from 3×10^5 to 8×10^5, although our results are nearly identical for p on the order of gene expression data ($p \approx 10^4$). We set $d = 1$ and assigned 50 samples to the treatment group

Table 1. *The τ_k and π_k values ($k = 1, \ldots, 10$) used to simulate L*

Factor no. (k)	1	2	3	4	5	6	7	8	9	10
τ_k	1	0.78	0.60	0.5	0.5	0.5	0.5	0.5	0.5	0.5
π_k	0	0	0	0.13	0.48	0.85	0.89	0.92	0.94	0.96
λ_k	98.0	58.9	35.4	21.3	12.8	3.8	2.7	1.9	1.4	1.0

and the rest to the control group so that $X = (1_{n/2}^{\mathrm{T}}, 0_{n/2}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^n$. We then set the eigenvalues $\lambda_1, \ldots, \lambda_K$ so that $\lambda_1 = n - 2$, $\lambda_K = 1$ and, for the others,

$$\lambda_k = \begin{cases} (n-2)^{(K-k)/(K-1)} & k \leqslant K/2, \\ \{(n-2)/5\}^{(K-k)/(K-1)} & k > K/2. \end{cases}$$

For a predefined value of $A \in \mathbb{R}^{1\times K}$ we simulated $B = \left(\beta_1 \cdots \beta_p\right)^{\mathrm{T}} \in \mathbb{R}^{p\times 1}$, $L \in \mathbb{R}^{p\times K}$, $C \in \mathbb{R}^{n\times K}$, $\Sigma = \mathrm{diag}\left(\sigma_1^2, \ldots, \sigma_p^2\right)$ and $\mathcal{E} \in \mathbb{R}^{p\times n}$ according to

$$\begin{aligned} \beta_g &\sim 0.95\delta_0 + 0.05N(0, 0.4^2) && (g = 1, \ldots, p), \\ \tau_k^2 &= \max\{\lambda_k/(n-2), 0.5^2\} && (k = 1, \ldots, K), \\ L_{gk} &\sim \pi_k\delta_0 + (1-\pi_k)\,N(0, \tau_k^2) && (g = 1, \ldots, p;\ k = 1, \ldots, K), \\ C &\sim MN_{n\times K}(XA, I_n, I_K), && \\ \sigma_g^2 &\sim \mathrm{Ga}(1/0.5^2, 1/0.5^2) && (g = 1, \ldots, p), \\ \mathcal{E}_{gi} &\sim 2^{-1/2}\sigma_g T_4 && (g = 1, \ldots, p;\ i = 1, \ldots, n), \end{aligned} \tag{23}$$

where π_k was chosen so that $E\left(L_{*k}^{\mathrm{T}} L_{*k}\right) = \lambda_k$ and T_4 is the t-distribution with four degrees of freedom. We then set the observed data to be $Y = BX^{\mathrm{T}} + LC^{\mathrm{T}} + \mathcal{E} \in \mathbb{R}^{p\times n}$. Although our theory from § 3 assumes the residuals $\mathcal{E}$ are normally distributed, we simulated t-distributed data to mimic real data with heavy tails. The values used for τ_k and π_k ($k = 1, \ldots, 10$) are given in Table 1. We show additional simulation results where we simulate B according $\beta_g \sim 0.80\delta_0 + 0.20N(0, 0.4^2)$ in the Supplementary Material.

We set the parameter A used to simulate C, where $A = E\,(\Omega)$ in (23), to be one of two values:

$$A_1 = \alpha\left(1_5^{\mathrm{T}}, 0_5^{\mathrm{T}}\right) \qquad A_2 = \alpha\left(0_5^{\mathrm{T}}, 1_5^{\mathrm{T}}\right),$$

with the scalar α chosen so that C explained 30% of the variability in group status X, on average. The choice of 30% was not arbitrary, as we estimated that over 30% of the variance in group status was explained by C in our data application in § 4.2.

As simulated, the eigenvalues $\lambda_1, \ldots, \lambda_5$ are large enough that the shrinkage terms $\lambda_k/\left(\lambda_k + \rho\right)$ ($k = 1, \ldots, 5$) from (19) in Proposition 2 are negligible. This implies that when $A = A_1$, $\hat{\Omega}^{\text{shrunk}}$ will likely be a suitable estimate of $\Omega \in \mathbb{R}^{1\times 10}$, since $\hat{\Omega}^{\text{shrunk}}$ will correctly estimate the largest and most important components of Ω, $\Omega_{*1}, \ldots, \Omega_{*5}$. The anticonservative nature of $\hat{\Omega}^{\text{shrunk}}$ implied by Corollary 1 does not apply when $A = A_1$ because Condition (ii) Corollary 1 will generally not hold. We would therefore expect our shrinkage-corrected method defined by Algorithm 2 to perform similarly to previous methods that ignore the shrinkage in their estimates of Ω in this simulation scenario. However, when $A = A_2$, $\hat{\Omega}^{\text{shrunk}}$ will not recover the largest and most

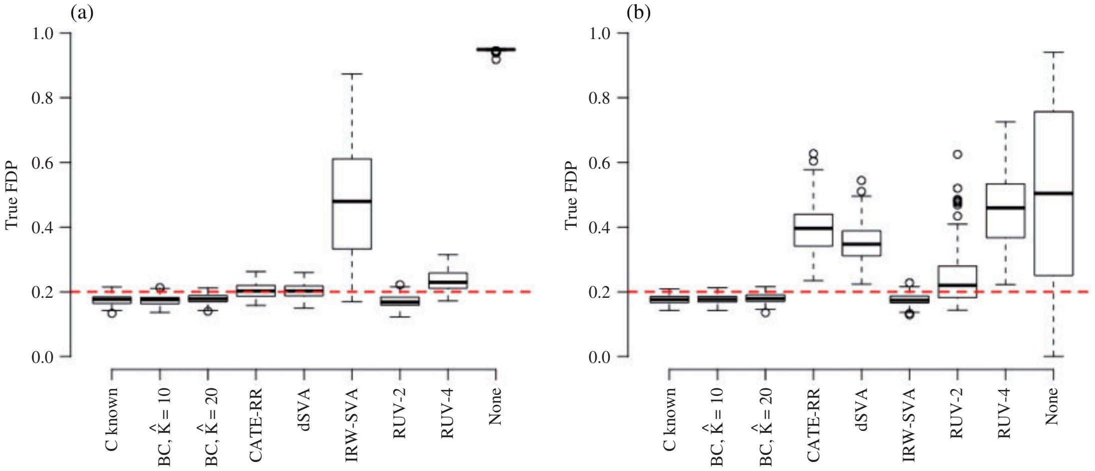

Fig. 1. The false discovery proportion, FDP, for each method at a q-value threshold of 0.2 in simulations when (a) $A = A_1$ and (b) $A = A_2$. BC is our shrinkage-corrected method defined in Algorithm 2 and $\hat{K}$ is the number of factors used to estimate C. CATE-RR, dSVA, IRW-SVA, RUV-2 and RUV-4 are the methods proposed in Leek & Storey (2008), Gagnon-Bartsch & Speed (2012), Gagnon-Bartsch et al. (2013), Lee et al. (2017) and Wang et al. (2017), respectively. These five methods were all applied with $\hat{K} = K = 10$. Inference with None was performed using the design matrix $(1_n X)$.

consequential components of Ω, $\Omega_{*6}, \ldots, \Omega_{*10}$, because of the substantial shrinkage caused by the relatively small eigenvalues $\lambda_6, \ldots, \lambda_{10}$. In this case, Corollary 1 and Remark 6 suggest that ignoring the shrinkage will lead to anticonservative inference on $\beta_1, \ldots, \beta_p$, whereas Theorems 1 and 2 imply that our shrinkage-corrected method will be asymptotically equivalent to ordinary least squares when C is observed.

We simulated 100 datasets with $A = A_1$ and another 100 with $A = A_2$. We found that we could perform the best inference on $\beta_1, \ldots, \beta_p$ with each method by performing ordinary least squares with the design matrix $(1_n \, X \, \hat{C})$, where $\hat{C}$ was C if C was known, or was estimated with any one of the six methods described above. Our shrinkage-corrected estimate of C was $\hat{C}_2 + X\hat{\Omega}$, where $\hat{C}_2$ was defined in Step (a) of Algorithm 2. We describe how the other five methods estimate C below. We compared the ordinary least squares t-statistics from all the methods to a t-distribution with $n - 2 - K$ degrees of freedom to compute p-values for the null hypotheses $H_{0,g} : \beta_g = 0$ $(g = 1, \ldots, p)$. We then judged the performance of each method by comparing their true false discovery proportion at a nominal 20% false discovery rate, estimated using q-values (Storey, 2001), because this is the inference method popular among biologists.

Figure 1 provides the simulation results. We see that our shrinkage-corrected method is able to control the false discovery rate both when K is known to be 10 and when we drastically overestimate it to be 20. Further, our method's power to detect units with nonzero β_g at this nominal 20% false discovery rate threshold was 13.6% when $\hat{K} = 10$ and 12.8% when $\hat{K} = 20$, which is compared to 13.6% when C was known. The power of all three methods was the same for both values of A. This is exactly what one would expect from Theorems 1 and 2, which prove that inference with our shrinkage-corrected estimator is asymptotically equivalent to that with ordinary least squares when C is known. This equivalence was also manifested when we overtly violated Assumption 3(a) and simulated $\beta_1, \ldots, \beta_p \sim 0.80\delta_0 + 0.20N(0, 0.4^2)$; see the Supplementary material for more detail.

It is also informative to study the performance of the other five methods, as this can be important to practitioners deciding which method to apply to their data.

The methods of Wang et al. (2017), CATE-RR, and Lee et al. (2017), dSVA, estimate C as $\hat{C}_2^{\text{cate}} + X\hat{\Omega}^{\text{cate}}$ and $\hat{C}_2^{\text{dSVA}} + X\hat{\Omega}^{\text{dSVA}}$, respectively, where their estimates of C_2, $\hat{C}_2^{\text{cate}}$ and $\hat{C}_2^{\text{dSVA}}$, are nearly identical to $\hat{C}_2$ as defined in Step (a) of Algorithm 2. However, their estimates of Ω, $\hat{\Omega}^{\text{cate}}$ and $\hat{\Omega}^{\text{dSVA}}$, ignore the shrinkage described in Proposition 2. We would therefore expect them to introduce more Type I errors when $A = A_2$. Both CATE-RR and dSVA's false discovery proportion estimates were closer to nominal values when $\beta_1, \ldots, \beta_p \sim 0.80\delta_0 + 0.20N(0, 0.4^2)$, since any rejection region was likely to have more genomic units with nonzero coefficients of interest.

The method of Leek & Storey (2008), IRW-SVA, estimates C by performing a factor analysis on $\text{diag}\left(\hat{\pi}_1, \ldots, \hat{\pi}_p\right) Y$, where $\hat{\pi}_g$ is an estimate of $\text{pr}\left(\ell_g \neq 0, \beta_g = 0 \mid Y\right)$ $(g = 1, \ldots, p)$, by iteratively estimating C and $\text{pr}\left(\ell_g \neq 0, \beta_g = 0 \mid Y\right)$. Since the first iteration assumes $\hat{C} = P_X^{\perp}\hat{C}$, $\hat{\pi}_g$ tends to be small if the marginal correlation between Y_{g*} and X is large, which occurs if $|\Omega L_{g*}|$ is large. Therefore, the latent factors that influence $\text{diag}\left(\hat{\pi}_1, \ldots, \hat{\pi}_p\right) Y$ will be different than those of Y if the latent factors with the largest effects are also correlated with X. This explains why IRW-SVA performs poorly when $A = A_1$. Unfortunately, there is no theory that states when IRW-SVA is expected to accurately recover C.

Both RUV-2 (Gagnon-Bartsch & Speed, 2012) and RUV-4 (Gagnon-Bartsch et al., 2013) assume the practitioner has prior knowledge of a subset $\mathcal{S} \subseteq [p]$ of control genomic units where $\beta_g = 0$ for all $g \in \mathcal{S}$. We selected $|\mathcal{S}| = 600 = 20 \times 30$ control units uniformly at random from the set of all genomic units with $\beta_g = 0$ across all simulations, because simulations in Wang et al. (2017) use 30 control units when $p = 5000 = 10^5/20$. RUV-2 estimates C via factor analysis using only data from genomic units in $\mathcal{S}$, whereas RUV-4 first estimates C_2 and L as $\hat{C}_2$ and $\hat{L}$ defined in Step (a) of Algorithm 2, and then estimates Ω as $\hat{\Omega}^{\text{RUV-4}} = Y_{1_{\mathcal{S}*}}^{\text{T}} \hat{L}_{\mathcal{S}*}(\hat{L}_{\mathcal{S}*}^{\text{T}}\hat{L}_{\mathcal{S}*})^{-1}$. Here, $Y_{1_{\mathcal{S}*}}$ and $\hat{L}_{\mathcal{S}*}$ are the submatrices of Y_1 and $\hat{L}$ restricted to the rows in $\mathcal{S}$. The RUV-4 estimate of C is then $\hat{C}_2 + X\hat{\Omega}^{\text{RUV-4}}$. The obvious caveat for RUV-2 and RUV-4 is that the practitioner must have a list of units whose coefficients of interest are zero and whose expression or methylation carries the latent factor signature, i.e., the first K eigenvalues of $C\left(|\mathcal{S}|^{-1}L_{\mathcal{S}*}^{\text{T}}L_{\mathcal{S}*}\right)C^{\text{T}}$ must be suitably large. For example, the large variability in RUV-2's false discovery proportion when $A = A_2$ is because the $|\mathcal{S}| = 600$ control units were not sufficient to capture the latent factor signature in many simulations.

4.2. *Data application*

In order to demonstrate the importance of using our shrinkage-corrected estimator, we applied our method to reanalyse data from Nicodemus-Johnson et al. (2016), which studied the correlation between adult asthma and DNA methylation in lung epithelial cells. The authors collected endobronchial brushings from 74 adult patients with a current doctor's diagnosis of asthma and 41 healthy adults, and quantified their DNA methylation at $p = 327\,271$ methylation sites, also referred to as CpGs, using the Infinium Human Methylation 450K Bead Chip (Dedeurwaerder et al., 2011). Nicodemus-Johnson et al. (2016) then used ordinary least squares to regress the methylation at each of the p sites on to the mean model subspace that included asthma status, age, ethnicity, sex and smoking status to estimate the effect due to asthma, $\left(\beta_1 \cdots \beta_p\right)^{\text{T}} \in \mathbb{R}^{p\times 1}$. They found 40 892 CpGs that were differentially methylated between asthmatics and healthy patients at a nominal false discovery rate of 5%.

We investigated whether or not the strong association between DNA methylation and asthma status was in part due to unobserved covariates. In particular, lung cell composition may differ between asthmatics and nonasthmatics, with asthmatic patients generally having a greater proportion of airway goblet cells that excrete mucus (Rogers, 2002; Bai & Knight, 2005). We

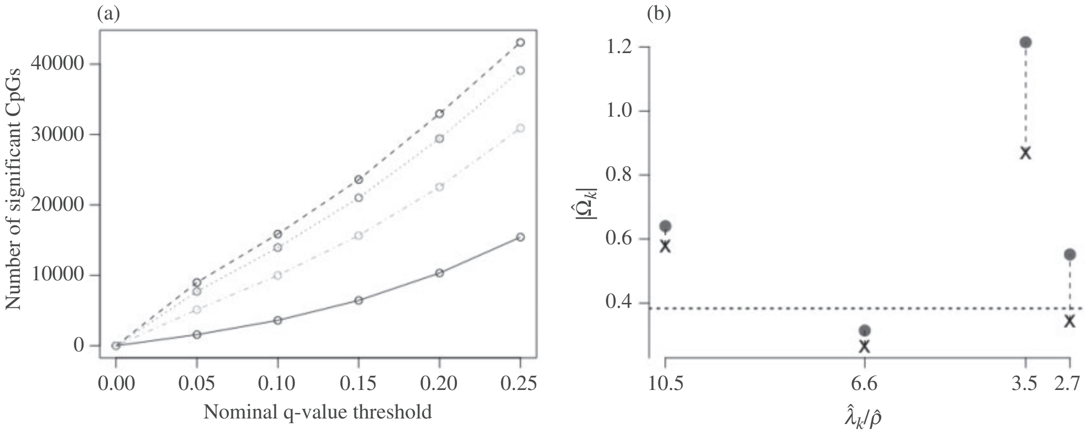

Fig. 2. Results from our analysis of lung DNA methylation data from Nicodemus-Johnson et al. (2016). (a) The number of asthma-related CpGs at a given q-value cut-off using our shrinkage-corrected estimator (solid line), as well as the estimators proposed in Lee et al. (2017) (dot-dashed line), Wang et al. (2017) (dotted line) and Leek & Storey (2008) (dashed line). (b) The $K = 4$ components of $\hat{\Omega}^{\text{shrunk}}$ ($\times$) and $\hat{\Omega}$ ($\bullet$) as a function of $\hat{\lambda}_1, \ldots, \hat{\lambda}_4$. The dashed line is the 0.95 quantile of the $\tilde{n}^{-1/2}\chi_1$ distribution, where $\tilde{n}$ is defined such that $\tilde{n}\hat{\Omega}\hat{\Omega}^{\mathrm{T}}$ converges to a chi-squared random variable with $K = 4$ degrees of freedom under the null hypothesis from Theorem 3.

therefore reanalysed these data to account for latent covariates with our shrinkage-corrected method defined by Algorithm 2, and compared the results to those obtained using the methods proposed in Leek & Storey (2008), Lee et al. (2017) and Wang et al. (2017). We could not apply the methods proposed in Gagnon-Bartsch & Speed (2012) and Gagnon-Bartsch et al. (2013) because we did not have access to control CpGs. We first used bi-crossvalidation (Owen & Wang, 2016) to estimate that there were $K = 4$ latent factors in these data, and subsequently estimated $C \in \mathbb{R}^{115\times 4}$ using the four different methods. We then computed p-values for the null hypotheses $H_{0,g} : \beta_g = 0$ $(g = 1, \ldots, p)$ using ordinary least squares with the design matrix $(X\ Z\ \hat{C})$, where $X \in \{0, 1\}^n$ was asthma status and Z contained the observed nuisance covariates age, ethnicity, sex and smoking status. The total number of asthma-related CpGs returned by each method as a function of q-value cut-offs (Storey et al., 2015), as well as the uncorrected and shrinkage-corrected estimates of $\Omega \in \mathbb{R}^{1\times 4}$, are given in Fig. 2. At a q-value threshold of 20%, our method identifies 10 324 asthma-related CpGs, while the methods proposed in Leek & Storey (2008), Lee et al. (2017) and Wang et al. (2017) ostensibly identify 32 952, 29 415 and 22 545 asthma-related CpGs, respectively. These numbers changed only slightly when we let K be as high as 7.

We estimated that approximately 36% of the variance in asthma status was explained by C, which, using Theorem 3, corresponds to a p-value for the null hypothesis $E(C \mid X) = 0$ of 3.2×10^{-12}. Moreover, assuming $(L, C) \in \Theta_{(1)}$, the largest component of Ω appeared to load on to the third column of $L \in \mathbb{R}^{p\times 4}$, where $\lambda_3/\rho \approx 2.5$. Since this was much smaller than $n^{1/2} = 10.7$ and we estimated $L_{g3} \neq 0$ at over 40% of the studied CpGs $g \in [p]$, Proposition 2, Corollary 1 and simulations connote that the methods proposed in Lee et al. (2017) and Wang et al. (2017) are likely underestimating the fraction of CpGs with $\beta_g = 0$ at any nominal q-value threshold. It is likely the case that λ_3, the third largest eigenvalue of $\mathcal{I} = P_X^{\perp} C \left(p^{-1} L^{\mathrm{T}} L\right) C^{\mathrm{T}} P_X^{\perp}$, was small even though the third factor explained a significant portion of the variability in methylation levels because its strong correlation with asthma status dampened $P_X^{\perp} C$.

We next sought to determine if differences in lung cell composition between asthmatic and healthy patients were responsible for some of the correlation between asthma status and the latent factors, since understanding the origin of the latent covariates could help practitioners determine which method is most appropriate for their data. To do so, we fit a topic model with $r = 7$ topics on the same individual's gene expression data, which has been shown to cluster bulk RNA-seq samples by tissue and cell type (Taddy, 2012; Dey et al., 2017). We then used the n-dimensional factor whose corresponding loading was the largest on the *MUC5AC* gene as a proxy for the proportion of goblet cells in each sample, as *MUC5AC* is a unique identifier for goblet cells (Zuhdi Alimam et al., 2000). Just as one would expect, asthmatics tended to have a higher proportion of estimated goblet cells than healthy controls, and we rejected the null hypothesis that asthmatics and healthy controls had the same mean estimated goblet cell proportion at the significance level of $\alpha = 0.01$. This indicates that cell composition is presumably driving much of the observed correlation between methylation levels and asthma status in Nicodemus-Johnson et al. (2016), as well as the results from the reanalysis with the methods proposed in Lee et al. (2017) and Wang et al. (2017).

These conclusions also help to explain why the method proposed in Leek & Storey (2008) is likely underestimating the number of false discoveries. We estimated that $p^{-1}\left\{I\left(L_{1*} \mp 0\right) + \cdots + I\left(L_{p*} \mp 0\right)\right\} \approx 0.90$ in these data, which is precisely what one would expect if cellular heterogeneity were among the unobserved factors, since changes in methylation help drive cellular differentiation. And since we have already shown that X is correlated with C, the method proposed in Leek & Storey (2008) would not be expected to control the false discovery rate, as the simulations in § 4.1 showed exactly this when $|\Omega L_{g*}|$ was large for many genomic units $g \in [p]$.

5. Discussion

The prevalence of unobserved covariates in high-throughput omic data has precipitated the development of methods that account for unobserved factors C in downstream inference. While these methods perform well when the data are strongly informative for C, they are not amenable to inference when the data are only moderately informative for C. On the other hand, we prove that inference using estimates from our shrinkage-corrected method in Algorithm 2 is asymptotically equivalent to ordinary least squares when C is observed.

Our method is not a cure-all for inference with unobserved covariates. For example, Assumption 3(a) restricts the number of units with nonzero main effect in DNA methylation data to be on the order of hundreds to thousands when the data are only moderately informative C. Even though simulations show we can potentially relax this number substantially to tens or even hundreds of thousands in practice, it begs the question as to whether or not practitioners should spend time and money to measure nuisance variables like cellular heterogeneity, or estimate them directly from the data. If the practitioner is concerned that C is correlated with X, but has reason to believe B is sparse, our theory suggests the effort should be spent collecting more samples. However, if C is correlated with X and B is dense, it may be worthwhile to attempt to measure some of the latent factors with other technologies. We are currently working with the authors of Nicodemus-Johnson et al. (2016) to use external sources of information to potentially better account for cellular heterogeneity in their data.

Acknowledgement

We thank Carole Ober and Michelle Stein for comments that have substantially improved this manuscript. The research was supported in part by the National Institutes of Health.

Supplementary material

Supplementary material available at *Biometrika* online includes additional simulation results and proofs of all the propositions, lemmas and theorems presented in this paper. An `R` package implementing our method, together with instructions and code to reproduce the simulations from § 4.1, are available from `https://github.com/chrismckennan/BCconf`.

References

Bai, J. & Li, K. (2012). Statistical analysis of factor models of high dimension. *Ann. Statist.* **40**, 436–65.

Bai, T. R. & Knight, D. A. (2005). Structural changes in the airways in asthma: observations and consequences. *Clin. Sci.* **108**, 463–77.

Cangelosi, R. & Goriely, A. (2007). Component retention in principal component analysis with application to cDNA microarray data. *Biol. Direct* **2**, 2.

Dedeurwaerder, S., Defrance, M., Calonne, E., Denis, H., Sotiriou, C. & Fuks, F. (2011). Evaluation of the Infinium Methylation 450K technology. *Epigenomics* **3**, 771–84.

Dey, K. K., Hsiao, C. J. & Stephens, M. (2017). Visualizing the structure of RNA-seq expression data using grade of membership models. *PLOS Genetics* **13**, e1006599.

Fahy, J. V. (2002). Goblet cell and mucin gene abnormalities in asthma. *Chest* **122**, 320S–26S.

Fan, J. & Han, X. (2017). Estimation of the false discovery proportion with unknown dependence. *J. R. Statist. Soc.* B **79**, 1143–64.

Gagnon-Bartsch, J. A., Jacob, L. & Speed, T. P. (2013). Removing unwanted variation from high dimensional data with negative controls. Tech. rep. 820, UC Berkeley.

Gagnon-Bartsch, J. A. & Speed, T. P. (2012). Using control genes to correct for unwanted variation in microarray data. *Biostatistics* **13**, 539–52.

Houseman, E. A., Accomando, W. P., Koestler, D. C., Christensen, B. C., Marsit, C. J., Nelson, H. H., Wiencke, J. K. & Kelsey, K. T. (2012). DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13**, 86.

Houseman, E. A., Molitor, J. & Marsit, C. J. (2014). Reference-free cell mixture adjustments in analysis of DNA methylation data. *Bioinformatics* **30**, 1431–9.

Jaffe, A. E. & Irizarry, R. A. (2014). Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol.* **15**, R31.

Johnson, W. E., Li, C. & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–27.

Lee, S., Sun, W., Wright, F. A. & Zou, F. (2017). An improved and explicit surrogate variable analysis procedure by coefficient adjustment. *Biometrika* **104**, 303–16.

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K. & Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Rev. Genet.* **11**, 733–9.

Leek, J. T. & Storey, J. D. (2008). A general framework for multiple testing dependence. *Proc. Nat. Acad. Sci.* **105**, 18718–23.

Liu, C., Marioni, R. E., Hedman, Å. K., Pfeiffer, L., Tsai, P. C., Reynolds, L. M., Just, A. C., Duan, Q., Boer, C. G., Tanaka, T., et al. (2018). A DNA methylation biomarker of alcohol consumption. *Molec. Psychiatry* **23**, 422–33.

McKennan, C. & Nicolae, D. (2018). Estimating and accounting for unobserved covariates in high dimensional correlated data. *arXiv*:1808.05895v1.

Morales, E., Vilahur, N., Salas, L. A., Motta, V., Fernandez, M. F., Murcia, M., Llop, S., Tardon, A., Fernandez-Tardon, G., Santa-Marina, L., et al. (2016). Genome-wide DNA methylation study in human placenta identifies novel loci associated with maternal smoking during pregnancy. *Int. J. Epidemiol.* **45**, 1644–55.

Nicodemus-Johnson, J., Myers, R. A., Sakabe, N. J., Sobreira, D. R., Hogarth, D. K., Naureckas, E. T., Sperling, A. I., Solway, J., White, S. R., Nobrega, M. A., et al. (2016). DNA methylation in lung cells is associated with asthma endotypes and genetic risk. *JCI Insight* **1**, e90151.

Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *Rev. Econom. Statist.* **92**, 1004–16.

Owen, A. B. & Wang, J. (2016). Bi-cross-validation for factor analysis. *Statist. Sci.* **31**, 119–39.

Rogers, D. F. (2002). Airway goblet cell hyperplasia in asthma: Hypersecretory and anti-inflammatory? *Clin. Experim. Allergy* **32**, 1124–7.

Stein, M. M., Hrusch, C. L., Gozdz, J., Igartua, C., Pivniouk, V., Murray, S. E., Ledford, J. G., Marques Dos Santos, M., Anderson, R. L., Metwali, N., et al. (2016). Innate immunity and asthma risk in Amish and Hutterite farm children. *New Engl. J. Med.* **375**, 411–21.

Storey, J. D. (2001). A direct approach to false discovery rates. *J. R. Statist. Soc.* B **63**, 479–98.

Storey, J. D., Bass, A. J., Dabney, A. & Robinson, D. (2015). *qvalue: Q-value Estimation for False Discovery Rate Control*. R package version 2.10.0. http://github.com/jdstorey/qvalue [last accessed 14 June 2019].

Sun, Y., Zhang, N. R. & Owen, A. B. (2012). Multiple hypothesis testing adjusted for latent variables, with an application to the AGEMAP gene expression data. *Ann. Appl. Statist.* **6**, 1664–88.

Taddy, M. (2012). On estimation and selection for topic models. *Proc. Mach. Learn. Res.* **22**, 1184–93.

van Iterson, M., van Zwet, E. W. & Heijmans, B. T. (2017). Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution. *Genome Biol.* **18**, 19.

Wang, J., Zhao, Q., Hastie, T. & Owen, A. B. (2017). Confounder adjustment in multiple hypothesis testing. *Ann. Statist.* **45**, 1863–94.

Yang, I. V., Pedersen, B. S., Liu, A. H., O'Connor, G. T., Pillai, D., Kattan, M., Misiak, R. T., Gruchalla, R., Szefler, S. J., Khurana Hershey, G. K., et al. (2017). The nasal methylome and childhood atopic asthma. *J. Allergy Clin. Immunol.* **139**, 1478–88.

Zhang, X., Biagini Myers, J. M., Burleson, J., Ulm, A., Bryan, K. S., Chen, X., Weirauch, M. T., Baker, T. A., Butsch Kovacic, M. S. & Ji, H. (2018). Nasal DNA methylation is associated with childhood asthma. *Epigenomics* **10**, 629–41.

Zuhdi Alimam, M., Piazza, F. M., Selby, D. M., Letwin, N., Huang, L. & Rose, M. C. (2000). Muc-5/5ac mucin messenger RNA and protein expression is a marker of goblet cell metaplasia in murine airways. *Am. J. Respir. Cell Molec. Biol.* **22**, 253–60.

[*Received on* 23 *January* 2018. *Editorial decision on* 9 *January* 2019]