Biometrika (2017), **104**, 2, *pp.* 303–316
Printed in Great Britain
doi: 10.1093/biomet/asx018
Advance Access publication 21 April 2017

An improved and explicit surrogate variable analysis procedure by coefficient adjustment

By SEUNGGEUN LEE

Department of Biostatistics, University of Michigan, 1415 Washington Heights, Ann Arbor, Michigan 48109, U.S.A.

leeshawn@umich.edu

WEI SUN

Public Health Sciences Division, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave. N., Seattle, Washington 98109, U.S.A.

wsun@fredhutch.org

FRED A. WRIGHT

Bioinformatics Research Center, North Carolina State University, 1 Lampe Drive, Raleigh, North Carolina 27607, U.S.A.

fred_wright@ncsu.edu

and FEI ZOU

Department of Biostatistics, University of Florida, 2004 Mowry Rd, Gainesville, Florida 32611, U.S.A.

fayzou@ufl.edu

Summary

Unobserved environmental, demographic and technical factors can adversely affect the estimation and testing of the effects of primary variables. Surrogate variable analysis, proposed to tackle this problem, has been widely used in genomic studies. To estimate hidden factors that are correlated with the primary variables, surrogate variable analysis performs principal component analysis either on a subset of features or on all features, but weighting each differently. However, existing approaches may fail to identify hidden factors that are strongly correlated with the primary variables, and the extra step of feature selection and weight calculation makes the theoretical investigation of surrogate variable analysis challenging. In this paper, we propose an improved surrogate variable analysis, using all measured features, that has a natural connection with restricted least squares, which allows us to study its theoretical properties. Simulation studies and real-data analysis show that the method is competitive with state-of-the-art methods.

Some key words: Batch effect; High-dimensional data; Principal component analysis; Surrogate variable analysis.

1. INTRODUCTION

In regression analysis, the existence of unobserved factors can cause biases in estimating parameters. Suppose that the true relationship in the data is

$$y = X\beta + Z\delta + \epsilon,$$

where y is a vector of outcome measurements, X is a matrix of the observed covariates including the primary variables, and Z is a matrix of the unobserved factors. We are interested in estimating the regression parameter β. Since Z is not observed, in practice we use the misspecified model

$$y = X\beta^* + \epsilon^*,$$

which can negatively impact inference on β.

With the development of high-throughput technologies in biomedical sciences, high-dimensional data are routinely collected and analysed to find biologically meaningful features. Unobserved factors can cause adverse effects, including inflation of Type I error and/or power loss (Stegle et al., 2010). Although in practice great efforts are made to control confounders, such efforts may be insufficient to avoid all confounding issues (Leek et al., 2010).

Principal component analysis on the original or residualized features after removing the effects of observed dependent variables has often been used to identify hidden factors, and has been successful in identifying and controlling for population stratification in genome-wide association studies (Price et al., 2006). However, principal component analysis-based approaches are less effective for gene expression studies, where the hidden factors can affect a subset of features with relatively large effects (Leek & Storey, 2007). To overcome this limitation, surrogate variable analysis has been proposed (Leek & Storey, 2007, 2008; Teschendorff et al., 2011; Chakraborty et al., 2012) for microarray data. Leek & Storey (2007) initially developed a two-step approach which involves first identifying a subset of features that may be affected by hidden factors but not by primary variables, and then performing principal component analysis on the selected features. Later, they modified the approach to a weighted principal component analysis, where each feature is weighted according to its probability of being affected by the hidden factors only (Leek & Storey, 2008). Surrogate variable analysis has been extended to factor analysis (Friguet et al., 2009) and mixed-effect models (Listgarten et al., 2010). Recently, assuming that negative control genes are known, Gagnon-Bartsch & Speed (2012) proposed a surrogate variable method.

Surrogate variable analysis has been successfully applied to many genomic studies (Dumeaux et al., 2010; Teschendorff et al., 2010), but existing methods may fail to identify hidden factors. Strong correlation between hidden factors and primary variables can prevent the two-step and weighted principal component-based surrogate variable methods from identifying features that are affected by hidden factors only. If negative control genes are affected by primary variables or if the observed variation in negative control genes does not reflect unwanted variations in the entire genome, the methods for removing unwanted variation can also fail to identify true hidden factors.

In this paper, we propose a simple and straightforward method for identifying hidden factors and adjusting for their effects. Our approach, called direct surrogate variable analysis, is based on the observation that naïve estimators of the effects of the primary variables are biased when the effects of hidden factors are ignored in the analysis, but the bias can be estimated and removed using singular value decomposition on residuals. We derive the asymptotic properties of our estimators using techniques recently developed for the ultrahigh-dimensional regime (Lee et al., 2014) and the connection between our estimating procedure and the restricted least-squares

method (Greene & Seaks, 1991). An R package (R Development Core Team, 2017) implementing the proposed approach, dSVA, can be downloaded from the comprehensive R archive network.

2. Methods

2·1. *Direct surrogate variable analysis*

Suppose that Y is an $n \times m$ matrix of measured features, where m is the number of features and n is the number of samples. For gene expression data, Y represents RNA expression levels on m genes. Further, suppose that X is an $n \times p$ matrix of observed covariates, including an intercept, and Z is an $n \times q$ matrix of unobserved hidden factors. The following model represents the true relationship between Y and (X, Z):

$$y_i = X\beta_i + Z\delta_i + \epsilon_i, \tag{1}$$

where y_i denotes the ith column of Y, $\beta_i = (\beta_{1i}, \ldots, \beta_{pi})^{\mathrm{T}}$ is a $p \times 1$ vector of regression coefficients associated with X, $\delta_i = (\delta_{1i}, \ldots, \delta_{qi})^{\mathrm{T}}$ is a $q \times 1$ vector of regression coefficients associated with Z, and ϵ_i is an $n \times 1$ random vector which follows $N(0, \sigma_i^2 I)$. We further define $B = (\beta_1, \ldots, \beta_m)$ and $\Delta = (\delta_1, \ldots, \delta_m)$, which are $p \times m$ and $q \times m$ matrices of regression coefficients associated with X and Z, respectively. In this model, β_i and δ_i are assumed to be fixed. Later, to generate large numbers of β_i and δ_i values for the simulation studies, we use a specified correlation between β_i and δ_i. However, we emphasize that the proposed method is frequentist: β_i and δ_i are considered fixed and unknown.

In practice, since Z is not observed, we effectively use the misspecified model

$$y_i = X\beta_i^* + \epsilon_i^* \tag{2}$$

instead of (1). Under (2), the least-squares estimator of β_i^* is

$$\hat{\beta}_i^* = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y_i = \beta_i + (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Z\delta_i + (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\epsilon_i, \tag{3}$$

with residual vector

$$r_i = (I - M)y_i = (I - M)Z\delta_i + (I - M)\epsilon_i, \tag{4}$$

where $M = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ is the projection matrix onto the column space of X. Equations (3) and (4) indicate that $\hat{\beta}_i^*$ is a biased estimator of β_i with bias $(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Z\delta_i$. The conditional mean of the residual vector given δ_i is $(I - M)Z\delta_i$, which allows us to estimate $Z\delta_i$ via, for example, singular value decomposition.

Suppose that singular value decomposition is performed on the residual matrix $R = (r_1, \ldots, r_m)$, where $R = UDV^{\mathrm{T}}$, with D being a diagonal matrix of ordered singular values, and U and V being matrices of left- and right-singular vectors. The first q left-singular vectors can be viewed as estimators of linear combinations of the columns of $(I - M)Z$, which we denote by $(I - M)\Gamma$ where $\Gamma = ZT$, with T being a $q \times q$ orthonormal matrix. Let $\psi_i = T^{\mathrm{T}}\delta_i$. For any T, the matrices Γ and Z have the same column space, so $\Gamma\psi_i$ is identical to $Z\delta_i$. With an additional assumption that the row vectors of B and the row vectors of Δ are asymptotically orthogonal after mean centring, we can estimate Γ and use it to remove the bias in $\hat{\beta}_i^*$. The proposed method is as follows.

Step 1. Carry out singular value decomposition on the residual matrix $R = UDV^{\mathrm{T}}$. Let U_q be the matrix comprising the first q columns of U that are equivalent to the q left-singular vectors corresponding to the q largest singular values.

Step 2. Obtain $\hat{\beta}_i^*$ and $\hat{\psi}_i$ from the model $y_i = X\beta_i^* + U_q\psi_i + \epsilon_i$. Since X and U_q are orthogonal to each other, $\hat{\beta}_i^*$ from this model equals that from model (3).

Step 3. Let $\hat{B}^* = (\hat{\beta}_1^*, \ldots, \hat{\beta}_m^*)$ and $\hat{\Psi} = (\hat{\psi}_1, \ldots, \hat{\psi}_m)$. We propose to estimate the surrogate variables Γ as

$$\hat{\Gamma} = U_q + X\hat{B}^*(I - M_J)\hat{\Psi}^{\mathrm{T}}\{\hat{\Psi}(I - M_J)\hat{\Psi}^{\mathrm{T}}\}^{-1},$$

where $M_J = J(J^{\mathrm{T}}J)^{-1}J^{\mathrm{T}}$ is a projection matrix with $J = (1, \ldots, 1)^{\mathrm{T}}$.

Step 4. Estimate and test β_i from the model

$$y_i = X\beta_i + \hat{\Gamma}\psi_i + \epsilon_i. \tag{5}$$

This method requires estimation of q, the number of surrogate variables, which can be obtained by permutation (Buja & Eyuboglu, 1992) or by analytical-asymptotic approaches (Johnstone, 2001; Leek, 2011). In this paper, we use the method of Buja & Eyuboglu (1992) for all numerical work. Since Γ and ψ_i can be always rescaled, they are not identifiable, so we set $\sum_{i=1}^{n}\sum_{j=1}^{q}\Gamma_{i,j}^2 = nq$, where $\Gamma_{i,j}$ is the (i,j)th element of Γ, and adjust $\hat{\Gamma}$ to satisfy this restriction. In the Supplementary Material we show that $\hat{B} = (\hat{\beta}_1, \ldots, \hat{\beta}_m)$ from (5) in Step 4 is the same as

$$\hat{B} = \hat{B}^* - \hat{B}^*(I - M_J)\hat{\Psi}^{\mathrm{T}}\{\hat{\Psi}(I - M_J)\hat{\Psi}^{\mathrm{T}}\}^{-1}\hat{\Psi}, \tag{6}$$

in which $\hat{B}^*(I - M_J)\hat{\Psi}^{\mathrm{T}}\{\hat{\Psi}(I - M_J)\hat{\Psi}^{\mathrm{T}}\}^{-1}\hat{\Psi}$ is an estimate of the bias of the naïve estimator $\hat{B}^*$. In § 2·3, we show that (6) is related to the restricted least-squares method.

2·2. *Consistency of the proposed estimators*

Important questions are under what conditions does the proposed $\hat{\Gamma}$ span the same column space as Z, and whether $\hat{\beta}_i$ is a consistent estimator of β_i. For high-dimensional data, the number of features, m, can be substantially larger than the number of samples, n, and thus asymptotic results derived from the traditional low-dimensional setting where m is fixed are inappropriate (Johnstone & Lu, 2009; Jung & Marron, 2009; Lee et al., 2010). Lee et al. (2014) considered a regime in which both m and n increase to infinity with $m/n = \gamma_m \to \infty$. This regime is well-suited to high-throughput biomedical data, where the number of genes is in the tens of thousands and the number of samples is in the range of several dozens to hundreds. We work in this regime and investigate the asymptotic properties of the proposed method under the spiked-eigenvalue model of Johnstone (2001).

Before presenting our main results, let us define some additional notation. Suppose that a_m and b_m are two sequences. We write $a_m \asymp b_m$ if $a_m = O(b_m)$ and $b_m = O(a_m)$, and write $a_m \ll b_m$ if $a_m/b_m = o(1)$. We also define $\varphi_v(\cdot)$ to be the function that returns the vth largest singular value of an input matrix. Without loss of generality we assume that $\|X_i\| \asymp n$ and $\|Z_i\| \asymp n$, where X_i and Z_i are the ith columns of X and Z, respectively, and $\|\cdot\|$ is the vector norm. We introduce the following conditions.

Condition 1. Both m and n increase to ∞ with $m/n = \gamma_m \to \infty$.

Condition 2. Let $\lambda_v = \varphi_v\{(I - M)Z\Delta\}$, where $\Delta = (\delta_1, \ldots, \delta_m)$. Then $\lambda_1 \asymp \lambda_2 \asymp \cdots \asymp \lambda_q$, $\lambda_1 > \cdots > \lambda_q$, and $(n\gamma_m)^{-1/2}\lambda_v \to \infty$ for $v = 1, \ldots, q$.

Condition 3. Let $\phi(k) = m^{-1}\sum_{i=1}^m (\sigma_i^2 - \bar{\sigma}^2)^k$, where σ_i is the standard deviation of ϵ_i and $\bar{\sigma}^2 = m^{-1}\sum_{i=1}^m \sigma_i^2$. Then either of the following is satisfied: (i) $\phi(2) = o(n^{-2}m)$; or (ii) $\phi(2) = o(n^{-3/2}m)$, $\phi(4) = O(1)$ and $\phi(4) = o(n^{-4}m^3)$.

Condition 4. Let $A_n = (X, Z)$ be the matrix with $p + q$ columns formed by concatenating X and Z. Then $A_n^{\mathrm{T}}A_n$ is nonsingular with $\varphi_1(A_n) \asymp n$ and $\varphi_{p+q}(A_n) \asymp n$.

Condition 5. Suppose that β_{ki} and δ_{ki} are the (k, i)th elements of B and Δ, respectively, and that $\bar{\beta}_k = \sum_{i=1}^m \beta_{ki}/m$ and $\bar{\delta}_k = \sum_{i=1}^m \delta_{ki}/m$. For all (k, l),

$$\frac{1}{m}\sum_{i=1}^m (\beta_{ki} - \bar{\beta}_k)(\delta_{li} - \bar{\delta}_l) = o_{\mathrm{p}}\left\{\frac{1}{m}\sum_{i=1}^m (\delta_{li} - \bar{\delta}_l)^2\right\}.$$

Condition 2 assumes the spiked-eigenvalue model (Johnstone, 2001), which ensures that the effects of hidden factors are large enough to be identified by singular value decomposition. Condition 3 comprises the sphericity conditions on nonspiked singular values (Lee et al., 2014). The relative growth rates of m and n play a key role in this condition. For example, when $m^{-1}\sum_{i=1}^m (\sigma_i - \bar{\sigma})^k$ is greater than zero, m must grow at a faster rate than n^2 to satisfy the first condition in Condition 3. The second part of Condition 3 relaxes the assumption on $\phi(2)$ but adds an assumption on $\phi(4)$. Condition 5 requires that the row vectors of B and the row vectors of Δ be asymptotically orthogonal after mean centring.

THEOREM 1. *Suppose that Conditions 1–5 are satisfied. Then the columns of* $\hat{\Gamma}$ *span the same column space as the columns of* Z *with probability* 1, *and* $\hat{\beta}_i - \beta_i = o_{\mathrm{p}}(1)$ *for* $i = 1, \ldots, m$.

Theorem 1 shows that the proposed method produces consistent estimates of β_i and the hidden factors. The proof can be found in the Supplementary Material.

2·3. *Relationship to the restricted least-squares method*

We now show the connection between the proposed method and the restricted least-squares procedure of Greene & Seaks (1991). Suppose that $C\beta = c$ is a linear restriction on β. The restricted least-squares estimator $\hat{\beta}_{\mathrm{RLS}}$ of β is the solution of

$$\text{minimize } (y - X\beta)^{\mathrm{T}}(y - X\beta) \quad \text{subject to } C\beta = c.$$

It can be shown that $\hat{\beta}_{\mathrm{RLS}} = \hat{\beta}^* - H(C\hat{\beta}^* - c)$, where $\hat{\beta}^* = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y_i$ is the ordinary least-squares estimator and $H = (X^{\mathrm{T}}X)^{-1}C^{\mathrm{T}}\{C(X^{\mathrm{T}}X)^{-1}C^{\mathrm{T}}\}^{-1}$.

When estimating β_i, we impose a restriction on β and δ, and hence on β and ψ, such that they are asymptotically orthogonal after mean centring. While this is not a linear restriction as in the restricted least-squares procedure, the similarity of the two approaches can be illustrated as follows. Let $\mathrm{vec}(\cdot)$ be a function on a matrix which stacks the columns of the matrix into one long vector. Then model (2) for all m features can be re-expressed as

$$\mathrm{vec}(Y) = (I_m \otimes X)\mathrm{vec}(B^*) + \mathrm{vec}(E^*),$$

where $B^* = (\beta_1^*, \ldots, \beta_m^*)$, $\otimes$ is the Kronecker product, and I_m is the $m \times m$ identity matrix. We further define $C = \Psi(I - M_J) \otimes I_p$ and $\hat{C} = \hat{\Psi}(I - M_J) \otimes I_p$. The solution of

$$\begin{aligned} &\text{minimize} \quad \{\text{vec}(Y) - (I_m \otimes X)\text{vec}(B^*)\}^{\mathrm{T}}\{\text{vec}(Y) - (I_m \otimes X)\text{vec}(B^*)\} \\ &\text{subject to} \quad \hat{C}\,\text{vec}(B^*) = 0 \end{aligned}$$

is $\text{vec}(\hat{B}_{\text{RLS}}) = \text{vec}(\hat{B}^*) - H\hat{C}\,\text{vec}(\hat{B}^*)$, where

$$\begin{aligned} H\hat{C} &= \{(I_m \otimes X)^{\mathrm{T}}(I_m \otimes X)\}^{-1}\hat{C}^{\mathrm{T}}[\hat{C}\{(I_m \otimes X)^{\mathrm{T}}(I_m \otimes X)\}^{-1}\hat{C}^{\mathrm{T}}]^{-1}\hat{C} \\ &= (I - M_J)\hat{\Psi}^{\mathrm{T}}\{\hat{\Psi}(I - M_J)\hat{\Psi}^{\mathrm{T}}\}^{-1}\hat{\Psi}(I - M_J) \otimes I_p. \end{aligned}$$

Now $\text{vec}(\hat{B}_{\text{RLS}})$ can be written as

$$\text{vec}(\hat{B}_{\text{RLS}}) = \text{vec}(\hat{B}^*) - \left[(I - M_J)\hat{\Psi}^{\mathrm{T}}\{\hat{\Psi}(I - M_J)\hat{\Psi}^{\mathrm{T}}\}^{-1}\hat{\Psi}(I - M_J) \otimes I_p\right]\text{vec}(\hat{B}^*),$$

which leads to

$$\hat{B}_{\text{RLS}} = \hat{B}^* - \hat{B}^*(I - M_J)\hat{\Psi}^{\mathrm{T}}\{\hat{\Psi}(I - M_J)\hat{\Psi}^{\mathrm{T}}\}^{-1}\hat{\Psi}(I - M_J). \tag{7}$$

Clearly, $\hat{B}$ in (6) and $\hat{B}_{\text{RLS}}$ in (7) are identical if $\hat{\Psi}(I - M_J) = \hat{\Psi}$. Hence the proposed method and the restricted least-squares method are identical if the row means of $\hat{\Psi}$ are zero. Gene expression data are commonly normalized so that the row means of Y are equal (Bolstad et al., 2003); this makes the row means of the residual matrix R, and consequently the row means of $\hat{\Psi}$, all zero. Thus, this zero-mean condition is easily satisfied by gene expression data. If $\hat{\Psi}(I - M_J) \neq \hat{\Psi}$, then $\hat{B}$ and $\hat{B}_{\text{RLS}}$ will be different.

Since $\hat{C}$ is a random matrix, our procedure is not the same as the restricted least-squares procedure, which assumes that the restriction matrix C is fixed. However, the discussion above highlights the similarity between the two approaches. The restricted least-squares estimator can have smaller mean squared error than the ordinary least-squares estimator if the restriction is satisfied (Greene & Seaks, 1991). From our simulations, we observe that our estimators tend to have smaller mean squared errors than the estimators from the true regression model, where the restriction is not utilized.

3. NUMERICAL STUDIES

3·1. *Simulation studies*

We performed simulations to compare the proposed approach with existing methods in a wide range of scenarios. For each simulated dataset, 5000 features and 100 samples were generated from the regression model

$$y_{ji} = \beta_i x_j + z_j^{\mathrm{T}}\delta_i + \epsilon_{ji} \quad (j = 1, \ldots, 100;\ i = 1, \ldots, 5000),$$

where ϵ_{ji} was generated from $N(0, \sigma_i^2)$ with σ_i^2 following an IG$(10, 9)$ distribution, which yields $E(\sigma_i^2) = 1$ and $\text{var}(\sigma_i^2) = 0{\cdot}125$.

Table 1. *Simulation parameters; the total number of simulation settings is* 864

Parameter	Values
μ_z	0, 0·5, 1
Percentage of nonzero δ	10, 20, 40, 60
Percentage of nonzero β	5, 10, 20
Overlap among nonzero δ	total overlap, independent
Number of hidden factors	2, 4
Type of X	binary, continuous
Correlation between nonzero β and nonzero δ	0, 0·4, 0·7

A total of 864 simulation settings are summarized in Table 1. The binary and continuous x_j were simulated respectively from

$$x_j = \begin{cases} 0, & j \leqslant 50, \\ 1, & j > 50, \end{cases} \qquad x_j = \begin{cases} N(-1, 0{\cdot}5), & j \leqslant 50, \\ N(1, 0{\cdot}5), & j > 50, \end{cases}$$

and the first two hidden factors were simulated from

$$z_{j1} \sim \begin{cases} N(\mu_z, 1), & j \leqslant 50, \\ N(-\mu_z, 1), & j > 50, \end{cases} \qquad z_{j2} \sim \mathrm{Ber}(0{\cdot}5).$$

When the number of hidden factors is four, i.e., $q = 4$, (z_{j3}, z_{j4}) were independently generated from $N(0, 1)$. The parameter μ_z determines the correlation between the primary variable, x_j, and the first hidden factor, z_1. Three different values of μ_z were considered: 0, 0·5 and 1. The regression coefficients β and δ were generated from the distribution

$$\begin{pmatrix} \beta_i \\ \delta_{i1} \\ \delta_{i2} \\ \vdots \\ \delta_{iq} \end{pmatrix} = \begin{pmatrix} a_{i0}\zeta_{i0} \\ a_{i1}\zeta_{i1} \\ a_{i2}\zeta_{i2} \\ \vdots \\ a_{iq}\zeta_{iq} \end{pmatrix}, \quad \begin{pmatrix} a_{i0} \\ a_{i1} \\ a_{i2} \\ \vdots \\ a_{iq} \end{pmatrix} \sim N\left\{ 0, \begin{pmatrix} 1 & \rho & 0 & \cdots & 0 \\ \rho & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \right\}$$

where ζ_{il} $(l = 0, \ldots, q)$ are indicator variables, $\zeta_{il} = 0$ or 1, that determine which of the β and δ have nonzero values. To mimic real biological data where the primary variables and hidden factors are not associated with all features, we assumed that 5%, 10% or 20% of the β_i were nonzero and that 10%, 20%, 40% or 60% of the δ_{ik} were nonzero. The value of ζ_{i0} was independently assigned. For ζ_{il} $(l = 1, \ldots, q)$, we considered situations in which nonzero δ values were totally overlapping, $\zeta_{i1} = \cdots = \zeta_{iq}$, or independently selected. In the first situation, each feature either had no associated hidden factors or was associated with all q hidden factors. The correlation between the nonzero β_i and the nonzero δ_{i1}, namely ρ, was set to $\rho = 0, 0{\cdot}4$ and $0{\cdot}7$, representing scenarios in which Condition 5 ranged from being satisfied to severely violated.

Nine different methods were compared: direct surrogate variable analysis; regression model (1), where the hidden factors are assumed to be known and included in the analysis; a no-adjustment model, i.e., regression model (2), where the hidden factors are ignored in the analysis;

the iteratively reweighted surrogate variable analysis of Leek & Storey (2008); the two-step surrogate variable analysis of Leek & Storey (2007); principal component analysis on the residuals; principal component analysis on the original measurements of the features; latent effect adjustment after primary projection (Sun et al., 2012); and four-step remove unwanted variation (Gagnon-Bartsch et al., 2017). Latent effect adjustment after primary projection uses an outlier detection approach after initial data projection to adjust for hidden factors. We treated the second method as a gold standard. In both principal component analyses, top principal components were selected and treated as surrogate variables.

For the four-step remove unwanted variation method, we assumed that 6% of features were negative control genes, close to the proportion of housekeeping genes in the genome (Gagnon-Bartsch & Speed, 2012). We considered situations in which negative control genes were selected only among features with $\beta_i = 0$, i.e., high-quality control genes, or were randomly selected among all features, i.e., poor-quality control genes. In the second case, the assumption of negative control genes was violated. A method to estimate the number of surrogate variables for the four-step remove unwanted variation method has been developed (Gagnon-Bartsch et al., 2017). We used this method in conjunction with the method of Buja & Eyuboglu (1992) to estimate q for the four-step remove unwanted variation method; for all other methods, the approach of Buja & Eyuboglu (1992) was used to estimate q.

For each simulation set-up, 200 datasets were generated and the performance of each method was evaluated based on (i) empirical false discovery rates, where the significant findings were determined by the Benjamini & Hochberg (1995) procedure for a targeted false discovery rate of 0·05; (ii) the mean squared errors of the β_i; and (iii) the area under the receiver operating characteristic curve. For calculation of the false discovery rate and the area under the receiver operating characteristic curve, we define true and false positives as follows. If $\beta_i \neq 0$ and a statistical test for $\beta_i = 0$ is significant after applying the Benjamini–Hochberg procedure, it is a true positive. If the test is significant when $\beta_i = 0$, it is a false positive. In addition to the mean false discovery rates, we calculated the proportion of datasets with an empirical false discovery rate greater than 0·5.

Figure 1 shows simulation results from a scenario where Condition 5 was satisfied, i.e., $\rho = 0$. Direct surrogate variable analysis performed well, as the observed area under the receiver operating characteristic curve and the mean squared errors are similar to those obtained from the approach assuming that the hidden factors are known, and the observed false discovery rates are only slightly inflated in a few simulation settings. Among 288 simulation settings, only four had mean false discovery rates higher than 0·1. As expected, the no-adjustment and principal component analysis-based approaches performed very poorly. When the negative control gene assumption was satisfied, the remove unwanted variation method performed only slightly worse than direct surrogate variable analysis: in ten of the simulation settings, the mean empirical false discovery rates were larger than 0·1; however, when this assumption was violated, the remove unwanted variation method had substantially inflated false discovery rates. The latent effect adjustment after primary projection method had mean empirical false discovery rates above 0·3 in some simulation settings. Since this method was not developed to estimate β, we did not obtain the mean squared errors. When the method developed for four-step remove unwanted variation was used to estimate the number of surrogate variables, the overall performance of the remove unwanted variation method declined substantially, indicating that the method of Buja & Eyuboglu (1992) performs better in estimating the number of surrogate variables. We compared different approaches to estimating q, and the method of Buja & Eyuboglu (1992) outperformed the others; see the Supplementary Material. In Fig. 2, we directly compare the two top-performing approaches: direct surrogate variable analysis and the four-step remove unwanted

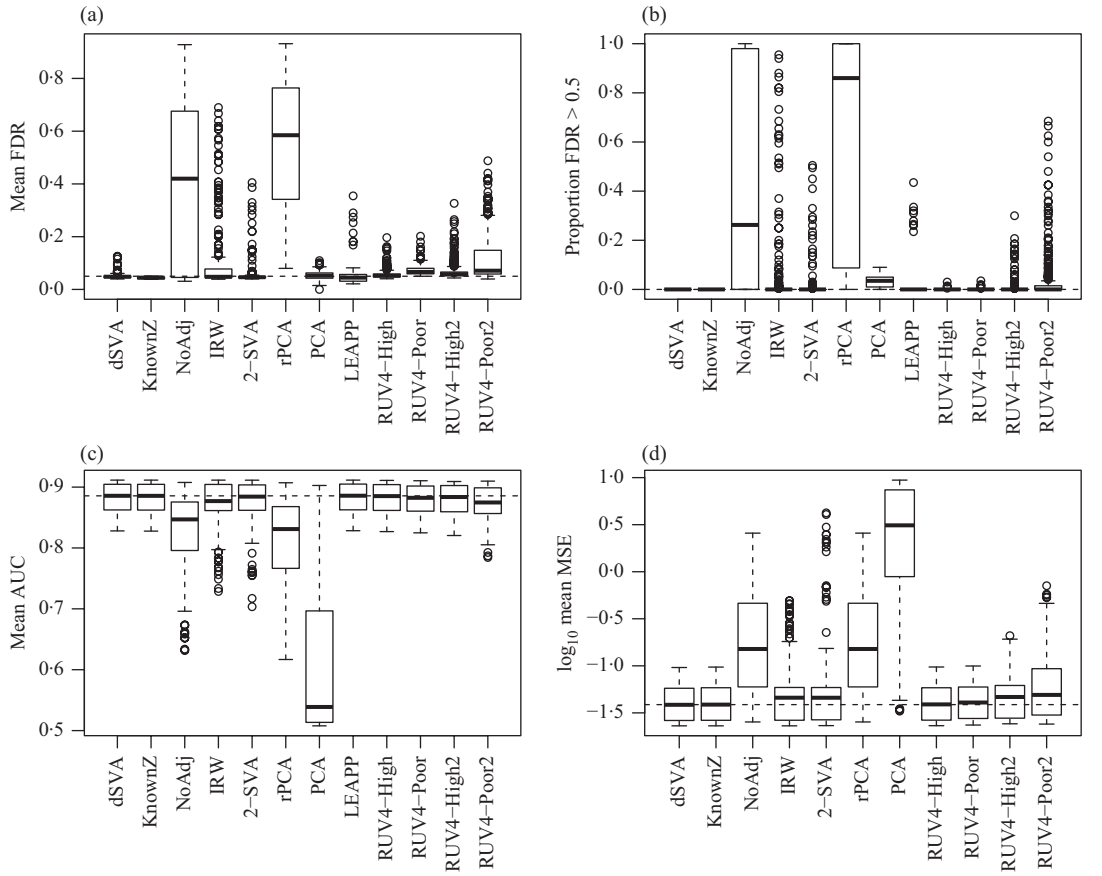

Fig. 1. Comparisons of the proposed and competing methods when $\rho = 0$. Each bar summarizes results from 288 different simulation settings, and in each setting 200 datasets were generated to calculate: (a) mean empirical false discovery rates, FDR; (b) the proportion of datasets with empirical FDR higher than 0·5; (c) the mean area under the receiver operating characteristic curve, AUC; and (d) the mean squared errors, MSE. The methods compared are: dSVA, direct surrogate variable analysis; KnownZ, hidden factors known and included in the model; NoAdj, no adjustment for hidden factors; IRW, iteratively reweighted surrogate variable analysis; 2-SVA, two-step surrogate variable analysis; rPCA, principal component analysis on the residuals; PCA, principal component analysis on the original measured features; LEAPP, latent effect adjustment after primary projection; RUV4-High, four-step remove unwanted variation method with high-quality control genes; RUV4-Poor, four-step remove unwanted variation method with poor-quality control genes; RUV4-High2, RUV4-High with $\hat{q}$ from Gagnon-Bartsch et al. (2017); RUV4-Poor2, RUV4-Poor with $\hat{q}$ from Gagnon-Bartsch et al. (2017).

variation method with high-quality control genes. Direct surrogate variable analysis clearly does better in controlling the false discovery rates.

To investigate the effect of each simulation parameter on the performance of the methods when $\rho = 0$, we created plots for each parameter value. Since the no-adjustment and principal component analysis-based methods performed substantially worse than the other methods, we did not include them in these plots. Among the parameters, μ_z and the percentage of nonzero δ had large effects on the performance of some methods. Figure 3(a) shows boxplots of the false discovery rates with different μ_z values. The iteratively reweighted and two-step surrogate variable analysis approaches had well-controlled false discovery rates when $\mu_z = 0$ and 0·5, but had inflated rates when $\mu_z = 1$. Therefore, these two methods cannot efficiently estimate the hidden factors in the presence of a strong correlation between X and Z. Figure 3(b) shows that when the percentage of nonzero δ was 10%, direct surrogate variable analysis had slightly inflated

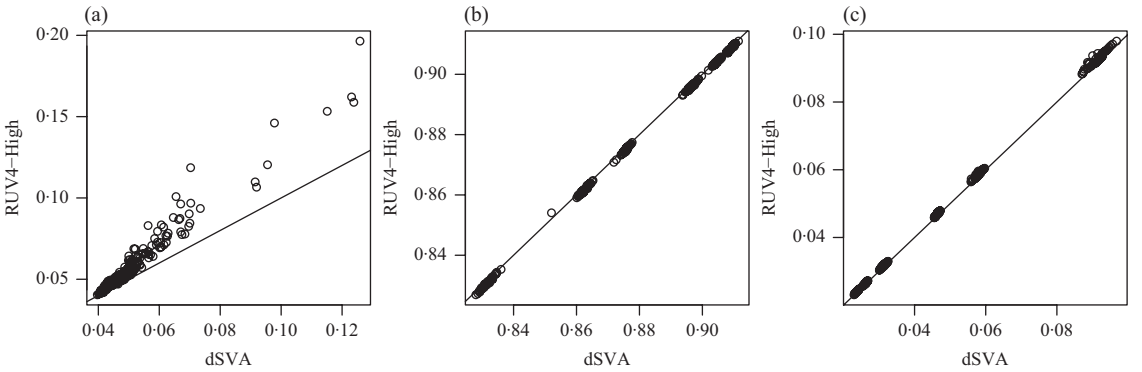

Fig. 2. Comparison of direct surrogate variable analysis, dSVA, and the four-step remove unwanted variation method with high-quality control genes, RUV4-High, when $\rho = 0$: (a) mean empirical false discovery rate; (b) mean area under the receiver operating characteristic curve; (c) mean squared errors.

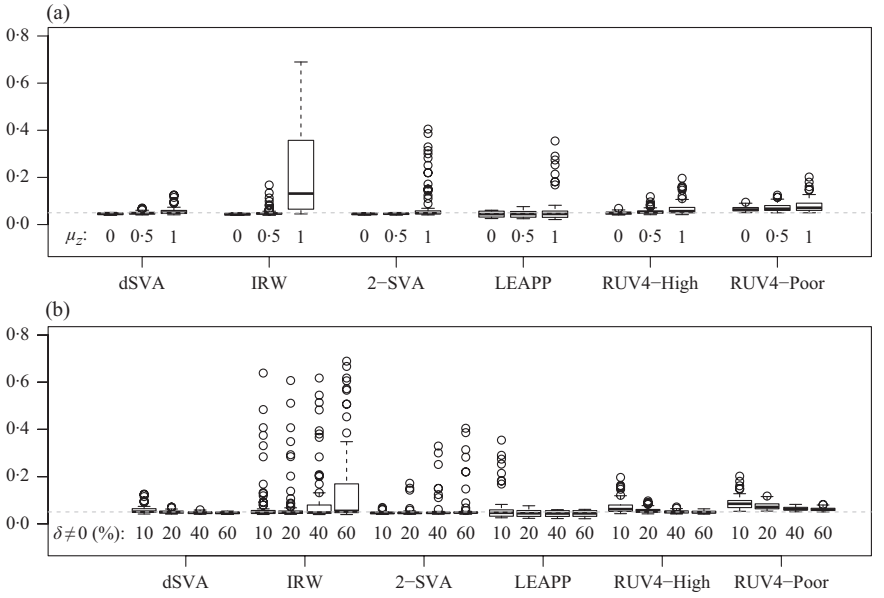

Fig. 3. Comparison of mean empirical false discovery rates when $\rho = 0$ for: (a) $\mu_z = 0, 0{\cdot}5$ or 1; and (b) different proportions of nonzero δ; 10%, 20%, 40% or 60%. In each simulation setting, 200 datasets were generated to obtain the mean empirical false discovery rates. The methods compared are: dSVA, direct surrogate variable analysis; IRW, iteratively reweighted surrogate variable analysis; 2-SVA, two-step surrogate variable analysis; LEAPP, latent effect adjustment after primary projection; RUV4-High, four-step remove unwanted variation method with high-quality control genes; RUV4-Poor, four-step remove unwanted variation method with poor-quality control genes.

false discovery rates, perhaps because direct surrogate variable analysis uses all features, instead of selecting features with nonzero δ. Since the four-step remove unwanted variation approach uses a small fraction of features to estimate the hidden factors, it had more inflated false discovery rates when the percentage of nonzero δ was small. The performances of the different methods in terms of areas under the receiver operating characteristic curves and mean squared errors were largely similar.

Additional simulation results are presented in the Supplementary Material. Our proposed approach was observed to perform well even when q was overestimated and Condition 5 was moderately violated. Overall, our simulation study shows that the proposed method can outperform existing methods in diverse scenarios.

Table 2. *Proportion of variability in year explained by the estimated surrogate variables; for a fair comparison, the same number of surrogate variables was used in all the methods*

Type	Number of surrogate variables	dSVA	IRW	2-SVA	RUV4
EUR vs (JPT + CHI)	25	0·70	0·41	0·64	0·73
JPT vs (EUR + CHI)	25	0·78	0·68	0·72	0·78
CHI vs (EUR + JPT)	25	0·80	0·64	0·71	0·80
JPT vs CHI	16	0·86	0·79	0·79	0·86

EUR, JPT and CHI, individuals of European, Japanese and Chinese ancestry, respectively; dSVA, direct surrogate variable analysis; IRW, iteratively reweighted surrogate variable analysis; 2-SVA, two-step surrogate variable analysis; RUV4, four-step remove unwanted variation method.

3·2. *Application to real data*

We downloaded the Hapmap dataset GSE5859 from the National Center for Biotechnology Information gene expression omnibus website to investigate differentially expressed genes between European and Asian populations (Spielman et al., 2007). This dataset contains 8793 genes, or features, and 208 samples from three continental populations: 102 European, 65 Chinese, and 41 Japanese. The affy package (Gautier et al., 2004) was used for background correction and quantile normalization (Bolstad et al., 2003). In the Supplementary Material we perform an analysis without quantile normalization as a sensitivity analysis. Similar to the original study, we restricted the analysis to 4044 reliably expressed genes in at least 80% of the samples in one population (Spielman et al., 2007).

The original study showed that nearly 70% of genes were differentially expressed across the European and Asian samples (Spielman et al., 2007), but it was subsequently discovered that the calendar year in which each sample was processed was a strong confounding factor (Akey et al., 2007; Leek et al., 2010), and many of the positive findings could potentially be false. In this analysis, we considered a scenario where the researchers did not record the calendar year of sample collection and investigated whether the proposed surrogate variable analysis could capture the year effect. We treated year as a categorical response variable and estimated the proportion of variability that can be explained by the surrogate variables.

Table 2 shows the proportion of the variability explained by surrogate variables estimated by four different methods. Since the estimated variability would increase with the number of surrogate variables, for a fair comparison we used $q = 25$ for all the methods, which was estimated by the method of Buja & Eyuboglu (1992). Both direct surrogate variable analysis and the remove unwanted variation method performed well, as 70% and 73% of the variability was explained by the surrogate variables estimated from these respective methods. In contrast, the surrogate variables from the iteratively reweighted and two-step surrogate variable analysis approaches explained only 41% and 64% of the variability in year, respectively. We also considered different combinations of the populations. Direct surrogate variable analysis and the four-step remove unwanted variation method again consistently outperformed the other methods.

Without any hidden variable adjustment, 73% and 65% of genes were found to be differentially expressed between the European and Asian populations at false discovery rates of 0·05 and 0·01, respectively. As pointed out elsewhere, it seems implausible that so many genes would be differentially expressed between the two populations (Akey et al., 2007). When direct surrogate variable analysis was applied, only 29% and 18% of genes were found to be significant at

false discovery rates of 0·05 and 0·01, respectively. Li et al. (2010) have reported that approximately 20% of genes in lymphoblastoid cell lines are differentially expressed between Hapmap2 European and African samples at a false discovery rate of 0·01. Given that the genetic difference between the European and African populations is greater than that between the European and Asian populations, 18% of genes differentially expressed between the European and Asian populations seems a reasonable estimate.

When we applied two-step surrogate variable analysis and the four-step remove unwanted variation method to the Hapmap data, 15% and 18% of genes, respectively, were declared to be differentially expressed between the European and Asian populations at false discovery rate 0·01. In contrast, 65% of genes were found to be significant by iteratively reweighted surrogate variable analysis at the same false discovery rate, indicating that this method fails to identify the effects of the hidden factors. When we included year as a covariate in the regression analysis, only 28 genes, i.e., 0·7% of the tested genes, were significant at false discovery rate 0·01, because year was nearly nested within each population. All Asian samples were processed in 2005 and 2006, but only three European samples were processed in those two years. Among these 28 genes, 15 were significant according to direct surrogate variable analysis. On the other hand, 12 and 14 genes, respectively, were significant by two-step surrogate variable analysis and the four-step remove unwanted variation method.

We carried out an additional analysis using the same dataset to identify genes differentially expressed by gender. Since genes in sex chromosomes can be used as positive control genes, this additional analysis can be used to directly evaluate the performance of each method. The results show that our method had comparable or slightly better performance than the competing methods; see the Supplementary Material for details.

4. DISCUSSION AND CONCLUSION

Surrogate variable analysis was originally proposed for gene expression data, but it has since been applied to epigenetic data as well (Teschendorff et al., 2011; Maksimovic et al., 2015). Recently, surrogate variable analysis has been extended to prediction and clustering problems. For example, Parker et al. (2014) developed frozen surrogate variable analysis to remove batch effects for prediction problems, and Jacob et al. (2016) extended the remove unwanted variation method to unsupervised learning. Direct surrogate variable analysis was mainly developed for differential expression analysis, but it can be extended to other types of -omics data, as well as to prediction problems, by adopting the approaches used in frozen surrogate variable analysis. We leave such extensions for future research.

One key assumption of the proposed method is Condition 5, which requires that the vector of β values across m genes and the vector of δ values across m genes be asymptotically orthogonal after mean centring. We think that this is reasonable for many biomedical datasets. In our real-data analysis, for example, batch effects are purely technical issues and their effect sizes would not be correlated with those of population differences. Moreover, our method is robust with respect to moderate violations of this condition. In simulation studies, for instance, our method shows better false discovery rate control than the competing methods when $\rho = 0{\cdot}4$. A similar assumption is implicitly used in existing methods. For example, in their simulation studies, Sun et al. (2012) generated β and δ independently. They also suggested that when β and δ are correlated, it will be difficult to identify β.

Principal component analysis was used to correct for batch effects and the effects of hidden confounders prior to the introduction of surrogate variable analysis. This approach has proven

very successful for genome-wide association studies (Price et al., 2006). However, our simulation results show that naïve use of principal components for hidden factor adjustment can result in severe power loss, because the top principal components identified can be highly correlated with the primary variables when the effects of the primary variables are not too weak. When this is the case, principal component analysis should be avoided.

Acknowledgement

We thank the editor, associate editor and referees for their valuable comments and suggestions, which have greatly helped to improve the quality of the paper. This research was supported by the U.S. National Institutes of Health.

Supplementary material

Supplementary material available at *Biometrika* online includes proofs of the theoretical results as well as additional simulation and real-data analysis results.

References

Akey, J. M., Biswas, S., Leek, J. T. & Storey, J. D. (2007). On the design and analysis of gene expression studies in human populations. *Nature Genet.* **39**, 807–8.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Statist. Soc.* B **57**, 289–300.

Bolstad, B. M., Irizarry, R. A., Åstrand, M. & Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185–93.

Buja, A. & Eyuboglu, N. (1992). Remarks on parallel analysis. *Mult. Behav. Res.* **27**, 509–40.

Chakraborty, S., Datta, S. & Datta, S. (2012). Surrogate variable analysis using partial least squares (SVA-PLS) in gene expression studies. *Bioinformatics* **28**, 799–806.

Dumeaux, V., Olsen, K. S., Nuel, G., Paulssen, R. H., Børresen-Dale, A. L. & Lund, E. (2010). Deciphering normal blood gene expression variation—The NOWAC postgenome study. *PLoS Genet.* **6**, e1000873.

Friguet, C., Kloareg, M. & Causeur, D. (2009). A factor model approach to multiple testing under dependence. *J. Am. Statist. Assoc.* **104**, 1406–15.

Gagnon-Bartsch, J. A., Jacob, L. & Speed, T. P. (2017). *Removing Unwanted Variation: Exploiting Negative Controls for High Dimensional Data Analysis*. IMS Monographs. Cambridge: Cambridge University Press, in press.

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

Gautier, L., Cope, L., Bolstad, B. M. & Irizarry, R. A. (2004). Affy-analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **20**, 307–15.

Greene, W. H. & Seaks, T. G. (1991). The restricted least squares estimator: A pedagogical note. *Rev. Econ. Statist.* **73**, 563–7.

Jacob, L., Gagnon-Bartsch, J. A. & Speed, T. P. (2016). Correcting gene expression data when neither the unwanted variation nor the factor of interest are observed. *Biostatistics* **17**, 16–28.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29**, 295–327.

Johnstone, I. M. & Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Am. Statist. Assoc.* **104**, 682–93.

Jung, S. & Marron, J. S. (2009). PCA consistency in high dimension, low sample size context. *Ann. Statist.* **37**, 4104–30.

Lee, S., Zou, F. & Wright, F. A. (2010). Convergence and prediction of principal component scores in high-dimensional settings. *Ann. Statist.* **38**, 3605–29.

Lee, S., Zou, F. & Wright, F. A. (2014). Convergence of sample eigenvalues, eigenvectors, and principal component scores for ultra-high dimensional data. *Biometrika* **101**, 484–90.

Leek, J. T. (2011). Asymptotic conditional singular value decomposition for high-dimensional genomic data. *Biometrics* **67**, 344–52.

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. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* **3**, e161.

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

Li, J., Liu, Y., Kim, T., Min, R. & Zhang, Z. (2010). Gene expression variability within and between human populations and implications toward disease susceptibility. *PLoS Comp. Biol.* **6**, e1000910.

Listgarten, J., Kadie, C., Schadt, E. E. & Heckerman, D. (2010). Correction for hidden confounders in the genetic analysis of gene expression. *Proc. Nat. Acad. Sci.* **107**, 16465–70.

Maksimovic, J., Gagnon-Bartsch, J. A., Speed, T. P. & Oshlack, A. (2015). Removing unwanted variation in a differential methylation analysis of Illumina HumanMethylation450 array data. *Nucleic Acids Res.* **43**, e106.

Parker, H. S., Bravo, H. C. & Leek, J. T. (2014). Removing batch effects for prediction problems with frozen surrogate variable analysis. *PeerJ* **2**, e561.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genet.* **38**, 904–9.

R Development Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, http://www.R-project.org.

Spielman, R. S., Bastone, L. A., Burdick, J. T., Morley, M., Ewens, W. J. & Cheung, V. G. (2007). Common genetic variants account for differences in gene expression among ethnic groups. *Nature Genet.* **39**, 226–31.

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

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.

Teschendorff, A. E., Menon, U., Gentry-Maharaj, A., Ramus, S. J., Weisenberger, D. J., Shen, H., Campan, M., Noushmehr, H., Bell, C. G., Maxwell, A. P. et al. (2010). Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome Res.* **20**, 440–6.

Teschendorff, A. E., Zhuang, J. & Widschwendter, M. (2011). Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. *Bioinformatics* **27**, 1496–505.

[*Received on* 5 *August* 2015. *Editorial decision on* 27 *January* 2017]