Biometrika (2019), **106**, 2, *pp.* 417–432
Printed in Great Britain
doi: 10.1093/biomet/asz010
Advance Access publication 22 April 2019

Sufficient direction factor model and its application to gene expression quantitative trait loci discovery

By F. JIANG

Department of Statistics, The University of Hong Kong, Pokfulam Road, Hong Kong
feijiang@hku.hk

Y. MA

Department of Statistics, Penn State University, 326 Thomas Building, University Park, Pennsylvania 16802, U.S.A.
yzm63@psu.edu

and Y. WEI

Department of Biostatistics, Columbia University, 722 West 168th St, New York, New York 10032, U.S.A.
yw2148@columbia.edu

Summary

Rapid improvement in technology has made it relatively cheap to collect genetic data, however statistical analysis of existing data is still much cheaper. Thus, secondary analysis of single-nucleotide polymorphism, SNP, data, i.e., reanalysing existing data in an effort to extract more information, is an attractive and cost-effective alternative to collecting new data. We study the relationship between gene expression and SNPs through a combination of factor analysis and dimension reduction estimation. To take advantage of the flexibility in traditional factor models where the latent factors are not required to be normal, we recommend using semiparametric sufficient dimension reduction methods in the joint estimation of the combined model. The resulting estimator is flexible and has superior performance relative to the existing estimator, which relies on additional assumptions on the latent factors. We quantify the asymptotic performance of the proposed parameter estimator and perform inference by assessing the estimation variability and by constructing confidence intervals. The new results enable us to identify, for the first time, statistically significant SNPs concerning gene-SNP relations in lung tissue from genotype-tissue expression data.

Some key words: Dimension reduction; Factor model; High dimension; Nonparametric method; Semiparametric method.

1. Introduction

Gene expression quantitative trait loci, eQTLs, are genetic variants that may explain variations in gene expression. Identifying eQTLs is an important area in genetics because it is the only way to understand how genetic variants function at the molecular level (Nica & Dermitzakis, 2013); it is also the most prominent way of discovering gene regulation networks (Gilad et al., 2008). Many genomic findings rely on eQTLs for meaningful interpretation. For example, numerous genome-wide association studies have been performed in recent years to identify genetic variants

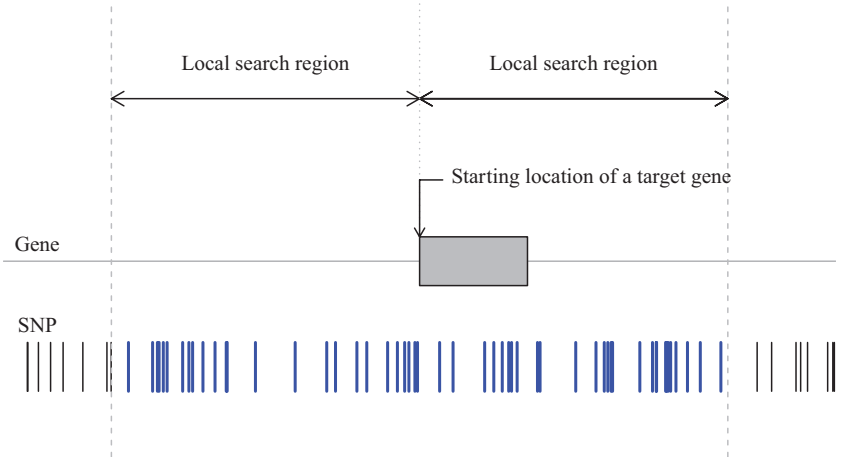

Fig. 1. Data structure of an eQTL study.

associated with complex diseases (Visscher et al., 2012; Lee et al., 2014). Such efforts have resulted in more than 2000 disease-associated variants being identified. However, most of them are noncoding, so their links to underlying diseases are likely to be through regulating gene expression. Consequently, understanding how these genetic variants are associated with gene expression is essential for interpreting these disease-associated variants.

Figure 1 illustrates the typical data structure of an eQTL study. The grey box indicates the location of a target gene in the genome, whose expression levels are measured either by microarray (Schena et al., 1995) or by more recently developed RNA sequencing techniques (Wang et al., 2009). The vertical bars underneath represent the locations of a set of pre-identified candidate SNPs within a local region centred at the gene, delineated by the vertical dashed lines. More candidate SNPs will be included as the width of the search window expands. Typical choices of window length include 20 kb, 100 kb and 1 Mb. The goal of eQTL studies is to identify which candidate SNPs are significantly associated with the target gene expression. To further illustrate the eQTL analysis, we retrieve a subset of data from a genotype-tissue expression pilot dataset collected in one of the major international projects on eQTL discovery (Lonsdale et al., 2013). The data are available at `https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000424.v5.p1`, and include samples of lung tissue collected from 278 subjects. For each tissue sample, the gene expression levels were measured across the entire genome using RNA-Seq technology; the tissue was also genotyped across the entire genome. We randomly select a gene, ENSG00000225880.4, and denote by Y_i its expression level in the ith tissue sample ($i = 1, \ldots, 278$). Using a window length of 20 kb, we identified 117 candidate SNPs, and use $X_i = (X_{i,1}, X_{i,2}, \ldots, X_{i,p})^{\mathrm{T}}$ to denote the genotypes of these SNPs in the ith tissue sample. The goal of the analysis is to identify the genetic variations across individuals that could explain the different gene expression levels in their lung tissue.

Due to the high cost of measuring gene expression levels, most eQTL data have limited sample size. Constrained by the limited sample sizes, most eQTL analyses are conducted separately for each gene-SNP pair (Ardlie et al., 2015). Typically, at any one time researchers will study the mean gene expression level given a single SNP while ignoring other covariates. Previous studies using this approach have identified the eQTLs in lymphocytes and in the adrenal gland, thyroid, arterial and skin tissue for the gene ENSG00000225880.4; see the reported single-tissue eQTLs at `http://gtexportal.org/`. When we applied this kind of modelling and estimation approach to the lung tissue data in § 5, we identified 76 significant SNPs after the

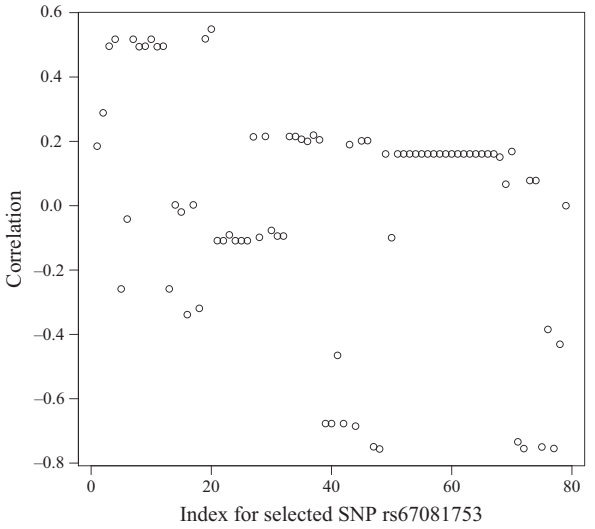

Fig. 2. Correlation between the unselected SNPs and a selected SNP, rs67081753.

Bonferroni correction. However, because the analysis ignores the correlation between the SNPs, it cannot control the overall false discovery rate. For example, two strongly associated SNPs can be selected together, although they explain the same variation in the gene expression. To overcome this deficiency, recent efforts in eQTL analysis have focused on combined analysis of all the genetic variations, see Kendziorski et al. (2006) and Gelfond et al. (2007) for examples. In addition, it is unlikely that gene expression can be regulated by a single SNP, or by multiple SNPs independently. Therefore, it is also desirable to estimate the joint associations between the gene expression Y and all of the candidate SNPs. However, it is challenging to estimate a full model, as p could be comparable to or even much greater than the sample size n.

Current practice in the high-dimensional setting relies on putting penalties on the parameters and assuming sparsity of the data. Following this general approach, we implemented several penalization methods. Again, none of them yielded satisfactory results. For example, we used the adaptive lasso method (Zou, 2006) to select a subset of SNPs and computed the correlation between the selected and unselected SNPs. The correlation can be as high as 0.8, as seen in Fig. 2, which casts doubt on the sparsity assumption. In fact, sparsity may not always hold in practice (De Mol et al., 2008; Giannone et al., 2017), and care needs to be taken before making such an assumption (Barigozzi & Hallin, 2017). The possible violation of the sparsity assumption here motivated us to consider an alternative to the penalization method.

The alternative that we propose is to extend sufficient dimension reduction methods (Li, 1991; Cook, 1998; Ma & Zhu, 2012) under a semiparametric framework to handle high-dimensional challenges in eQTL analysis. With semiparametric modelling, one can further leave the relation between the gene expression and the SNPs unspecified to avoid the risk of model misspecification.

Sufficient dimension reduction has gained much attention since its original introduction by Li (1991). A very comprehensive survey of this area can be found in Cook (1998). See also Ma & Zhu (2013b) for a review of more recent developments. The vast majority of the work in this area deals with a moderate number of covariates, despite the claim of high dimension; this is partly due to the difficulty of the problem. Indeed, when the regression function is left unspecified, finding the sufficient dimension reduction space is already difficult when the number of covariates is moderate. The scarcity of research dealing with very high-dimensional covariates in this framework is also because a sparsity and penalization approach is almost unavoidable

(Tan et al., 2018) if the number of covariates is truly very large, such as in the situation where the number of covariates is larger than the number of observations, even when the sufficient dimension reduction assumption has already been made. This presents a somewhat awkward situation because one of the original goals of sufficient dimension reduction was to provide an alternative approach to handling a large number of covariates, that avoids the popular sparsity assumption and penalization technique.

This awkward situation is successfully avoided by a creative modelling approach proposed recently by Fan et al. (2017), which jointly uses the factor model and the sufficient dimension reduction technique to analyse high-dimensional data in a time series context under the linearity condition. We adapt the idea of combining factor analysis and dimension reduction estimation to performing secondary analysis of SNP-gene expression data, and we further relax some of the assumptions of Fan et al. (2017) and enhance their results. Specifically, we relax the distributional requirements on the covariates, rigorously establish the identifiability property, and establish first-order asymptotic properties that enable inference.

We study the identifiability of the model in the sense that it is uniquely defined, and we illustrate how the distribution of the covariate vector affects model identifiability. Bai & Ng (2013) claimed that a factor model is computationally identifiable if the number of unknown parameters is the same as the number of equations to solve. We further establish that under the same condition, the model is asymptotically identifiable as $p, n \to \infty$.

In addition, we establish the asymptotic normality of our estimators obtained from combining the factor analysis and the sufficient dimension reduction models, and derive their asymptotic variances, which were not provided in Fan et al. (2017), for forecasting. This result allows us to perform statistical inference and calculate p-values, which is crucial in identifying significant SNPs. Furthermore, the detailed asymptotic analysis describes how the ratio p/n affects the estimation variance. Taking advantage of the double robustness property, the estimation variance in estimating the sufficient direction is not inflated by the estimation error in the factor analysis when p grows sufficiently faster than n. A similar result was also shown in Stock & Watson (2002).

2. MODEL SPECIFICATION

Recall that Y_i is the expression level of a target gene from the ith subject ($i = 1, \ldots, n$) and X_i is a p-dimensional vector of covariates, which include the subject's p_1 SNP values within a local region around the target gene, along with p_2 controlling covariates. Let $p = p_1 + p_2$. We assume that the observations (X_i, Y_i) are independent and identically distributed, and that the association between Y_i and X_i is fully captured by a latent factor f_i, i.e., Y_i is independent of X_i when f_i is given. More specifically, let $X_i = (X_{i1}, \ldots, X_{ip})^{\mathrm{T}}$ with

$$X_{il} = b_l^{\mathrm{T}} f_i + u_{il} \quad (1 \leqslant l \leqslant p;\ 1 \leqslant i \leqslant n), \tag{1}$$

where b_l is a q-dimensional vector. The relation in (1) can be written in the matrix form

$$X_i = B f_i + u_i, \tag{2}$$

where $f_i = (f_{i1}, \ldots, f_{iq})^{\mathrm{T}}$ is a q-dimensional vector of factors, $B = (b_1, \ldots, b_p)^{\mathrm{T}}$ is a $p \times q$ deterministic matrix, $u_i = (u_{i1}, \ldots, u_{ip})^{\mathrm{T}}$ and $u = (u_1, \ldots, u_n)$. We require $q < p$ and require u_i to be independent of f_i and $E(u_i) = 0$. Let $F = (f_1, \ldots, f_n)^{\mathrm{T}}$ be an $n \times q$ matrix and $X = (X_1, \ldots, X_n)$ an $p \times n$ matrix. We further consider a sufficient dimension reduction model of the factors $f_1, \ldots, f_n$,

$$Y_i = \psi\left(\beta^{\mathrm{T}} f_i, \epsilon_i\right), \tag{3}$$

where ψ is an unknown function, ϵ_i is a random variable independent of $\beta^{\mathrm{T}} f_i$ and u_i, and β is a $q \times d$ dimensional parameter vector with $d < q$.

Considering jointly (2) and (3), and ignoring the error u_i in (2), $f_i = (B^{\mathrm{T}}B)^{-1}B^{\mathrm{T}}X$, so $\beta^{\mathrm{T}} f_i$ in (3) can be written as $\{B(B^{\mathrm{T}}B)^{-1}\beta\}^{\mathrm{T}}X$. In other words, the covariate effect of X on Y_i can be summarized through $\alpha \equiv B(B^{\mathrm{T}}B)^{-1}\beta$, which essentially allows the reduction of the covariate dimension from p to d. The first p_1 rows of α correspond to the effects of the first p_1 SNPs on the gene expression level in the sufficient direction. We can determine whether the jth SNP in eQTL is significant by testing the null hypothesis $\alpha_{j\cdot} = 0$.

The idea of combining the factor model (1) and the sufficient dimension reduction model (3) when there are a large number of predictors and an unknown link function ψ was first proposed by Fan et al. (2017) in the context of statistical forecasting. The dimension reduction was performed in two steps. First, the dimensionality was reduced from p to q via the high-dimensional factor model (1). Second, using the extracted factors, Fan et al. (2017) developed a link-free sufficient forecasting method based on sliced inverse regression to further reduce the dimension from q to d and to deliver additional predictive power.

The drawback of sliced inverse regression is that it requires the linearity condition on the covariates f_i, i.e., it requires that $E(f_i \mid \beta^{\mathrm{T}} f_i) = \beta(\beta^{\mathrm{T}}\beta)^{-1}\beta^{\mathrm{T}} f_i$ for all f_i. Since the factor model and its subsequent estimation procedure do not rely on such restriction of the factors f_i, and since the f_i themselves cannot be observed directly, it is desirable to allow as much flexibility as possible and avoid any structural assumptions on the distribution of the f_i. Thus, to relax the linearity condition on the latent variables, we adopt the semiparametric approach introduced in Ma & Zhu (2012) for the estimation in the dimension reduction step. The essential idea in relaxing the linearity condition is to reformulate the sliced inverse regression so that it can be written equivalently in an estimating equation form, where the estimating function has a product form. The linearity condition leads to the zero mean of one multiplier of the product form, and one can apply a centring procedure to the other multiplier to achieve zero mean as well if the linearity condition is violated, hence retaining the consistency of the estimating equation. In addition, replacing the linear form of $E(f \mid \beta^{\mathrm{T}} f)$, which is assumed by the linearity condition, with its nonparametric estimate can achieve a double robustness property and reduce estimation variance. The generality of the semiparametric dimension reduction method allows us to study a wide range of sufficient dimension reduction estimates within a unified framework and results in a rich class of estimators, including the classical dimension reduction techniques as special cases. We will show that all the desired properties in Fan et al. (2017) can be achieved without the linearity condition. Most importantly, in addition to the results on convergence order, we derive the specific forms of the asymptotic variances, which were not given in Fan et al. (2017). This is crucial because the calculation of p-values and the identification of statistically significant covariates, i.e., the identification of eQTLs, relies on such properties.

3. Estimation

3.1. *Estimation algorithm*

We first use factor analysis on (1) to obtain $\hat{f}_i$, an estimator for f_i, and then plug $\hat{f}_i$ into (3) in place of f_i and find the sufficient direction by semiparametric dimension reduction techniques. Specifically, the estimation procedure is as follows.

Step 1. Following Fan et al. (2017), we solve the constrained least squares problem

$$(\hat{B}, \hat{F}) = \arg\min_{B,F} \|X - BF^{\mathrm{T}}\|_{\mathrm{F}}^2$$
$$\text{subject to} \quad n^{-1}F^{\mathrm{T}}F = I_q, \quad B^{\mathrm{T}}B \text{ is diagonal}$$

to obtain the estimators $\hat{F} = (\hat{f}_1, \ldots, \hat{f}_n)^{\mathrm{T}}$ for F and $\hat{B} = (\hat{b}_1, \ldots, \hat{b}_p)^{\mathrm{T}}$ for B, where $\hat{f}_i$ and $\hat{b}_l$ are the estimators for f_i and b_l, respectively. This is a classical principal components problem. The estimated factor matrix $\hat{F}$ is $n^{1/2}$ times the eigenvectors corresponding to the q largest eigenvalues of the $n \times n$ matrix $X^{\mathrm{T}}X$, and $\hat{B} = n^{-1}X\hat{F}$ is the corresponding factor loading matrix.

Step 2. Treating the $\hat{f}_i$ as the covariates, following Ma & Zhu (2012) we then solve

$$n^{-1}\sum_{i=1}^{n}\left[g(Y_i, \beta^{\mathrm{T}}\hat{f}_i) - \hat{E}\{g(Y_i, \beta^{\mathrm{T}}\hat{f}_i) \mid \beta^{\mathrm{T}}\hat{f}_i\}\right]\left[\eta(\hat{f}_i) - \hat{E}\{\eta(\hat{f}_i) \mid \beta^{\mathrm{T}}\hat{f}_i\}\right] = 0 \tag{4}$$

for β. The resulting $\hat{\beta}$ is the estimator for β_0. Here, g and η are user-chosen smooth functions, and

$$\hat{E}\{g(Y_i, \beta^{\mathrm{T}}\hat{f}_i) \mid \beta^{\mathrm{T}}\hat{f}_i\} = \frac{\sum_{j=1}^{n} K_h(\beta^{\mathrm{T}}\hat{f}_j - \beta^{\mathrm{T}}\hat{f}_i)g(Y_j, \beta^{\mathrm{T}}\hat{f}_j)}{\sum_{j=1}^{n} K_h(\beta^{\mathrm{T}}\hat{f}_j - \beta^{\mathrm{T}}\hat{f}_i)}$$
$$\hat{E}\{\eta(\hat{f}_i) \mid \beta^{\mathrm{T}}\hat{f}_i\} = \frac{\sum_{j=1}^{n} K_h(\beta^{\mathrm{T}}\hat{f}_j - \beta^{\mathrm{T}}\hat{f}_i)\eta(\hat{f}_j)}{\sum_{j=1}^{n} K_h(\beta^{\mathrm{T}}\hat{f}_j - \beta^{\mathrm{T}}\hat{f}_i)},$$

where for vector $x = (x_1, \ldots, x_d)^{\mathrm{T}}$, $K_h(x) = (1/h^d)\prod_{l=1}^{d} K(x_l/h)$ is a product kernel function with a unified bandwidth h, which only needs to be in the range between $n^{-1/(2d)}$ and $n^{-1/(4m)}$, where m is the kernel order. We assume that we have arranged g and η properly so that $g(Y, \beta^{\mathrm{T}}f)\eta(f)$ is a vector of length $(q-d)d$. Some special choices of g and η lead to semiparametric sliced inverse regression, semiparametric principal Hessian directions, semiparametric sliced average variance estimation, and semiparametric dimension reduction (Ma & Zhu, 2012).

Step 3. Estimate the individual covariate effect α by $\hat{\alpha} \equiv \hat{B}(\hat{B}^{\mathrm{T}}\hat{B})^{-1}\hat{\beta}$, where $\hat{B}$ and $\hat{\beta}$ are, respectively, the estimated factor loading matrix and estimated parameters from the previous steps.

Considering the equivalence between the relations $X = BF^{\mathrm{T}} + U$ and $X^{\mathrm{T}} = FB^{\mathrm{T}} + U^{\mathrm{T}}$, we could also reverse the treatment of B and F as in Stock & Watson (2002) and Fan et al. (2013). We have opted for the current treatment so that we only need to handle an $n \times n$ matrix $X^{\mathrm{T}}X$, instead of a possibly much larger $p \times p$ matrix XX^{T}. Combining the facts that the $\hat{f}_i$, the kernel estimators and the sufficient dimension reduction estimator are consistent, we show the consistency of our proposed procedure in Theorem 1. Moreover, in Theorem 2, we show that the estimation variation in $\hat{f}_i$ and that arising from the kernel estimators do not inflate the variation in $\hat{\beta}$ when $n^{1/2}p^{-1} \to 0$, a condition that is readily satisfied in our setting.

3.2. *Selection of the functions g and η and the tuning parameters*

In Step 2 of the above estimation procedure, users have the freedom to choose the functions g and η. The general requirement is that g be a function of Y and $\beta^{\mathrm{T}}f$ only, that η be a function of

f only, and that they should be sufficiently rich that the dimension of $g\eta$ is at least $(q-d)d$. For example, we could select the components of g to be polynomials of $(Y, \beta^{\mathrm{T}}f)$ and those of η to be monomials of f, i.e., $g = \{Y, f^{\mathrm{T}}\beta, Y^2, Y(f^{\mathrm{T}}\beta), (f^{\mathrm{T}}\beta)\otimes(f^{\mathrm{T}}\beta), \ldots, Y^k, Y^{k-1}(f^{\mathrm{T}}\beta), \ldots, (f^{\mathrm{T}}\beta)\otimes\cdots\otimes(f^{\mathrm{T}}\beta)\}^{\mathrm{T}}$ and $\eta = \{f^{\mathrm{T}}, (f\otimes f)^{\mathrm{T}}, \ldots, (f\otimes\cdots\otimes f)^{\mathrm{T}}\}$. Because the number of parameters in this step is $(q-d)d$, we need the dimension of $g\eta$ to be at least $(q-d)d$. If more than $(q-d)d$ equations are obtained from $g\eta$, we bring the number down to exactly $(q-d)d$ by applying the generalized method of moments. Different choices of g and η will affect the estimation variability of β, while the consistency of the β estimation is retained regardless of the choices of g and η. The optimal choice consists of $g = \partial \log f_{Y|\beta^{\mathrm{T}}f}(y, \beta^{\mathrm{T}}f)/\partial(\beta^{\mathrm{T}}f)$ and taking η to be a $(p-d)$-dimensional subvector of f (Ma & Zhu, 2013a), for which the efficient estimator of β will be obtained. The price associated with the optimal choice is the need to estimate the conditional density function $f_{Y|\beta^{\mathrm{T}}f}(y, \beta^{\mathrm{T}}f)$ and its derivative with respect to $\beta^{\mathrm{T}}f$, so other non-optimal choices are also used in practice. The bandwidth h does not play a critical role, and can be chosen as any value that satisfies Condition 10 in § 4.2. In practice, a common choice is $h = n^{-1/(2m+d)}$, where m is the order of the kernel function.

The proposed method includes determination of the dimension of the latent factor q and the structure dimension d. For the choice of d, Ma & Zhang (2015) proposed a validated information criterion, which selects d consistently through minimizing a validation of the goodness-of-fit.

The selection of the dimension q has been discussed extensively in the literature; see Bai & Ng (2002), Alessi et al. (2010) and Ahn & Horenstein (2013) for the independent data case and Hallin & Liška (2007) and Lam & Yao (2012) for the case of time series data. Compared to the traditional factor analysis, the proposed estimation method is less sensitive to the selection of q, because the subsequent sufficient dimension reduction method refines the dimension reduction by using information from the Y_i. Here, we propose to use the recently proposed criteria described in Ahn & Horenstein (2013).

4. Main results

4.1. *Uniqueness of the model in the ultrahigh-dimensional setting*

The computational identifiability of the linear factor model is usually considered and mostly relies on the equality of the number of unknown parameters and the number of equations constructed (Bai & Ng, 2013). In this work, we show the identifiability of the model in the sense that the true model is uniquely defined. Only after such an identifiability property is established does estimation become meaningful. Otherwise, estimation will not have a well-defined target and it will become unclear what one is estimating.

To fix notation, let $\|W\|_1$ be the 1-norm of an arbitrary matrix W, i.e., the maximum of the absolute column sums. Let $\|W\|_2$ be the 2-norm of the matrix, i.e., the maximum singular value of W or the square root of the maximum eigenvalue of $W^{\mathrm{T}}W$. Let $\|W\|_\infty$ be the sup-norm of the matrix, i.e., the maximum of the absolute row sums. Finally, let $\|W\|_{\mathrm{F}}$ be the Frobenius norm. For the identifiability of B and F as $n, p \to \infty$, we require the following regularity conditions.

Condition 1. There exists a constant M, not depending on p and n, such that $E(\|f_i\|_2^4) \leqslant M$. In addition, $E(f_i) = 0$ and $\mathrm{cov}(f_i) = I_q$, where I_q is the $q \times q$ identity matrix.

Condition 2. Let b_l be deterministic and such that $\|b_l\| \leqslant M$, where M is a constant independent of n and p. The matrix $p^{-1}B^{\mathrm{T}}B$ is diagonal, with distinct positive entries arranged in decreasing order. As $p \to \infty$, $p^{-1}B^{\mathrm{T}}B \to \Sigma_\Lambda$, where Σ_Λ is a $q \times q$ diagonal nonrandom matrix with positive distinct diagonal elements. In addition, the first nonzero element in each column of B is positive.

Condition 3. For some $c > 0$, the loading b_l is such that $\|b_l\|_2 \leqslant c$ for $l = 1, \ldots, p$. As $p \to \infty$, there are two positive constants c_1 and c_2 such that

$$c_1 < \lambda_{\min}(B^{\mathrm{T}}B/p) < \lambda_{\max}(B^{\mathrm{T}}B/p) < c_2.$$

Throughout the paper, $\lambda_{\min}(M)$ and $\lambda_{\max}(M)$ denote the minimum and maximum eigenvalues of a symmetric matrix M, respectively.

Condition 4. The random variables u_i are independent of each other, and each u_i is independent of b_i and f_i, with $E(u_i) = 0$ and $nE(u_{il}^2) \leqslant M$ for all $l = 1, \ldots, p$. Further, for all $i = 1, \ldots, n$, $\sum_{l=1}^{p} E(u_{il}^2) \leqslant M$ and $p^{-1/2} \sum_{l=1}^{p} |u_{il}^2 - E(u_{il}^2)|^4 \leqslant M$.

Condition 5. As $p \to \infty$, $p^{-1/2} \sum_{l=l}^{p} b_l u_{il} \to N(0, \Gamma)$ in distribution, where

$$\Gamma \equiv \lim_{p\to\infty} p^{-1} \sum_{l=1}^{p} \sum_{k=1}^{p} b_l b_k^{\mathrm{T}} E(u_{il}u_{ik})$$

is a bounded variance matrix.

Condition 6. The random variables f_i and u_i are mutually independent conditional on B. In addition, $E(f_i f_i^{\mathrm{T}} u_{il}^2) = \Phi_l$.

Condition 4 implies $\|\Sigma_U\|_1 \leqslant M$, where $\Sigma_U = E(u_i u_i^{\mathrm{T}})$. Conditions 1–4 are needed for the identifiability and consistency of estimation, while Conditions 5 and 6 are needed for the asymptotic distribution of the estimators. We first state the identifiability result as Proposition 1; its proof is given in the Supplementary Material.

PROPOSITION 1. *Under Conditions 1–4, B and F are unique as $p \to \infty$.*

4.2. *Theoretical properties*

Fan et al. (2017) established the consistency of $\hat{\beta}$ when sliced inverse regression is used in the dimension reduction step. Their result requires the linearity condition. We adopt the semiparametric approach and show the consistency of the resulting estimator without imposing such a linearity condition. This is an important step forward, as a key feature of the factor model is that assumptions on the latent factor, including the linearity condition, need not be imposed. In addition, we also show the asymptotic normality and derive the asymptotic variance of the estimators. These results are crucial in genetic studies because they are required for inference, p-value calculation, and selection of the significant SNPs. Our results are established in a very general context and can be readily applied regardless of whether semiparametric sliced inverse regression, semiparametric sliced average variance estimation, semiparametric dimension reduction, semiparametric principal Hessian directions, or any other choices of g and η are used to conduct the second step of the factor analysis and the dimension reduction estimation.

Regularity conditions for the asymptotic properties are as follows.

Condition 7. The univariate kernel function $K(\cdot)$ is Lipschitz and has compact support. It satisfies

$$\int K(v)\,\mathrm{d}v = 1, \quad \int v^t K(v)\,\mathrm{d}v = 0 \ \ (1 \leqslant t \leqslant m-1), \quad 0 \neq \int v^m K(v)\,\mathrm{d}v < \infty.$$

The d–dimensional kernel function is a product of d univariate kernel functions, i.e., $K_h(v) = K(v/h)/h^d = \prod_{l=1}^d K_h(v_l) = \prod_{l=1}^d K(v_l/h)/h^d$ for $v = (v_1, \ldots, v_d)^{\mathrm{T}}$.

Condition 8. The density functions of f_i and $\beta^{\mathrm{T}} f_i$, denoted by $\pi_f(f_i)$ and $\pi(\beta^{\mathrm{T}} f_i)$, are bounded away from zero and infinity. Each entry in the matrices $E\{g(Y_i, \beta^{\mathrm{T}} f_i) g(Y_i, \beta^{\mathrm{T}} f_i)^{\mathrm{T}} \mid \beta^{\mathrm{T}} f_i\}$ and $E\{\eta(f_i)\eta(f_i)^{\mathrm{T}} \mid \beta^{\mathrm{T}} f_i\}$ is locally Lipschitz continuous and bounded from above as a function of $\beta^{\mathrm{T}} f_i$.

Condition 9. Let $r_1(\beta^{\mathrm{T}} f_i) = E\{\eta(f_i) \mid \beta^{\mathrm{T}} f_i\}\pi(\beta^{\mathrm{T}} f_i)$ and $r_2(\beta^{\mathrm{T}} f_i) = E\{g(Y_i, \beta^{\mathrm{T}} f_i) \mid \beta^{\mathrm{T}} f_i\}\pi(\beta^{\mathrm{T}} f_i)$. The mth derivatives of $r_1(\beta^{\mathrm{T}} f_i)$, $r_2(\beta^{\mathrm{T}} f_i)$ and $\pi(\beta^{\mathrm{T}} f_i)$ are locally Lipschitz continuous.

Condition 10. The bandwidth $h = O(n^{-\kappa})$ for $1/(4m) < \kappa < 1/(2d)$.

Condition 11. For the identification of β, further assume that the upper $d \times d$ matrix of β is an identity matrix and the lower $(p-d) \times d$ matrix of β is arbitrary.

Condition 12. Let $E[g(Y_i, \beta^{\mathrm{T}} f_i) - E\{g(Y_i, \beta^{\mathrm{T}} f_i) \mid \beta^{\mathrm{T}} f_i\}][\eta(f_i) - E\{\eta(f_i) \mid \beta^{\mathrm{T}} f_i\}]$ be a smooth function of β that has a unique root for β.

Condition 13. The random vectors f_i, u_i and ϵ_i are mutually independent.

Condition 7 is a typical assumption on the kernel function, where m is usually referred to as the order of the kernel. Conditions 8 and 9 impose sufficient smoothness requirements on several functions. Condition 10 adds a constraint on the bandwidth related to the kernel order and the dimension d. It can be seen that as long as $d \leqslant 3$, the common second-order kernel function is sufficient. Condition 11 guarantees the identifiability of β (Ma & Zhu, 2013a). Condition 12 ensures the global consistency of $\hat{\beta}$ (White, 1982). Condition 13 contains standard independence assumptions from the factor model and dimension reduction model formulation. These conditions are all moderate and are commonly assumed.

Under the true model, we have

$$E\big[g(Y_i, \beta_0^{\mathrm{T}} f_i) - E\{g(Y_i, \beta_0^{\mathrm{T}} f_i) \mid \beta_0^{\mathrm{T}} f_i\}\big]\big[\eta(f_i) - E\{\eta(f_i) \mid \beta_0^{\mathrm{T}} f_i\}\big] = 0. \tag{5}$$

Therefore, we show the convergence of $\hat{\beta}$ by showing that (4) converges to (5).

We are now ready to establish the main theorems of this article. Theorem 1 gives the consistency property of the sufficient reduction directions, and Theorem 2 further states the asymptotic properties of these directions. Specifically, the asymptotic normality is proven and the asymptotic variance is derived. These results are established under the setting that both the dimension of the covariates and the number of observations are growing, and that the covariate dimension is much larger than the number of observations. These results are new and stronger than those in Fan et al. (2017), and they are derived under more flexible conditions. The proofs are given in the Supplementary Material.

THEOREM 1. *Assume that Conditions 1–12 hold, and let $\hat{\beta}$ satisfy*

$$n^{-1} \sum_{i=1}^{n} \big[g(Y_i, \hat{\beta}^{\mathrm{T}} \hat{f}_i) - \hat{E}\{g(Y_i, \hat{\beta}^{\mathrm{T}} \hat{f}_i) \mid \hat{\beta}^{\mathrm{T}} \hat{f}_i\}\big]\big[\eta(\hat{f}_i) - \hat{E}\{\eta(\hat{f}_i) \mid \hat{\beta}^{\mathrm{T}} \hat{f}_i\}\big] = 0.$$

Then $\hat{\beta} - \beta_0 = o_{\mathrm{p}}(1)$.

THEOREM 2. *Assume that Conditions 1–12 hold, and let* $\hat{\beta}$ *solve*

$$n^{-1}\sum_{i=1}^{n}\big[g(Y_i,\hat{\beta}^{\mathrm{T}}\hat{f}_i)-\hat{E}\{g(Y_i,\hat{\beta}^{\mathrm{T}}\hat{f}_i)\mid\hat{\beta}^{\mathrm{T}}\hat{f}_i\}\big]\big[\eta(\hat{f}_i)-\hat{E}\{\eta(\hat{f}_i)\mid\hat{\beta}^{\mathrm{T}}\hat{f}_i\}\big]=0.$$

Then

$$n^{1/2}\operatorname{vecl}(\hat{\beta}-\beta_0)=T_0^{-1}n^{-1/2}\sum_{i=1}^{n}\big[g(Y_i,\beta_0^{\mathrm{T}}f_i)-E\{g(Y_i,\beta_0^{\mathrm{T}}f_i)\mid\beta_0^{\mathrm{T}}f_i\}\big]\big[\eta(f_i)-E\{\eta(f_i)\mid\beta_0^{\mathrm{T}}f_i\}\big]$$
$$+O_{\mathrm{p}}\{h^m+n^{1/2}h^{2m}+\log^2 n/(n^{1/2}h^d)+p^{-1/2}+n^{1/2}p^{-1}+n^{-1/2}\},$$

where $T_0=E\Big(\partial\big[g(Y_i,\beta_0^{\mathrm{T}}f_i)-E\{g(Y_i,\beta_0^{\mathrm{T}}f_i)\mid\beta_0^{\mathrm{T}}f_i\}\big]\big[\eta(f_i)-E\{\eta(f_i)\mid\beta_0^{\mathrm{T}}f_i\}\big]\big/\partial\operatorname{vecl}(\beta_0)^{\mathrm{T}}\Big)$. *Here* $\operatorname{vecl}(M)$ *denotes the vector formed by concatenating the columns of the lower* $(q-d)\times d$ *portion of a* $q\times d$ *matrix* M.

Therefore, as $n,p\to\infty$ *and* $n^{1/2}p^{-1}\to 0$, $n^{1/2}\operatorname{vecl}(\hat{\beta}-\beta_0)$ *converges to a normal vector with mean* 0 *and variance*

$$\Sigma_\beta=T_0^{-1}E\Big\{\big([g(Y_i,\beta_0^{\mathrm{T}}f_i)-E\{g(Y_i,\beta_0^{\mathrm{T}}f_i)\mid\beta_0^{\mathrm{T}}f_i\}]\big[\eta(f_i)-E\{\eta(f_i)\mid\beta_0^{\mathrm{T}}f_i\}\big]\big)^{\otimes 2}\Big\}(T_0^{-1})^{\mathrm{T}}$$

where $a^{\otimes 2}=aa^{\mathrm{T}}$ *for an arbitrary matrix* a.

Remark 1. For the case where $d=1$ and $\psi(\cdot)$ is linear, Bai (2003) argued empirically that the regression estimators converge to the true values at a root-n rate as $n^{1/2}p^{-1}\to 0$. Here we establish the result rigorously and extend it to the cases where $d>1$ and $\psi(\cdot)$ is an unknown function.

5. NUMERICAL EVALUATION

5.1. *Simulations*

In our simulation studies we let $q=6,d=2$ and $p=50$. We took the sample size to be $n=300$ and repeated our simulation 1000 times. To generate X_{il} from model (1), we consider two cases: in Case I we simulate f_i from a multivariate normal distribution with mean zero and covariance matrix $(\sigma_{ij})_{q\times q}$ where $\sigma_{ij}=0.5^{|i-j|}$; in Case II we simulate f_{i1} and f_{i2} from a multivariate normal distribution with mean zero and covariance matrix $(\sigma_{ij})_{2\times 2}$ where $\sigma_{ij}=0.5^{|i-j|}$. We let $f_{i3}=|f_{i1}+f_{i2}|+f_{i1}\xi_{i1}$ and $f_{i4}=|f_{i1}+f_{i2}|^2+|f_{i2}|\xi_{i2}$, where ξ_{i1} and ξ_{i2} are independently generated from the standard normal distribution, and we generate f_{i5} from a Bernoulli distribution with success probability $\exp(f_{i2})/\{1+\exp(f_{i2})\}$ and f_{i6} from a Bernoulli distribution with success probability $\Phi(f_{i2})$, where Φ is the standard normal distribution function. We centre and normalize F by its mean and covariance so that F satisfies Condition 1. To construct the matrix B, we first generate n samples of p-dimensional random vectors Z_i from a normal distribution with mean zero and covariance matrix Σ_z, where $\Sigma_{zij}=0.5^{|i-j|}$ for $1\leqslant i,j\leqslant p$. Let $Z=(Z_1,\ldots,Z_n)^{\mathrm{T}}$. We perform eigendecomposition on the matrix ZZ^{T}, and retain the $n\times q$ orthogonal matrix E that spans the eigenspace corresponding to the q largest eigenvalues. We form $B=1/6^{1/2}Z^{\mathrm{T}}E$. This construction yields the eigenvalues of $B^{\mathrm{T}}B/p$ in the range $(2,3)$, which ensures that $B^{\mathrm{T}}B=O_{\mathrm{p}}(p)$, as required by Condition 3. To ensure Conditions 4 and 5, we simulate u_{il} from a normal distribution with mean zero and variance $1/(2n)$.

Further, in model (3), we let $\beta_1 = (1, 0, 1, 1, 1, 1)/6^{1/2}$ and $\beta_2 = (0, -1, 1, -1, 1, -1)/6^{1/2}$. We evaluated the performance of the methods on the following models:

(i) $Y_i = (f_i^{\mathrm{T}}\beta_1)/\{0.5 + (f_i^{\mathrm{T}}\beta_2 + 1.5)^2\} + 0.5\epsilon_i$;
(ii) $Y_i = \exp(f_i^{\mathrm{T}}\beta_1) + 2|f_i^{\mathrm{T}}\beta_2 + 1| + 0.1|f_i^{\mathrm{T}}\beta_1|\epsilon_i$;
(iii) $Y_i = (f_i^{\mathrm{T}}\beta_1)^2 + 2|f_i^{\mathrm{T}}\beta_2 + 1| + 0.1(f_i^{\mathrm{T}}\beta_1)^2\epsilon_i$.

Here ϵ_i follows the standard normal distribution. For Case I, we evaluated the semiparametric sliced inverse regression and semiparametric principal Hessian directions under model (i), and evaluated the semiparametric sliced average variance estimation and semiparametric dimension reduction under model (ii). For Case II, we evaluated the semiparametric sliced inverse regression and semiparametric principal Hessian directions under model (i), and evaluated the semiparametric sliced average variance estimation and semiparametric dimension reduction under model (iii). We designed these simulations to limit the resulting Y_i values to within 200 to avoid numerical instability. For each method, the computation time is roughly 10 seconds when the initial value is randomly selected, and 3 seconds when the initial value is near the truth. The above simulation settings are summarized in the Supplementary Material.

We compared the proposed semiparametric method and the original dimensional reduction techniques in terms of the Euclidean distances between the resulting estimators and the true values; the results are shown in the Supplementary Material. We also evaluated the asymptotic performances of the estimators, and report the results in the Supplementary Material. Further, we compared the empirical distribution of the estimator with the normal distribution using the Kolmogorov–Smirnov normal test. The results show that the estimators are close to the true values and that the confidence intervals have coverage probabilities close to the nominal level. In addition, most of the estimators achieve asymptotic normality, with the p-values from the Kolmogorov–Smirnov normal test being less than $0.00625 = 0.05/8$, the bound adjusted for multiple testing.

5.2. *Analysis on eQTL discovery*

In this subsection we illustrate the application of the proposed semiparametric method to eQTL discoveries. Recall that the illustrative genotype-tissue expression dataset contains $n = 278$ subjects. The expression levels of the gene ENSG00000225880.4 in the subjects' lung tissue were measured by the RNA-Seq technique. The subjects were also genotyped on 117 SNPs within 20 kb of the target gene. In addition, we consider 40 controlling covariates, including gender, platform, three principal components of genome-wide gene expressions, and 35 principal components of genome-wide SNPs. These covariates were included in previous genotype-tissue expression analyses to control for population stratification. In total, we have $p = 157$ covariates.

The approach proposed here has the capacity to include all the SNPs in one model which can take the inter-SNP correlations into account. It consequently enhances the power in identifying eQTLs and may provide new insights into the SNP functionals. To apply the proposed method to the genotype-tissue expression data, we first perform a principal component analysis on the 157 covariates. Following the eigenvalue ratio method discussed in Ahn & Horenstein (2013), we compute

$$\tilde{q} = \underset{1 \leqslant q \leqslant 157}{\arg\max} \frac{\log(V_{q-1}/V_q)}{\log(V_q/V_{q+1})}$$

and obtain $\tilde{q} = 5$, where V_q is the average of the first q eigenvalues of the matrix $B^{\mathrm{T}}B$. This suggests picking the first five factors for the second-stage analysis. To be more conservative, we

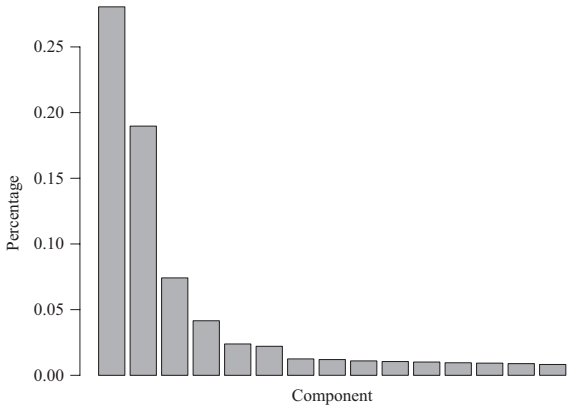

Fig. 3. Bar graph of principal components.

plot in Fig. 3 the percentage of variance explained by each of the first 15 principal components.

We can see that the results for the fifth and the sixth components are very similar. Taking this into account, $q = 6$ seems a reasonable choice in our analysis. To avoid carrying out futile analysis, we perform a test of the null hypothesis that none of the first six factors is related to the response using a method in the spirit of Zhu et al. (2011). To this end, we first calculate $\hat{R}_k = n^{-1}\sum_{j=1}^{n}\{n^{-1}\sum_{i=1}^{n}\hat{f}_{ik}I(Y_i < Y_j)\}^2$ for each estimated factor $\hat{f}_{ik}$ $(k = 1, \ldots, 6)$, resulting in $(\hat{R}_k : k = 1, \ldots, 6) = (0.0061, 0.00026, 0.0037, 0.0022, 0.0009, 0.0002)$. Then we select the threshold by permuting the rows of $\hat{F}$ 100 times. In the lth permutation, let $\tilde{F}_l$ be the permuted $\hat{F}$ and $\tilde{f}_{ikl}$ its (i, k) element, and compute $\tilde{R}_{kl} = n^{-1}\sum_{j=1}^{n}\{n^{-1}\sum_{i=1}^{n}\tilde{f}_{ikl}I(Y_i < Y_j)\}^2$. Then we take $\max_{k,l}\tilde{R}_{kl} = 0.0039$ over the 100 replicates to be our threshold for rejecting the null model that there is no factor with an effect on the response. Clearly, $\hat{R}_1 > 0.0039$. This ensures that at least one factor has an effect on the response even when considered separately. Moreover, we use the validated information criterion proposed in Ma & Zhang (2015) to select the structural dimension d. The validated information criterion values for the four semiparametric dimension reduction methods, i.e., semiparametric sliced inverse regression, semiparametric sliced average variance estimation, semiparametric dimension reduction, and semiparametric principal Hessian directions, are presented in the upper part of Table 1.

The validated information criterion values are smallest at $d = 1$ except for semiparametric sliced average variance estimation, which achieves the minimum at $d = 2$. We adopted majority voting and set $d = 1$ for the model to describe the association between the gene expressions and genetic variants. We subsequently estimated $\hat{\beta}$ using the four semiparametric dimension reduction methods and report the corresponding estimates along with their standard errors in the lower part of Table 1. The four sets of estimation results are similar.

We further compared the semiparametric dimension reduction methods and the classical dimension reduction methods through two-fold crossvalidation. Specifically, we randomly split the data into two equal parts, the training and testing datasets, and computed the mean predictive errors for each method. The averages of the mean predictive errors over 100 random splits were 1.0715, 1.0755, 1.0378, and 1.0294 for the semiparametric sliced inverse regression, semiparametric sliced average variance estimation, semiparametric dimension reduction, and semiparametric principal Hessian directions methods, while they were 1.0985, 1.1170, 1.1043, and 1.1138 for the original sliced inverse regression, sliced average variance estimation, directional regression, and principal Hessian directions methods, respectively. It is clear that the semiparametric methods outperform the classical dimension reduction methods in terms of prediction in this dataset.

Table 1. *Validated information criterion values at* $d = 1, \ldots, 4$ *together with estimates and standard errors under* $d = 1$ *for semiparametric sliced inverse regression, semiparametric sliced average variance estimation, semiparametric principal Hessian directions, and semiparametric dimension reduction in the gene-SNP association analysis*

	S-SIR	S-SAVE	S-DR	S-PHD
		Validated information criterion		
$d = 1$	103.2497	200.1980	118.9829	78.1923
$d = 2$	160.3691	133.8987	144.3405	122.2833
$d = 3$	186.1471	152.0223	150.8336	142.5507
$d = 4$	194.0230	220.2505	186.5501	183.7236
		Estimate (standard error)		
β_{12}	−0.124 (0.151)	−0.058 (0.151)	−0.056 (0.118)	−0.112 (0.295)
β_{13}	−0.539 (0.178)	−0.576 (0.120)	−0.668 (0.193)	−0.465 (0.214)
β_{14}	−0.534 (0.145)	−0.700 (0.248)	−0.674 (0.095)	−0.462 (0.149)
β_{15}	−0.134 (0.144)	−0.266 (0.136)	−0.270 (0.120)	−0.141 (0.149)
β_{16}	−0.055 (0.065)	−0.061 (0.079)	−0.088 (0.135)	−0.021 (0.126)

S-SIR, semiparametric sliced inverse regression; S-SAVE, semiparametric sliced average variance estimation; S-PHD, semiparametric principal Hessian directions; S-DR, semiparametric dimension reduction.

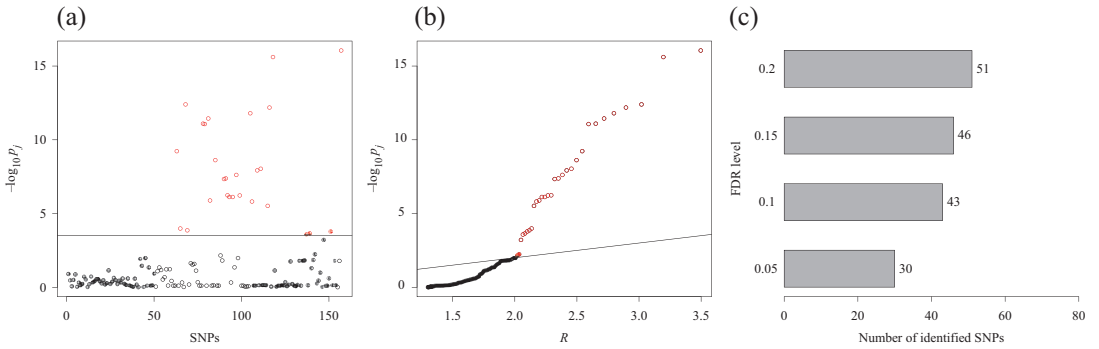

Fig. 4. (a) Base-10 log-transformed p-values ($-\log_{10} p_j$) for each estimated covariate effect. (b) Sorted base-10 log-transformed p-values versus $0.05j/157$. (c) False discovery rate level versus the number of identified SNPs.

As the semiparametric principal Hessian directions method has the best performance, in that it has the smallest mean predictive error as shown in Table 1, we carry out further analysis based on this estimator.

To assess the effect of individual SNPs on the gene expression, we estimate the α coefficients $\hat{\alpha} \equiv \hat{B}(\hat{B}^{\mathrm{T}}\hat{B})^{-1}\hat{\beta}$, where $\hat{B}$ is the factor loading obtained from the first-step principal component analysis. The first p_1 components of vector α correspond to the effects of SNPs on the gene expression level in the sufficient direction. To test the null hypothesis of $\alpha_j = 0$, we calculate the p-values via $p_j \equiv 2[1 - \Phi\{|\hat{\alpha}_j|/\hat{\mathrm{sd}}(\hat{\alpha}_j)\}]$, where $\hat{\alpha}_j$ is the jth component in $\hat{\alpha}$ and Φ is the standard normal distribution function. The $-\log_{10}$ of the resulting p-values are plotted in Fig. 4 and are compared with $-\log_{10}(0.05j/157)$ to adjust for multiple comparisons. As shown in Fig. 4(a), we identified 27 variants at loci in the Supplementary Material which are significantly associated with the gene expression level after Bonferroni correction. These SNPs are also reported in genotype-tissue expression as eQTLs through marginal regressions. Since Bonferroni correction is known to be overly conservative, we further performed an analysis to control the false discovery rate (Benjamini & Hochberg, 1995) to within 0.05 by treating the p-values as independent. We present the results of the false discovery rate-based analysis in Fig. 4(b). Compared with the traditional

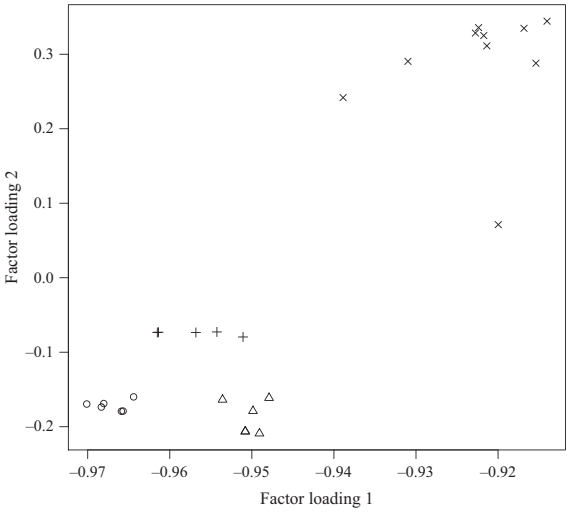

Fig. 5. Scatterplot of the first and second factor loadings of the 27 identified SNPs. ○, Cluster 1; △, Cluster 2; +, Cluster 3; ×, Cluster 4.

pairwise analysis, the proposed joint analysis has great potential for studying the connections among the eQTLs.

To further validate the 27 identified SNPs, we extracted their functional annotation scores across 13 tissue types, including lung, adipose, aorta, liver, brain, intestine, esophagus, pancreas, gastric, heart, ovary, thymus, and spleen. Functional annotation scores were recently developed by Backenroth et al. (2018) to predict the functional effect of noncoding genetic variants in different cell and tissue types. These scores are estimated from independent roadmap datasets (The ENCODE Project Consortium, 2012; Kundaje et al., 2015) and measure the probability of an SNP regulating gene expression in certain cell and tissue types. On average, about 5% of SNPs have functional scores exceeding 0.01 in lung tissue, estimated from the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2012).

Of the 27 SNPs that we identified in the lung tissue, 23 have positive functional annotation scores, which further confirms their function in regulating gene expression in lung tissue. In addition, further investigations of the factor loadings of the identified eQTLs also provide useful insights into how those eQTLs function. Figure 5 shows the distributions of the first and second factor loadings of the 27 SNPs. They naturally cluster the SNPs into four groups. To investigate whether the factor loadings provide a meaningful grouping of the SNPs, we plotted their functional annotation scores across the 13 tissue types by cluster in Fig. 6. We observe distinctive patterns of tissue-specific functional effects across the four clusters. Specifically, clusters 2 and 4 both have a strong effect in lung tissue, but have different effect patterns across other tissues. Cluster 2 has stronger effects in liver, brain, intestine and heart tissue, but the effect of cluster 4 is not strong in any tissue types other than lung. On the other hand, cluster 1 has a moderate effect in lung tissue and strong effects in some other tissues such as adipose, liver, intestine and heart, whereas cluster 3 has a very weak effect in lung tissue but strong effects in brain and thymus tissue. Hence, the factor loadings do provide meaningful groupings, and can help us to understand the underlying potential functional pathways of the identified SNPs.

Remark 2. In our analysis, we used a false discovery rate level of 0.05 to guide the selection of eQTLs. In practice, this level may not be suitable for screening in as many true signals as possible for follow-up studies (Craiu & Sun, 2008). We examine the relationship between the number of

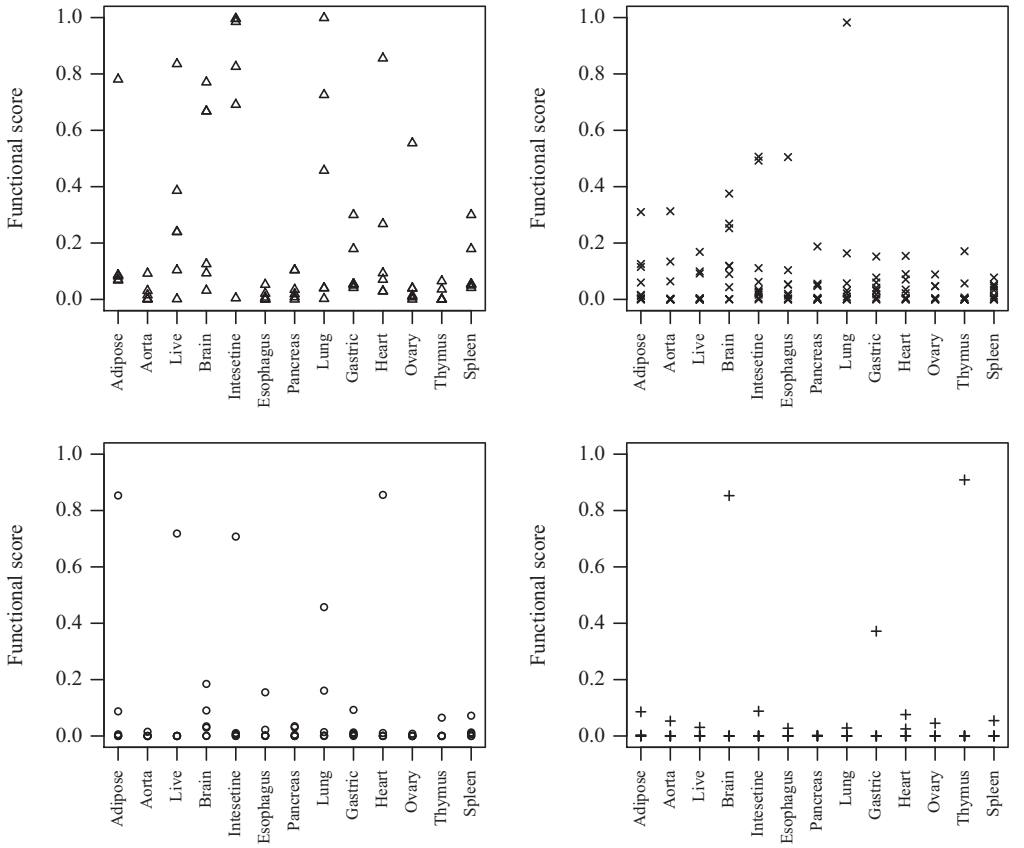

Fig. 6. Functional annotation scores of the 27 SNPs by cluster.

identified SNPs and the false discovery rate level in Fig. 4(c). Researchers should determine the proper false discovery rate level based on their objectives for the follow-up studies.

Acknowledgement

This work was partially supported by the U.S. National Science Foundation, National Institute of Neurological Disorders and Stroke, and National Human Genome Research Institute, and by the Research Grants Council of Hong Kong.

Supplementary material

Supplementary material available at *Biometrika* online includes further details on the simulations.

References

Ahn, S. C. & Horenstein, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica* **81**, 1203–27.

Alessi, L., Barigozzi, M. & Capasso, M. (2010). Improved penalization for determining the number of factors in approximate factor models. *Statist. Prob. Lett.* **80**, 1806–13.

Ardlie, K. K., Dermitzakis, E. T. & GTEx Consortium (2015). The genotype-tissue expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348**, 648–60.

Backenroth, D., He, Z., Kiryluk, K., Boeva, V., Pethukova, L., Khurana, E., Christiano, A., Buxbaum, J. & Ionita-Laza, I. (2018). FUN-LDA: A latent Dirichlet allocation model for predicting tissue-specific functional effects of noncoding variation. *Am. J. Hum. Genet.* **102**, 920–42.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica* **71**, 135–71.

BAI, J. & NG, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70**, 191–221.
BAI, J. & NG, S. (2013). Principal components estimation and identification of static factors. *J. Economet.* **176**, 18–29.
BARIGOZZI, M. & HALLIN, M. (2017). A network analysis of the volatility of high dimensional financial series. *Appl. Statist.* **66**, 581–605.
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.
COOK, D. (1998). *Regression Graphics: Ideas for Studying Regressions through Graphics*. New York: Wiley.
CRAIU, R. V. & SUN, L. (2008). Choosing the lesser evil: Trade-off between false discovery rate and non-discovery rate. *Statist. Sinica* **18**, 861–79.
DE MOL, C., GIANNONE, D. & REICHLIN, L. (2008). Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components? *J. Economet.* **146**, 318–28.
FAN, J., LIAO, Y. & MINCHEVA, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *J. R. Statist. Soc.* B **75**, 603–80.
FAN, J., XUE, L. & YAO, J. (2017). Sufficient forecasting using factor models. *J. Economet.* **201**, 292–306.
GELFOND, J. A., IBRAHIM, J. G. & ZOU, F. (2007). Proximity model for expression quantitative trait loci (eQTL) detection. *Biometrics* **63**, 1108–16.
GIANNONE, D., LENZA, M. & PRIMICERI, G. (2017). Economic predictions with big data: The illusion of sparsity. CEPR Discussion Paper No. DP12256. Available at SSRN: https://ssrn.com/abstract=3031893.
GILAD, Y., RIFKIN, S. A. & PRITCHARD, J. K. (2008). Revealing the architecture of gene regulation: The promise of eQTL studies. *Trends Genet.* **24**, 408–15.
HALLIN, M. & LIŠKA, R. (2007). Determining the number of factors in the general dynamic factor model. *J. Am. Statist. Assoc.* **102**, 603–17.
KENDZIORSKI, C., CHEN, M., YUAN, M., LAN, H. & ATTIE, A. (2006). Statistical methods for expression quantitative trait loci (eQTL) mapping. *Biometrics* **62**, 19–27.
KUNDAJE, A., MEULEMAN, W., ERNST, J., BILENKY, M., YEN, A., KHERADPOUR, P., ZHANG, Z., HERAVI-MOUSSAVI, A., LIU, Y., AMIN, V. et al. (2015). Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–30.
LAM, C. & YAO, Q. (2012). Factor modeling for high-dimensional time series: Inference for the number of factors. *Ann. Statist.* **40**, 694–726.
LEE, M. N., YE, C., VILLANI, A. C., RAJ, T., LI, W., EISENHAURE, T. M., IMBOYWA, S. H., CHIPENDO, P. I., RAN F. A., SLOWIKOWSKI. K. et al (2014). Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science*. **343**, 1246980.
LI, K. C. (1991). Sliced inverse regression for dimension reduction (with Discussion). *J. Am. Statist. Assoc.* **86**, 316–42.
LONSDALE, J., THOMAS, J., SALVATORE, M., PHILLIPS, R., LO, E., SHAD, S., HASZ, R., WALTERS, G., GARCIA, F., YOUNG, N. et al. (2013). The genotype-tissue expression (GTEx) project. *Nature Genet.* **45**, 580–5.
MA, Y. & ZHANG, X. (2015). A validated information criterion to determine the structural dimension in dimension reduction models. *Biometrika* **102**, 409–20.
MA, Y. & ZHU, L. (2012). A semiparametric approach to dimension reduction. *J. Am. Statist. Assoc.* **107**, 168–79.
MA, Y. & ZHU, L. (2013a). Efficient estimation in sufficient dimension reduction. *Ann. Statist.* **100**, 371–83.
MA, Y. & ZHU, L. (2013b). A review on dimension reduction. *Int. Statist. Rev.* **81**, 134–50.
NICA, A. C. & DERMITZAKIS, E. T. (2013). Expression quantitative trait loci: Present and future. *Phil. Trans. R. Soc. Lond.* B **368**, 20120362.
SCHENA, M., SHALON, D., DAVIS, R. W. & BROWN, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–70.
STOCK, J. H. & WATSON, M. W. (2002). Forecasting using principal components from a large number of predictors. *J. Am. Statist. Assoc.* **97**, 1167–79.
TAN, K. M., WANG, Z., ZHANG, T., LIU, H. & COOK, R. D. (2018). A convex formulation for high-dimensional sparse sliced inverse regression. *arXiv:* 1809.06024.
THE 1000 GENOMES PROJECT CONSORTIUM (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65.
THE ENCODE PROJECT CONSORTIUM (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74.
VISSCHER, P. M., BROWN, M. A., MCCARTHY, M. I. & YANG, J. (2012). Five years of GWAS discovery. *Am. J. Hum. Genet.* **90**, 7–24.
WANG, Z., GERSTEIN, M. & SNYDER, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nature Rev. Genet.* **10**, 57–63.
WHITE, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–25.
ZHU, L., LI, L., LI, R. & ZHU, L.-X. (2011). Model-free feature screening for ultrahigh dimensional data. *J. Am. Statist. Assoc.* **106**, 1464–75.
ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Am. Statist. Assoc.* **101**, 1418–29.

[*Received on* 28 *February* 2018. *Editorial decision on* 11 *October* 2018]