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

Fei Jiang, Yanyuan Ma and Ying Wei *

SUMMARY

Rapid improvement in technology has made it relatively cheap to collect genetic data. However statistical analysis is still much cheaper. Thus, secondary analysis of single-nucleotide polymorphism (SNP) data, re-analyzing 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 to the existing estimator that relies on additional assumptions on the latent factors. We quantify the asymptotic performance of the parameter estimators and perform inference by assessing the estimation variability and constructing confidence intervals. The new results enable us to identify statistically significant SNPs concerning gene-SNPs relation in lung tissues for the first time from the Genotype-Tissue Expression data.

Keywords: Dimension reduction, high dimension, factor model, nonparametrics, semiparametrics.

1. INTRODUCTION

Gene expression quantitative trait loci (eQTLs) are the genetic variants that may explain the variations of gene expressions. 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), and it is also the most prominent way to discover the gene regulation network (Gilad et al., 2008). Many genomic findings rely on eQTLs for a meaningful interpretation. For example, numerous Genome-Wide Association Studies have been performed in recent years to identify genetic variants that are associated with complex diseases (Lee et al., 2014; Visscher et al., 2012). Such efforts have resulted in more than 2000 disease associated variants being identified. However, most of them are non-coding, and hence their links to underlying diseases are likely to be through regulating gene expression. Consequently, understanding how these genetic variants associated with gene expressions is essential for interpreting these disease associate variants.

Figure 1 illustrates the typical data structure of an eQTL study. The gray box indicates the location of a target gene on the genome, whose gene 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 stand for the locations of a set of pre-identified candidate single-nucleotide polymorphisms (SNPs) within a local region centered around the gene (the blue vertical dotted lines). More candidate SNPs will be included as the width of the search window expands. Typical choices of the length of windows include 20kb, 100kb, and 1Mb. The goal of eQTL studies is to identify which candi-

* Fei Jiang is assistant professor, Department of Statistics, University of Hongkong, Hongkong. feijiang@hku.hk. Yanyuan Ma is professor, Department of Statistics, Penn State University, University Park, PA, 16802, yzym63@psu.edu. Ma’s work is supported by NSF grant DMS-1608540. Ying Wei is associate professor, Department of Biostatistics, Columbia University, New York, NY 10032. Wei’s work is supported by NIH grant R01 HG008980-01. Fei’s work is supported by RGC grant ECS 27304117.
date SNPs are significantly associated with the target gene expressions. To further illustrate
the eQTL analysis, we retrieve a subset of data from the Genotype-Tissue Expression pilot
data, which is one of major international projects in eQTL discovery (Lonsdale et al., 2013).
The data is available upon request in the database of Genotypes and Phenotypes, the NIH
sponsored data achieve for genetic studies, at https://www.ncbi.nlm.nih.gov/projects/gap/cgi-
bin/study.cgi?study_id=phs000424.v5.p1. The data set includes 278 lung tissues collected from
278 subjects. For each tissue, a scientist measured its gene expression levels across the entire
gene using RNA-Seq technology, and also genotyped the tissue sample across the entire
genome. We randomly select a gene ENSG00000225880.4, and denote its expression level $Y_i$
for the $i$-th tissue, $i = 1, \ldots, 278$. Using the window length of 20kb, we identified 117 candidate
SNPs, and use $X_i = (X_{i,1}, X_{i,2}, \ldots, X_{i,p})^T$ to denote the genotypes of these SNPs in the $i$-th
tissue. The goal of the analysis is to identify the genetic variations across individuals that could
explain the different gene-expression levels in the lung tissues.

Due to the high cost in measuring gene expression levels, most eQTL data have limited sam-
ple sizes. Restrained by the limited sample sizes, most eQTL analyses are conducted separately
for each gene-SNP pair (Ardlie et al., 2015). Typically, researchers study the mean gene ex-
pression level given a single SNP each time, while ignoring other covariates. Previous studies
using this approach have identified the eQTLs in lymphocytes, adrenal gland, thyroid, artery
and skin tissues for the gene ENSG00000225880.4, see the reported single tissue eQTLs at
http://gtexportal.org/. When we applied this kind of modeling and estimation approach to the
lung tissue data in Section 5, we identified 76 significant SNPs after the 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 fix this deficiency, recent ef-
forts for eQTL analysis focus on combined analysis of all the genetic variations (see Kendziorski
et al. 2006, 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 the candidate SNPs.
However, it is challenging to estimate a full model as $p$ could be comparative to or even much
greater than the sample size $n$.

The current practice in the high dimension setting relies on putting penalties on the parameters
and assume sparsity of the data. Following this general approach, we implemented several penal-
ization methods. Again, none of them provided satisfactory results. For example, we performed
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 large as 0.8 (see Figure 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
leads us to consider an alternative to the penalization method.

The alternative that we propose is to extend the 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 modeling, 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 in
Li (1991). A very comprehensive introduction to 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 moderate number of covariates, despite the claim of high dimension.
The phenomenon is partly caused by 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 works dealing with very high dimensional covariates in this framework is also because a sparsity and penalisation approach is almost unavoidable (Tan et al., 2018) if the number of covariates is indeed 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 is already made. This brings a somewhat awkward situation because one of the original goals of sufficient dimension reduction is to provide an alternative approach to handling large number of covariates, one that avoids the popular sparsity assumption and penalization technique.

This awkward situation is successfully avoided by a creative modeling approach proposed in Fan et al. (2017) recently, which jointly uses the factor model and sufficient dimension reduction technique to analyze the high dimensional data in a time series context under the linearity condition. We adapt the idea of combining Factor Analysis and Dimension Reduction Estimation in performing secondary analysis of SNPs-gene expression data, while we further relax some of their assumptions 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 the model is uniquely defined, where 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 with the same condition, the model is asymptotically identifiable when $p, n \to \infty$.

In addition, we establish the asymptotic normality of our estimator from combining the factor analysis and the sufficient dimension reduction models, and derive the asymptotic variance of our estimators, 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 doubly robust 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 $i$-th subject, $i = 1, \ldots, n$, and $X_i$ is a $p$-dimensional vector of covariates, which include his or her $p_1$ SNP values within a local region around the target gene, and $p_2$ controlling covariates. Let $p = p_1 + p_2$. We assume that the observations $(X_i, Y_i)$ are independent identically distributed, and 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 $X_i = (X_{i1}, \ldots, X_{ip})^T$, $X_{il} = b_l^T f_i + u_{il}, 1 \leq l \leq p, 1 \leq i \leq n,$ (1) where $b_l$ is a $q$-dimensional vector. The relation in (1) can be written in the matrix form $X_i = B f_i + u_i,$ (2) where $f_i = \left( f_{i1}, \ldots, f_{iq} \right)^T$ is a $q$-dimensional vector of factors, $B = \left( b_1, \ldots, b_p \right)^T$ is a $p \times q$ deterministic matrix, $u_i = \left( u_{i1}, \ldots, u_{ip} \right)^T$ and $u = \left( u_1, \ldots, u_n \right)$. We require $q < p$ and require
$u_i$ to be independent of $f_i$ and $E(u_i) = 0$. Let $F = (f_1, \ldots, f_n)^T$ be an $n \times q$ matrix, and $X = (X_1, \ldots, X_n)$ be an $p \times n$ matrix. We further consider a sufficient dimension reduction model of the factors $f_1, \ldots, f_n$.

\[ Y_i = \psi(\beta^T f_i, \epsilon_i), \tag{3} \]

where $\psi$ is an unknown function, $\epsilon_i$ is a random variable independent of $\beta^T f_i$ and $u_i$, $\beta$ is a $q \times d$ dimensional parameter vector with $d < q$.

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

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 $\psi$ 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, they developed a link-free sufficient forecasting method based on the sliced inverse regression to further reduce the dimension from $q$ to $d$ and to deliver additional predictive power.

The drawback of the sliced inverse regression is that it requires the linearity condition on the covariates $f_i$, i.e., it requires that $E(f_i \mid \beta^T f_i) = \beta(\beta^T \beta)^{-1} \beta^T f_i$ 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 $f_i$ themselves cannot be observed directly, it is desirable to allow the maximum flexibility and avoid any structural assumptions on the distribution of the $f_i$’s. Thus, to relax the linearity condition on the latent variables, we adopt the semiparametric method 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 equivalently written into 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, while one can perform a centering step to the other multiplier to achieve zero mean as well if the linearity condition is violated, hence retain the consistency of the estimating equation. In addition, replacing the linear form of $E(f \mid \beta^T f)$, which is assumed by the linearity condition, by 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 class of sufficient dimension reduction estimates under 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, we further derive the specific forms of the asymptotic variances in addition to the results on convergence order, which are not achieved in Fan et al. (2017). This is crucial because the calculation of p-values and the identification of statistically significant covariates, that is the identification of eQTLs, relies on such properties.
3. Estimation

3.1. Estimation Algorithm

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

- Following Fan et al. (2017), we solve the following constrained least square problem:
  
  \[
  (\hat{B}, \hat{F}) = \arg\min_{B,F} \|X - BF^T\|_F^2
  \]
  subject to
  
  \[
  n^{-1}F^T F = I_q, B^T B \text{ is diagonal},
  \]
  to obtain the estimators \( \hat{F} = (\hat{f}_1, \ldots, \hat{f}_n)^T \) for \( F \), and \( \hat{B} = (\hat{b}_1, \ldots, \hat{b}_p)^T \) for \( B \), where \( \hat{f}_i \) and \( \hat{b}_i \) are the estimators for \( f_i \) and \( b_i \) respectively. This is a classical principal components problem. The estimated factor matrix \( \hat{F} \) is \( n^{1/2} \) times eigenvectors corresponding to the \( q \) largest eigenvalues of \( n \times n \) matrix \( X^TX \), and \( \hat{B} = n^{-1}X\hat{F} \) are the corresponding factor loading.

- Treating \( \hat{f}_i \)'s as the covariates, following Ma & Zhu (2012), we then solve
  
  \[
  n^{-1} \sum_{i=1}^{n} [g(Y_i, \beta^T \hat{f}_i) - \hat{E}\{g(Y_i, \beta^T \hat{f}_i)|\beta^T \hat{f}_i\}][\eta(\hat{f}_i) - \hat{E}\{\eta(\hat{f}_i)|\beta^T \hat{f}_i\}] = 0
  \]  
  (4)

  for \( \beta \). The resulting \( \hat{\beta} \) is the estimator for \( \beta_0 \). Here, \( g \) and \( \eta \) are user chosen smooth functions,

  \[
  \hat{E}\{g(Y_i, \beta^T \hat{f}_i)|\beta^T \hat{f}_i\} = \frac{\sum_{j=1}^{n} K_h(\beta^T \hat{f}_j - \beta^T \hat{f}_i)g(Y_j, \beta^T \hat{f}_j)}{\sum_{j=1}^{n} K_h(\beta^T \hat{f}_j - \beta^T \hat{f}_i)},
  \]

  \[
  \hat{E}\{\eta(\hat{f}_i)|\beta^T \hat{f}_i\} = \frac{\sum_{j=1}^{n} K_h(\beta^T \hat{f}_j - \beta^T \hat{f}_i)\eta(\hat{f}_j)}{\sum_{j=1}^{n} K_h(\beta^T \hat{f}_j - \beta^T \hat{f}_i)},
  \]

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

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

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

In the second step of the above estimation procedure, users have the freedom to choose the functions $g$ and $\eta$. The general requirement is that $g$ is a function of $Y$ and $\beta^T f$ only, $\eta$ is a function of $f$ only, and they are sufficiently rich so 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 $(Y, \beta^T f)$ and those of $\eta$ to be monomials of $f$, i.e., $g = \{Y, f^T \beta, Y^2, Y(f^T \beta), (f^T \beta) \otimes (f^T \beta), \ldots, Y^k, Y^{k-1}(f^T \beta), \ldots, (f^T \beta) \otimes \cdots \otimes (f^T \beta)\}^T$, and $\eta = \{f^T, (f \otimes f)^T, \ldots, (f \otimes \cdots \otimes f)^T\}$. 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 it down to exactly $(q - d)d$ through generalized method of moment. Different choices of $g$ and $\eta$ will affect the estimation variability of $\beta$, while the consistency of the $\beta$ estimation is retained regardless of the choices of $g, \eta$. The optimal choice is $g = \partial \log f_{Y|\beta^T f}(y, \beta^T f)/\partial (\beta^T f)$ and $\eta$ is a $p - d$ dimensional subvector of $f$ (Ma & Zhu, 2013a), where the efficient estimator of $\beta$ will be obtained. The price associated with the optimal choice is the need to estimate the conditional density function $f_{Y|\beta^T f}(y, \beta^T f)$ and its derivative with respect to $\beta^T f$, hence other non-optimal choice is also used in practice. The bandwidth $h$ does not play a critical role, and can be chosen as any values that satisfy Condition B4 in Section 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 determinations of the dimension of the latent factor $q$ and the structure dimension $d$. In terms of the choice of $d$, Ma & Zhang (2015) proposed a validated information criterion which select $d$ consistently through minimizing a validation of the goodness-of-fit.

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

4. Main results

4.1. Uniqueness of model in the ultra-high dimension setting

Linear factor model is a mature research topic. The computational identifiability of the 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 identifiability property is established, estimation becomes 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 $W^TW$. Let $\|W\|_\infty$ be the sup-norm of the matrix, i.e., the maximum of the absolute row sums. Finally let $\|W\|_F$ be the Frobenius norm. For the identifiability of $B, F$ when $n, p \to \infty$, we require the following conditions.

**Regularity conditions for the identifiability of $B, F$.**

(A1) There exists a constant $M$, not depending on $p$ and $n$, such that $E(\|f_i\|_2^2) \leq M$. In addition, $E(f_i) = 0, \text{cov}(f_i) = I_q$, where $I_q$ is a $q \times q$ identity matrix.
(A2) Let \( b_l \)'s be deterministic and \( \|b_l\| \leq M \), where \( M \) is a constant independent of \( n \) and \( p \). The matrix \( p^{-1}B^TB \) is diagonal, with distinct positive entries arranged in decreasing order. When \( p \to \infty \), \( p^{-1}B^TB \to \Sigma_A \), where \( \Sigma_A \) is \( q \times q \) diagonal non-random matrix with positive distinct diagonal elements. In addition, the first nonzero element in each column of \( B \) is positive.

(A3) For some \( c > 0 \), the loading \( b_l \) satisfies that \( \|b_l\|_2 \leq c \) for \( l = 1, \ldots, p \). As \( p \to \infty \), there are two positive constant \( c_1, c_2 \) such that

\[
\frac{c_1}{\lambda_{\min}(B^TB/p)} < \frac{1}{\lambda_{\max}(B^TB/p)} < \frac{c_2}{\lambda_{\max}(B^TB/p)}.
\]

Here and throughout the text, \( \lambda_{\min}(M) \) and \( \lambda_{\max}(M) \) are the minimum and maximum eigenvalues of a symmetric matrix \( M \), respectively.

(A4) The random variables \( u_i \)'s are independent of each other and \( u_i \) is independent of \( b_l, f_i \), and \( E(u_{il}) = 0, nE(u_{il}^2) \leq M \) for all \( l = 1, \ldots, p \). Further, for all \( i = 1, \ldots, n \), \( \sum_{l=1}^p E(u_{il}^2) \leq M \) and \( p^{-1/2} \sum_{l=1}^p |u_{il}^2 - E(u_{il}^2)|^4 \leq M \).

(A5) When \( p \to \infty \), \( p^{-1/2} \sum_{l=1}^p b_l u_{il} \xrightarrow{d} N(0, \Gamma) \), where

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

is a bounded variance matrix.

(A6) The random variables \( f_i \) and \( u_i \) are mutually independent conditional on \( B \). In addition,

\[
E(f_i f_i^T u_{il}^2) = \Phi_i.
\]

Condition (A4) implies \( \|\Sigma_U\|_1 \leq M \), where \( \Sigma_U = E(u_i u_i^T) \). Conditions (A1)–(A4) are needed for the identifiability and consistency of estimation, while Conditions (A5) and (A6) are needed for the asymptotic distribution of the estimators. We first establish the identifiability result in Proposition 1, with its proof given in Section A.1 of the supplement.

**Proposition 1.** Under Conditions (A1)–(A4), \( B, \Phi \) are unique as \( p \to \infty \).

### 4.2. Theoretical properties

Fan et al. (2017) established the consistency of \( \hat{\beta} \) when using the sliced inverse regression in the dimension reduction step. Their result requires the linearity condition to hold. We adopt the semiparametric method and show the consistency of the resulting estimator without imposing such linearity condition. This is an important step forward because a key feature of factor model is not to impose assumptions, including the linearity condition, on the latent factor. In addition, we also show the asymptotic normality and derive the asymptotic variance of the estimators. These results are crucial in genetic studies because inference, \( p \)-value calculation and finally the selection of the significant SNPs all require these results. It is worthwhile to mention that our results are established in a very general context. They can be readily applied regardless the semiparametric sliced inverse regression, semiparametric sliced average variance estimation, semiparametric directional regression, semiparametric principal Hessian directions or any other choices of the \( q, \eta \) are used to conduct the second step of the Factor Analysis and Dimension Reduction Estimation.

**Regularity conditions for asymptotic properties**

(B1) The univariate kernel function \( K(\cdot) \) is Lipschitz, has compact support. It satisfies

\[
\int K(v)dv = 1, \quad \int v^t K(v)dv = 0, 1 \leq t \leq m - 1, \quad 0 \neq \int v^m K(v)dv < \infty.
\]
The $d$–dimensional kernel function is a product of $d$ univariate kernel function, that is $K_h(v) = K(v/h)/h^d = \prod_{i=1}^d K_h(v_i) = \prod_{i=1}^d (v_i/h)/h^d$ for $v = (v_1, \ldots, v_d)^T$.

(B2) The density functions of $f_i$ and $\beta_i f_i$ denoted by $\pi_f(f_i)$ and $\pi(\beta_i f_i)$, are bounded away from zero and infinity. Each entry in the matrices $E\{g(Y_i, \beta_i f_i) g(Y_i, \beta_i f_i)^T | \beta_i f_i\}$ and $E\{\eta(f_i) \eta(f_i)^T | \beta_i f_i\}$ is locally Lipschitz continuous and bounded from above as a function of $\beta_i f_i$.

(B3) Let $r_1(\beta_i f_i) = E\{\eta(f_i) | \beta_i f_i\} \pi(\beta_i f_i)$ and $r_2(\beta_i f_i) = E\{g(Y_i, \beta_i f_i) | \beta_i f_i\} \pi(\beta_i f_i)$. The $m$th derivatives of $r_1(\beta_i f_i), r_2(\beta_i f_i)$ and $\pi(\beta_i f_i)$ are locally Lipschitz continuous.

(B4) The bandwidth $h = O(n^{-\kappa})$ for $1/(4m) < \kappa < 1/(2d)$.

(B5) For the identification of $\beta$, we further assume that the upper $d \times d$ matrix of $\beta$ is an identity matrix and the lower $(p - d) \times d$ matrix of $\beta$ is arbitrary.

(B6) $E[g(Y_i, \beta_i f_i) - E\{g(Y_i, \beta_i f_i) | \beta_i f_i\}] \pi(\beta_i f_i)$ is a smooth function of $\beta$ and has unique root for $\beta$.

(B7) The random vectors $f_i$, $u_i$, and $\epsilon_i$ are mutually independent.

Condition (B1) is a usual requirement on the kernel function, where $m$ is usually referred to as the order of the kernel. Conditions (B2) and (B3) set sufficient smoothness requirement on several functions. Condition (B4) adds requirement on the bandwidth related to the kernel order and the dimension $d$. It is seen that as long as $d \leq 3$, the common second order kernel function is sufficient. Condition (B5) guarantees the identifiability of $\beta$ (Ma & Zhu, 2013a). Condition (B6) ensures the global consistency of $\hat{\beta}$ (White, 1982). Condition (B7) contains standard independence assumptions from factor model and dimension reduction model formulation. These conditions are all moderate and are commonly required.

Under the true model, we have

$$ E[g(Y_i, \beta_0 f_i) - E\{g(Y_i, \beta_0 f_i) | \beta_0 f_i\}] \eta(f_i) - E\{\eta(f_i) | \beta_0 f_i\} = 0. $$

(5)

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

We are now ready to establish the main theorems of this article. Theorem 1 derives the consistency property of the sufficient reduction directions, and Theorem 2 further establishes the asymptotic properties of these directions. Specifically, the asymptotic normality is proven, and the asymptotic variance is derived. These results are established under the situation that both the dimension of covariates and the number of observations are growing, and 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 setting. The proofs are given in Sections A.6 and A.7 respectively.

**Theorem 1.** Assume (A1)–(A4) and (B1)–(B6), let $\hat{\beta}$ satisfy

$$ n^{-1} \sum_{i=1}^n \{g(Y_i, \beta_0 f_i) - \hat{E}\{g(Y_i, \beta_0 f_i) | \beta_0 f_i\}\} \eta(f_i) - \hat{E}\{\eta(f_i) | \beta_0 f_i\} = 0. $$

Then $\hat{\beta} - \beta_0 = o_p(1)$.

**Theorem 2.** Assume (A1)–(A6) and (B1)–(B6), and let $\hat{\beta}$ solve

$$ n^{-1} \sum_{i=1}^n \{g(Y_i, \beta_0 f_i) - \hat{E}\{g(Y_i, \beta_0 f_i) | \beta_0 f_i\}\} \eta(f_i) - \hat{E}\{\eta(f_i) | \beta_0 f_i\} = 0. $$
ric sliced inverse regression and the semiparametric principal Hessian directions under model

where $\epsilon Y^I$.

We evaluated the performance of the methods on the models

Remark 1. When $d = 1$ and $\psi(\cdot)$ is linear, Bai (2003) empirically argued that the regression estimators converge to the true values in a root $n$ rate when $n^{1/2}p^{-1} \to 0$. Here we establish the result rigorously and extend it to the cases when $d > 1$ 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 used sample size $n = 300$ and repeated our simulation 1000 times. To generate $X_i$ from model (1), we considered two cases. Case I: we simulated $f_i$ from multivariate norm distribution with mean zero and covariance matrix $(\sigma_{ij})_{q \times q}$ with $\sigma_{ij} = 0.5^{i+j}$. Case II: we simulated $f_{i1}, f_{i2}$ from multivariate normal distribution with mean zero and covariance matrix $(\sigma_{ij})_{2 \times 2}$ with $\sigma_{ij} = 0.5^{i+j}$. We generate $f_{i3} = |f_{i1} + f_{i2}| + f_{i1}x_{i1}$, $f_{i4} = |f_{i1} + f_{i2}| + |f_{i2}x_{i2}|$, where $x_{i1}, x_{i2}$ are independently generated from standard normal distribution, $f_{i5}$ 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 $\Phi$ is the standard normal distribution function. We center and normalize $F$ by its mean and covariance so that $F$ indeed satisfies Condition (A1). To construct the matrix $B$, we first generate $n$ samples of $p$ dimensional random vector $Z_i$ from a normal distribution with mean 0 and covariance matrix $\Sigma_\varepsilon$, where $\Sigma_{\varepsilon ij} = 0.5^{i+j}$ for $1 \leq i, j \leq p$. Let $Z = (Z_1, \ldots, Z_n)^T$.

We perform eigen decomposition 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 $B = 1/6^{1/2}Z^TE$. This construction yields the eigenvalues of $B^TB/p$ in the range of $(2, 3)$, which ensures that $B^TB = O_p(p)$, as required in Condition (A3). To ensure Condition (A4) and (A5), we simulated $u_{id}$ from a normal distribution with mean 0 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 models

(I): $Y_i = (f_i^T\beta_1)/\{0.5 + (f_i^T\beta_2 + 1.5)^2\} + 0.5\epsilon_i$,

(II): $Y_i = \exp(f_i^T\beta_1) + 2|f_i^T\beta_2 + 1| + 0.1|f_i^T\beta_1|\epsilon_i$,

(III): $Y_i = (f_i^T\beta_1)^2 + 2|f_i^T\beta_2 + 1| + 0.1(f_i^T\beta_1)^2\epsilon_i$,

where $\epsilon_i$ follows the standard normal distribution. For Case I, we evaluated the semiparametric sliced inverse regression and the semiparametric principal Hessian directions under model
(I), while evaluated the semiparametric sliced average variance estimation and semiparametric directional regression under model (II). For Case II, we evaluated the semiparametric sliced inverse regression and semiparametric principal Hessian directions under model (I), while evaluated semiparametric sliced average variance estimation and semiparametric directional regression under model (III). We designed these simulations to limit the resulting $Y_i$ values 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 as in Table A.1 in the supplementary material.

We compare the proposed semiparametric method and the original dimensional reduction techniques through the Euclidean distances between the resulting estimators and the true values in Figures A.1 and A.2 in the supplementary material. We also evaluated the asymptotic performances of the estimators and the results are shown in Tables A.3 and A.4 in the supplementary material. Further we compare 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 the confidence intervals have coverage probabilities close to the nominal level. In addition, most of the estimators achieve the asymptotic normality, with p-values of the Kolmogorov-Smirnov normal test less than 0.00625=0.05/8, the bound adjusted for multiple testing.

5.2. Analysis on eQTL discovery

In this section, we illustrate the application of the proposed semiparametric method for eQTL discoveries. Recall that the illustrative Genotype-Tissue Expression data contains $n=278$ subjects. Their expression levels on the gene ENSG00000225880.4 in their lung tissues were measured by RNA-seq technique. They were also genotyped on 117 SNPs within 20kb from the target gene. In addition, we also have 40 controlling covariates include gender, platform, three principal components of expressions of genome-wide gene expressions, and 35 principal components of genome-wide SNPs. Those 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 into one model, which could take the inter-SNP correlations into account. It consequently enhances the power in identifying eQTLs and may provide new insights on the SNP functionals. To apply the proposed methods to the Genotype-Tissue Expression gene expression data, we first perform a principal component analysis on the 157 covariates. Following the eigen-value ratio method discussed in Ahn & Horenstein (2013), we compute

$$\tilde{q} = \arg\max_{1 \leq q \leq 157} \log \left( \frac{V_{q-1}/V_q}{V_q/V_{q+1}} \right)$$

and obtain $\tilde{q} = 5$, where $V_q$ is the average of the first $q$ eigenvalues of the matrix $B^T B$. This approach suggests to pick the first five factors for the second stage analysis. To be more conservative, we plot the percentage of variance explained by each principal component of the first fifteen principal components in Figure 3.

We can see that the results for the 5th and the 6th components are very similar. Taking into account these two aspects, $q = 6$ is a reasonable choice in our analysis. Further, to avoid carrying out futile analysis, we perform a test of the null hypothesis that none of the 6 factors are related to the response by using the method in the spirit of Zhu et al. (2011). To this end, we first calculate $\hat{R}_k = n^{-1} \sum_{j=1}^{n} \left( n^{-1} \sum_{i=1}^{n} \tilde{f}_{ik} I(Y_i < Y_j) \right)^2$ for each estimated factor $\tilde{f}_{ik}, k = 1, \ldots, 6$. The resulting $(\hat{R}_k, k = 1, \ldots, 6) = (0.0061, 0.00026, 0.0037, 0.0022, 0.0009, 0.0002)$. Further,
we select the threshold through permuting the rows of \( \hat{F} \) 100 times. In the \( l \)th permutation, let \( \tilde{F}_l \) be the permuted \( \hat{F} \), and \( \tilde{f}_{ikl} \) be its \((i,k)\) element, we 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, \( \tilde{R}_1 > 0.0039 \). 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 directional regression, 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 the 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 provided the corresponding estimates 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 a two-fold cross validation. Specifically, we randomly split the data into two equal parts as the training and testing data sets and computed the mean predictive errors for each method. The averages of the mean predictive errors over 100 random splits are 1.0715, 1.0755, 1.0378, 1.0294 for the semiparametric sliced inverse regression, semiparametric sliced average variance estimation, semiparametric directional regression and semiparametric principal Hessian directions methods while those are 1.0985, 1.1170, 1.1043, and 1.1138 for the original sliced inverse regression, sliced average variance estimation, semiparametric directional regression, semiparametric principal Hessian directions methods, respectively. It is clear that the semiparametric methods outperform the classical dimension reduction methods in terms of prediction in this data set. Considering that the results from semiparametric principal Hessian directions 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 the semiparametric principal Hessian directions estimator.

To assess the effect of individual SNPs on the gene expression, we estimate the \( \alpha \) coefficients \( \hat{\alpha} \equiv \hat{B}(\hat{B}^T \hat{B})^{-1} \hat{\beta} \), where \( \hat{B} \) is the factor loading obtained from the first step principal component analysis. The first \( p_1 \) components of vector \( \alpha \) correspond to the effects of SNPs on the gene expression level in the sufficient direction. To test the null hypothesis \( \alpha_j = 0 \), we calculate the p-values via \( p_j \equiv 2[1 - \Phi\{ |\hat{\alpha}_j|/\hat{\sigma}(\hat{\alpha}_j) \}] \), where \( \hat{\alpha}_j \) is the \( j \)th components in \( \hat{\alpha} \), \( \Phi \) is the standard normal distribution function. The \( -\log_{10} \) of the resulting p-values were plotted in Figure 4, and were compared with \( -\log_{10}(0.05/157) \) to adjust for multiple comparisons. As shown in the left plot of Figure 4, we identified 27 variants at loci in Table A.2 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 FDR (Benjamini & Hochberg, 1995) within 0.05 by treating the p-values as independent. We present the results of the false discovery rate based analysis in the middle plot of Figure 4. Compared to the traditional pair-wise 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, lung, gastric, heart, ovary, thymus and spleen. Functional annotation scores were recently developed in Backenroth et al. (2017) to predict the functional effect of noncoding genetic variants in different cell and tissue types. They are estimated from independent Roadmap datasets (Consortium et al., 2012b; Kundaje et al., 2015), and measure the probability of a SNP to regulate gene expressions in certain cell and tissue type. On average, about 5% of SNPs have functional scores > 0.01 in lung (estimated from the 1000 Genomes Project (Consortium et al., 2012a)).

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

Remark 2. In our analysis, we used FDR 0.05 level to build the selection of eQTLs. In practice, FDR 0.05 level may not be suitable to screen in as many true signals as possible for follow up studies (Craiu & Sun, 2008). We examine the relationship between number of identified SNPs and the FDR level in the right plot of Figure 4. Researchers may determine the proper FDR level based on their purposes for the follow up studies.

ACKNOWLEDGMENT

This work is partially supported by grants from National Sciences Foundation, National Institute of Neurologic Disorder and Stroke and National Human Genome Research Institute, Research Grants Council of Hong Kong.

REFERENCES

Bai, J. & Ng, S. (2002). Determining the number of factors in approximate factor models. Econometrica 70, 191–221.


Fig. 1. Data structure of eQTL study

Fig. 2. The correlation between the unselected SNPs and a selected SNP rs67081753

Fig. 3. The barplot of principal components.
Table 1. Validated information criterion values at $d = 1, \ldots, 4$ and estimates and standard errors under $d = 1$ for the semiparametric sliced inverse regression, semiparametric sliced average variance estimation, semiparametric principal Hessian directions, and semiparametric directional regression in Gene-SNP association analysis. S-SIR, S-SAVE, S-PHD, S-DR are short names for semiparametric sliced inverse regression, semiparametric sliced average variance estimation, semiparametric principal Hessian directions, and semiparametric directional regression, respectively.

<table>
<thead>
<tr>
<th>$d$</th>
<th>S-SIR</th>
<th>S-SAVE</th>
<th>S-DR</th>
<th>S-PHD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>103.2497</td>
<td>200.1980</td>
<td>118.9829</td>
<td>78.1923</td>
</tr>
<tr>
<td>2</td>
<td>160.3691</td>
<td>133.8987</td>
<td>144.3405</td>
<td>122.2833</td>
</tr>
<tr>
<td>3</td>
<td>186.1471</td>
<td>152.0223</td>
<td>150.8336</td>
<td>142.5507</td>
</tr>
<tr>
<td>4</td>
<td>194.0230</td>
<td>220.2505</td>
<td>186.5501</td>
<td>183.7236</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\beta$</th>
<th>Estimate (Standard Error)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_{12}$</td>
<td>-0.124 (0.151)</td>
</tr>
<tr>
<td>$\beta_{13}$</td>
<td>-0.539 (0.178)</td>
</tr>
<tr>
<td>$\beta_{14}$</td>
<td>-0.534 (0.145)</td>
</tr>
<tr>
<td>$\beta_{15}$</td>
<td>-0.134 (0.144)</td>
</tr>
<tr>
<td>$\beta_{16}$</td>
<td>-0.055 (0.065)</td>
</tr>
</tbody>
</table>

Fig. 4. Base 10 log-transformed p-values ($-\log_{10} p_j$) for each estimated covariate effect (left). Sorted base 10 log-transformed p-values ($-\log_{10} p_j$) versus 0.05j/157 (middle). FDR level versus the number of identified SNPs (right).

Fig. 5. Scatter plot of the first and second factor loadings of the 27 identified SNPs.
Fig. 6. Functional annotation scores of the 27 SNPs by cluster