Supervised structural learning of semiparametric regression on high-dimensional correlated covariates with applications to eQTL studies

Wei Liu[1], Huazhen Lin[1,2*], Li Liu[3], Yanyuan Ma[4], Ying Wei[5] and Yi Li[6]
[1]Center of Statistical Research and School of Statistics, Southwestern University of Finance and Economics, Chengdu, China
[2]New Cornerstone Science Laboratory, Shenzhen 518054, China
[3]School of Mathematics and Statistics, Wuhan University, Wuhan, China
[4]Department of Statistics, Penn State University, University Park, PA 16802, USA
[5]Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY 10000, USA
[6]Department of Biostatistics, University of Michigan, Ann Arbor, USA

Abstract

Expression quantitative trait loci (eQTL) studies utilize regression models to explain the variance of gene expressions with genetic loci or single nucleotide polymorphisms (SNPs). However, regression models for eQTL are challenged by the presence of high dimensional non-sparse and correlated SNPs with small effects, and nonlinear relationships between responses and SNPs. Principal component analyses are commonly conducted for dimension reduction without considering responses. Because of that, this non-supervised learning method often does not work well when the focus is on discovery of the response-covariate relationship. We propose a new supervised structural dimensional reduction method for semiparametric regression models with high dimensional and correlated covariates; we extract low-dimensional latent features from a vast number of correlated SNPs while accounting for their relationships, possibly nonlinear, with gene expressions. Our model identifies important SNPs associated with gene expressions and estimates the association parameters via a likelihood-based algorithm. A GTEx data application on a cancer related gene is presented with 18 novel eQTLs detected by our method. In addition, extensive simulations show that our method outperforms the other competing methods in bias, efficiency and computational cost.

Key words and phrases: Latent variables; Bias and Efficiency; Joint Models; SNP and Gene Expression; GTEx.

*Corresponding author: Center of Statistical Research, Southwestern University of Finance and Economics, 555, Liutai Avenue, Wenjiang District, Chengdu, Sichuan, China, 611130. Email: linhz@swufe.edu.cn.

This research is partially supported by National Key R&D Program of China (2022YFA1003702), National Natural Science Foundation of China (11931014 and 11829101), New Cornerstone Science Foundation, NSF grant (1608540), NINDS grant (010847) and NIH/NHGRI R01award (HG008980)

1 Introduction

An expression quantitative trait locus (eQTL) is a locus that explains variation in a gene expression phenotype[1]. Identification of eQTLs helps characterize functional sequence variation, understand basic processes underpinning gene regulation, facilitate interpretation of genome-wide association studies and decipher biology of complex diseases[1]. Genotype-Tissue Expression (GTEx) Program, as an NIH initiative, established a data resource and tissue depot for eQTL studies on multiple human tissues across individuals[2]. Of our particular interest is to investigate which loci are related to the expression of ENSG00000225880.4, a gene associated with lung cancer susceptibility. Our motivating dataset is from GTEx, including samples of lung tissue collected from 278 subjects. For each tissue sample, the expression of ENSG00000225880.4 was measured, along with 117 SNPs in the flanking region of this gene within a window of size 20kb[3]. eQTL studies, including ours, often suffer from power limitations[4] with a large number of candidate loci or single nucleotide polymorphisms (SNPs) to detect; variants in neighboring regions of a candidate gene can vary from hundreds to millions[5].

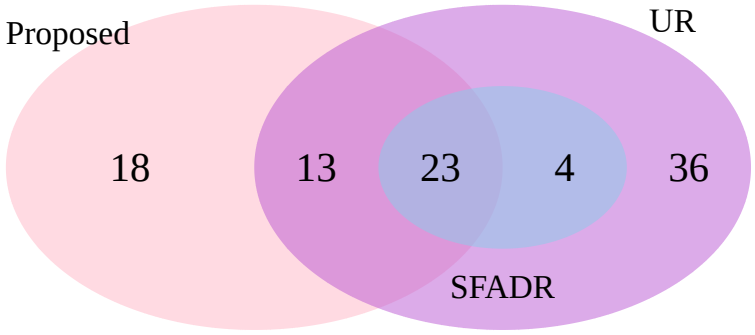
For extracting relevant features out of a massive number of candidates, commonly used are variable selection methods, including Bridge regression[6], Lasso[7], SCAD[8], Elastic net[9], adaptive Lasso[10], and Dantzig selector[11]. Variable screening methods, e.g., Fan and Lv[12], Fan et al.[13], Zhao and Li[14] , Fan et al.[15], Li et al.[16], and Ma et al.[17], have also emerged as a powerful means for effectively eliminating unimportant covariates.

Often, the validity of these selection/screening methods hinges upon a sparsity assumption, e.g., only a limited number of SNPs are associated with gene expressions, a beta-min condition that the effects of signals are well above from 0, and a partial faithfulness assumption, that is, signal variables are at most weakly correlated with noise variables. However, these conditions may fail in our eQTL analysis. For instance, we have applied the adaptive

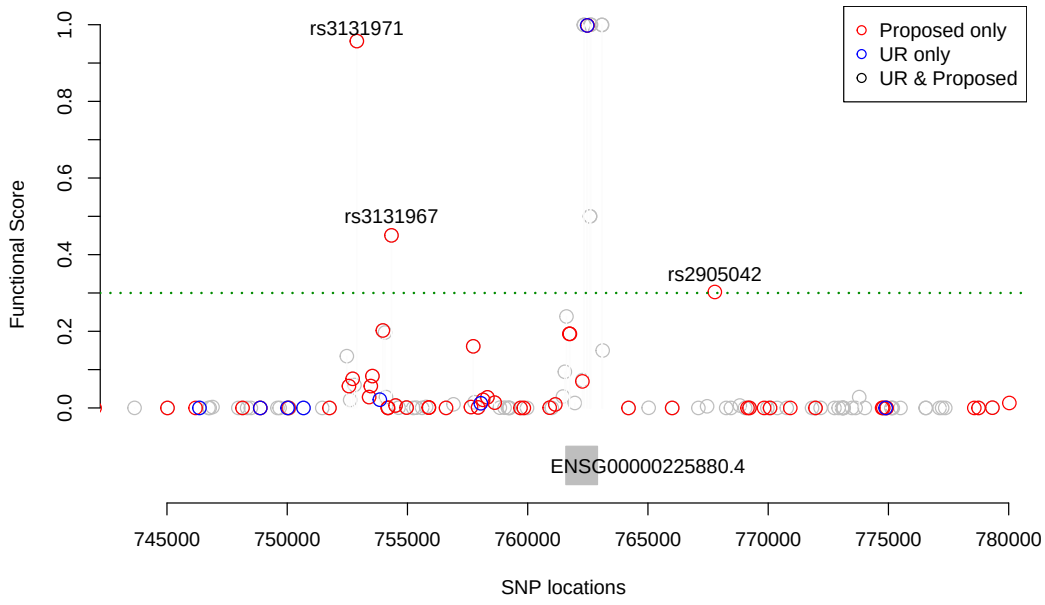
lasso[10] to the aforementioned GTEx data, and found the correlations between the selected and unselected SNPs could be as large as 0.8. Wu et al.[18] used a partial linear model and incorporated correlations among covariates via a network structure, while applying variable selection to determine important variables (i.e., SNPs) under a sparsity assumption.

To circumvent the sparsity assumption, Fan et al.[19], by borrowing deep learning ideas[20,21], proposed a factor model for dimension reduction (FADR) under a linearity condition on the latent variables[22]; to relax the linearity condition, Jiang et al.[3] proposed a semiparametric method, termed semiparametric factor model for dimension reduction (SFADR), which showed improved power in identifying eQTLs and provided useful insight into the dependence among candidate SNPs. However, both Fan et al.[19] and Jiang et al.[3] are non-supervised in the sense that latent factors are extracted without considering responses. Therefore, the obtained latent factors (e.g., linear combinations of SNPs) may not well quantify the dependence between gene expressions and SNPs[23].

We propose supervised structural learning of semiparametric regression models with high-dimensional covariates for eQTL analyses. As opposed to the literature, our method extracts low-dimensional latent features by using the correlations among the predictors as well as the relationships between the response and these predictors. We further adopt a flexible multi-index nonparametric model to capture the dependence between the response and the latent factors and derive a likelihood based estimator to achieve efficient estimation. Our method has several advantages. First, our method does not require the linearity condition on the latent variables. Second, our model allows the distributions of the response and the latent variables, as well as the forecasting function linking the response and the latent factors, to be unspecified. Third, our likelihood-based estimators are efficient, even with unspecified forecasting functions and response distribution functions. Finally, our estimators for high dimensional parameters have a closed-form at each iterative step, greatly facilitating computation.



(a)



(b)

Figure 1: (a) A Venn diagram of detected SNPs after the Bonferroni correction, where UR is for univariate regression; (b) Comparison of the identified eQTLs from UR and the proposed method (Proposed), where the x-axis denotes the location of each SNP, and the y-axis denotes the FUN-LDA functional annotation scores, and the dashed line represents 0.3 in the y-axis. SNPs are colored in red if identified as eQTLs by Proposed only, in blue if by UR only, in black if by both methods and in gray if not by any of them.

To find the eQTLs related to the expression of ENSG00000225880.4, we have applied the proposed method to analyze the aforementioned GTEx data and compared the results with those obtained by univariate association analyses[5] and by SFADR[3]; see Figure 1(a). A total of 18 SNPs are uniquely identified by the new method, while the SNPs identified by SFADR are a subset of those obtained from univariate association analyses. For validation, we have adopted a functional annotation score, termed FUN-LDA[24,25], ranging from 0 to 1 with larger values indicating higher likelihoods to be an eQTL; see Figure 1(b) for the FUN-LDA scores of the candidate SNPs by chromosomal locations, which shows a cluster of SNPs nearing the target gene (the gray box) have high FUN-LDA scores, and moreover, our method empowers detection of eQTLs outside the range of the target gene.

The paper is organized as follows. Section 2 describes the model and the proposed estimation method; Section 3 describes an iterative computational strategy where infinite dimensional parameters can be explicitly expressed in each step and further propose a BIC-type procedure to select the structural dimension; the theoretical properties of the proposed estimator are summarized in Section 4; the performance of the proposed estimation procedure is assessed through simulation studies in Section 5. Section 6 applies our model together with the proposed estimation procedure to analyze the data set regarding the eQTL study in the GTEx project. A brief discussion about further research in this direction is given in Section 7. Technical proofs are relegated to Supplementary Materials.

2 Model and Estimation

Consider n independent subjects. Let y_i be the outcome and $\mathbf{x}_i = (x_{i1}, \cdots, x_{ip})^{\mathrm{T}}$ be the covariates for subject $i = 1, \ldots, n$, where $p > n$. We assume the relationship between y_i and $\mathbf{x}_i$ is fully captured by a latent factor $\mathbf{h}_i$, i.e., y_i is independent with $\mathbf{x}_i$ given $\mathbf{h}_i$, and

stipulate the following models[19,3],

$$\mathbf{x}_i = \mathbf{B}\mathbf{h}_i + \mathbf{u}_i, \tag{1}$$

$$y_i = \psi(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{h}_i, \cdots, \boldsymbol{\beta}_d^{\mathrm{T}}\mathbf{h}_i, \epsilon_i), \tag{2}$$

where ψ is an unknown forecasting function, $\mathbf{h}_i = (h_{i1}, \cdots, h_{iq})^{\mathrm{T}}$ is a q-dimensional vector of latent factors, $\boldsymbol{\beta}_j, j = 1, \cdots, d$ are q-dimensional parameter vectors with $d \leqslant q$, $\mathbf{B} = (\mathbf{b}_1, \cdots, \mathbf{b}_p)^{\mathrm{T}}$ is a $p \times q$ deterministic matrix, $\mathbf{u}_i = (u_{i1}, \cdots, u_{ip})^{\mathrm{T}}$ with $cov(\mathbf{u}_i) = diag(\lambda_1, \cdots, \lambda_p)$ and $\sup_j |\lambda_j| \leqslant M < \infty$, and ϵ_i is a random error independent of $\mathbf{h}_i$ and $\mathbf{u}_i$. We require $d < q \ll p$ and $\mathbf{u}_i$,$\mathbf{h}_i$, and ϵ_i are independent. That is, y_i depends on the latent factors $\mathbf{h}_i$ only through d predictive indices $\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{h}_i, \cdots, \boldsymbol{\beta}_d^{\mathrm{T}}\mathbf{h}_i$. In the following, for compactness we write $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_d)^{\mathrm{T}}$, a $d \times q$ matrix, $\mathbf{H} = (\mathbf{h}_1, \cdots, \mathbf{h}_n)^{\mathrm{T}}$, an $n \times q$ matrix, $\mathbf{y} = (y_1, \cdots, y_n)^{\mathrm{T}}$, a vector, and $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^{\mathrm{T}}$, an $n \times p$ matrix.

Models (1) and (2) jointly describe the relationship between y_i and $\mathbf{x}_i$ by the quantity[3] $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_p)^T \equiv \mathbf{B}(\mathbf{B}^{\mathrm{T}}\mathbf{B})^{-1}\boldsymbol{\beta} \in R^{p\times d}$. By testing $H_0 : \boldsymbol{\alpha}_j = 0$, we identify whether variable j is significantly associated with y_i. If it is, the effect strength of x_{ij} can be measured by the norm of $\boldsymbol{\alpha}_j$, which ranks SNPs in the rest of the paper.

To make (1) and (2) identifiable, we impose three constraints: (E1) the upper $d \times d$ of $\boldsymbol{\beta}^{\mathrm{T}}$ is an identity matrix; $n^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{H} = \mathbf{I}_q$; (E2) $\mathbf{B}^{\mathrm{T}}\mathbf{B}$ is diagonal with decreasing diagonal elements; and (E3) the first non-zero element in each column of $\mathbf{B}$ is positive. Furthermore, without loss of generality, we assume $E(x_{ij}) = 0$ and $E(\mathbf{h}_i) = 0$.

Fitting models (1) and (2) is challenging as both ψ and $\mathbf{h}_i$ are unknown. Since (1) is a classical factor model, Fan et al.[19] and Jiang et al.[3] fitted (1) and (2) via a sequential approach: they first estimated latent factors $\mathbf{h}_i$ via principal component analysis (PCA); then they fitted (2) by plugging in the estimated $\mathbf{h}_i$. This unsupervised method, which estimates latent factors while ignoring their relationships with the response, may not fully

capture the dependence of responses on latent factors. On the other hand, estimating latent factors while accounting for their associations with y_i is complicated by that ψ is unknown. We address these issues by proposing a supervised learning method, which builds upon $f(y_i \mid \boldsymbol{\beta}\mathbf{h}_i)$, the conditional density of y_i given $\boldsymbol{\beta}\mathbf{h}_i$. We estimate $\Omega = (\mathbf{B}, \mathbf{H}, \boldsymbol{\beta})$ by maximizing

$$l(\Omega; \mathbf{y}, \mathbf{X}) = \sum_{i=1}^{n} \log\{\widehat{f}(y_i \mid \boldsymbol{\beta}\mathbf{h}_i)\} - w\|\mathbf{X} - \mathbf{H}\mathbf{B}^{\mathrm{T}}\|_F^2, \tag{3}$$

where $\widehat{f}(y_i \mid \boldsymbol{\beta}\mathbf{h}_i)$ is the kernel smoothing estimate of $f(y_i \mid \boldsymbol{\beta}\mathbf{h}_i)$. That is,

$$\widehat{f}(y_i \mid \boldsymbol{\beta}\mathbf{h}_i) = \frac{\sum_{j=1}^{n} K_{b_y}(Y_j - y_i)\mathcal{K}_b(\boldsymbol{\beta}\mathbf{h}_j - \boldsymbol{\beta}\mathbf{h}_i)}{\sum_{j=1}^{n} \mathcal{K}_b(\boldsymbol{\beta}\mathbf{h}_j - \boldsymbol{\beta}\mathbf{h}_i)},$$

where $K_{b_y}(y) = K(y/b_y)/b_y$ is a kernel function with a bandwidth b_y and $\mathcal{K}_b(\mathbf{v}) = 1/b^d \prod_{l=1}^{d} K(v_l/b)$ is a product kernel function with a bandwidth b for a vector $\mathbf{v} = (v_1, \cdots, v_d)^{\mathrm{T}}$. For simplicity of notation, the bandwidth b is not variable-specific; based on our numerical experience, component-specific bandwidths make little differences.

Models (1) and (2) imply $f(y_i, \mathbf{x}_i|\mathbf{h}_i) = f(y_i|\boldsymbol{\beta}\mathbf{h}_i)f(\mathbf{x}_i|\mathbf{h}_i) = f(y_i|\boldsymbol{\beta}\mathbf{h}_i)\Pi_{j=1}^{p} f(x_{ij}|\mathbf{h}_i)$. If $u_{ij} \sim N(0, \sigma^2)$, the loglikelihood can be written as $\sum_{i=1}^{n} \log\{f(y_i|\boldsymbol{\beta}\mathbf{h}_i)\} - \frac{1}{2\sigma^2}\|\mathbf{X} - \mathbf{H}\mathbf{B}^{\mathrm{T}}\|_F^2 - C$, which is equal to (3) up to a constant C. Effectively, (3) generalizes the normal likelihood function and does not require specifying the distributions of y_i or $\mathbf{u}_i$.

The first term on the right side of (3) characerizes the dependence of y_i on $\mathbf{h}_i$. Without it, our estimator for $\mathbf{h}_i$ would degenerate to the PCA-based estimator used in[19] and[3]; by including it in $l(\Omega; \mathbf{y}, \mathbf{X})$, our supervised estimator for $\mathbf{h}_i$ makes full use of both the information among $\mathbf{x}_i$ as well as the dependence of y_i on $\mathbf{x}_i$.

3 Implementation

3.1 An Algorithm to Estimate B and H Alternatingly

As $l(\Omega; \mathbf{y}, \mathbf{X})$ involves the high dimensional parameters of $\mathbf{B}$ and $\mathbf{H}$, a direct maximization can be prohibitive. We design an iterative algorithm to alternatingly estimate $\mathbf{B}$ and $\mathbf{H}$ at each step.

We first obtain the initial value $(\mathbf{B}^{(0)}, \mathbf{H}^{(0)})$ for $(\mathbf{B}, \mathbf{H})$ from the principal component method[19], and then obtain the initial value $\boldsymbol{\beta}^{(0)}$ for $\boldsymbol{\beta}$ from directional regression method[26] given $(y_i, \mathbf{h}_i^{(0)}), i = 1, \ldots, n$. Denote by $\mathbf{B}^{(k-1)}, \mathbf{H}^{(k-1)}$ and $\boldsymbol{\beta}^{(k-1)}$ the estimates of $\mathbf{B}, \mathbf{H}$ and $\boldsymbol{\beta}$ obtained after the $(k-1)$th iteration, respectively. In the kth iteration, we update these estimates alternatingly.

- Updating $\mathbf{B}$. Differentiating $l(\Omega; \mathbf{y}, \mathbf{X})$ with respect $\mathbf{b}_j, j = 1, \cdots, p$ and setting the derivatives to zero leads to the following estimation equations,

$$\mathbf{b}_j = \left(\sum_{i=1}^{n} \mathbf{h}_i \mathbf{h}_i^{\mathrm{T}} \right)^{-1} \sum_{i=1}^{n} x_{ij} \mathbf{h}_i, j = 1, \cdots, p.$$

  With the identification condition of $n^{-1} \sum_{i=1}^{n} \mathbf{h}_i \mathbf{h}_i^{\mathrm{T}} = \mathbf{I}_q$, we have

$$\mathbf{B}_n = \frac{1}{n} \mathbf{X} \mathbf{H}^{(k-1)}.$$

  To adhere to the identification condition on $\mathbf{B}$, we further perform a singular value decomposition (SVD) to get $\mathbf{B}_n = S_1 \Lambda^{1/2} D_1$ and update $\mathbf{B}^{(k-1)}$ by $\mathbf{B}^{(k)} = S_1 \Lambda^{1/2}$. It is easy to see that ${\mathbf{B}^{(k)}}^{\mathrm{T}} \mathbf{B}^{(k)} = \Lambda$ is a diagonal matrix with decreasing elements and $\mathbf{B}^{(k)}$ is a $p \times q$ matrix.

- Updating $\mathbf{H}$. Differentiating $l(\Omega; \mathbf{y}, \mathbf{X})$ with respect $\mathbf{h}_i, i = 1, \cdots, n$, and setting the

derivatives to zero leads to the following estimation equations:

$$\mathbf{h}_i = \left(\sum_{j=1}^{p} \mathbf{b}_j \mathbf{b}_j^{\mathrm{T}}\right)^{-1} \left\{ \frac{\boldsymbol{\beta}^{\mathrm{T}} \widehat{f}^{(01)}(y_i \mid \boldsymbol{\beta}\mathbf{h}_i)}{2w\widehat{f}(y_i \mid \boldsymbol{\beta}\mathbf{h}_i)} + \sum_{j=1}^{p} x_{ij}\mathbf{b}_j \right\}, i = 1, \cdots, n,$$

where $\widehat{f}^{(01)}(y_i \mid \boldsymbol{\beta}\mathbf{h}_i) = \partial \widehat{f}(y_i \mid \boldsymbol{\beta}\mathbf{h}_i)/\partial(\boldsymbol{\beta}\mathbf{h}_i)$. Thus we obtain

$$\mathbf{H}_n = \left\{ \tilde{M}(\boldsymbol{\beta}^{(k-1)}, \mathbf{H}^{(k-1)})/(2w) + \mathbf{X}\mathbf{B}^{(k)} \right\} (\mathbf{B}^{(k)^{\mathrm{T}}}\mathbf{B}^{(k)})^{-1}.$$

where $\tilde{M}(\boldsymbol{\beta}, \mathbf{H}) = \{m_1(\boldsymbol{\beta}, \mathbf{h}_1)^{\mathrm{T}}, m_2(\boldsymbol{\beta}, \mathbf{h}_2)^{\mathrm{T}}, \ldots, m_n(\boldsymbol{\beta}, \mathbf{h}_n)^{\mathrm{T}}\}^{\mathrm{T}}$ with $m_i(\boldsymbol{\beta}, \mathbf{h}_i) = \boldsymbol{\beta}^{\mathrm{T}} \widehat{f}^{(01)}(y_i \mid \boldsymbol{\beta}\mathbf{h}_i)/\widehat{f}(y_i \mid \boldsymbol{\beta}\mathbf{h}_i)$. Likewise, with the identification condition of $\mathbf{H}$, we perform SVD to get $\mathbf{H}_n = S_2 V D_2$ and update $\mathbf{H}^{(k-1)}$ by $\mathbf{H}^{(k)} = \sqrt{n} S_2$. It is easy to see that $\mathbf{H}^{(k)\prime}\mathbf{H}^{(k)} = n\mathbf{I}_q$.

- Updating $\boldsymbol{\beta}$ with

$$\boldsymbol{\beta}^{(k)} = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} \log\{\widehat{f}(y_i \mid \boldsymbol{\beta}\mathbf{h}_i^{(k)})\}, \tag{4}$$

subject to that the upper $d \times d$ block of $\boldsymbol{\beta}^{(k)^{\mathrm{T}}}$ is an identity matrix. We use the matlab function *fmincon* to carry out the constrained optimization, which incurs little computing time.

We repeat the iterations until convergence, that is, $|l(\Omega^{(k)}; \mathbf{y}, \mathbf{X}) - l(\Omega^{(k-1)}; \mathbf{y}, \mathbf{X})| \leqslant a_0$, where a_0 is a prespecified small number. We denote the final estimates by $(\widehat{\mathbf{B}}, \widehat{\mathbf{H}}, \widehat{\boldsymbol{\beta}})$.

With large p and n, calculating $\mathbf{B}$ and $\mathbf{H}$ can constitute the main bulk of computation. However, the closed-form updates of $\mathbf{B}$ and $\mathbf{H}$ at each step have made our algorithm computationally efficient.

3.2 Tuning

We use the BIC criterion or cross-validation to select tuning parameters, including the bandwidth parameters (b, b_y), the dimension of latent factors q, number of indices d and weighted parameter w. We detail the implementation for each of them. We also utilize simulations in Section 5 to verify the utility of our selection procedures.

We use the "kde" function in MATLAB to estimate the multivariate density function $f(Y|\boldsymbol{\beta}\mathbf{h})$, which provided several data-driven bandwidth selection methods, including likelihood cross validation, local cross validation, rule of thumb and plug-in estimator of[27], etc; see the details at `https://github.com/feiyoung/npReg/tree/master/%40kde`. Because of its satisfactory performance, we have opted to use the rule of thumb method for bandwidth selection in our implementation.

We selected the structural dimension (q, d) based on a BIC-type procedure[28,29]. Specifically, we select the optimal (q, d) through maximizing

$$\mathrm{BIC}(q,d) = \sum_{i=1}^{n} \log(\widehat{f}(Y_i \mid \widehat{\mathbf{h}}_i\widehat{\boldsymbol{\beta}})) - \frac{1}{2} df(d,q)\frac{\log(n)}{n}, \tag{5}$$

where $df(d, q)$ is the degree of freedom. To calculate $df(q, d)$, we approximate Y_i via $Y_i \approx \sum_{j=1}^{d} \gamma_{j,j}(\boldsymbol{\beta}_j^{\mathrm{T}}\mathbf{h}_i)^2 + \sum_{j>k} \gamma_{j,k}(\boldsymbol{\beta}_j^{\mathrm{T}}\mathbf{h}_i\boldsymbol{\beta}_k^{\mathrm{T}}\mathbf{h}_i) + \sum_{j=1}^{d} \gamma_j\boldsymbol{\beta}_j^{\mathrm{T}}\mathbf{h}_i + \gamma_0$ and get the fitted values $\widehat{Y}_i$ of $Y_i, i = 1, \ldots, n$. Then we follow the method of Ye[30] to evaluate $df(q, d)$ by $\hat{df}(q, d) = \sum_{i=1}^{n} \partial\widehat{Y}_i/\partial Y_i$ where $\partial\widehat{Y}_i/\partial Y_i$ is estimated by local linear regression of $\widehat{Y}_i$ on Y_i to obtain the derivative estimates. In simulation studies and in the analysis of real data example, we perform the selection of q and d on grids of the possible structural dimensions.

The weight w was chosen by maximizing quasi-loglikelihood $\sum_{i=1}^{n} \log\{\widehat{f}(y_i \mid \widehat{\boldsymbol{\beta}}(w)\widehat{\mathbf{h}}_i(w))\} - w\sum_{i=1}^{n}\sum_{j=1}^{p}(x_{ij} - \widehat{\mathbf{b}}_j(w)^{\mathrm{T}}\widehat{\mathbf{h}}_i(w))^2 + \log(w)np/2$ on the grid $\{p^{-t}, t \in [0, 1)\}$. We test the performance of our tuning procedure via simulation studies in Section 5, which show that the selection procedure works well.

Remark 1. Since the estimator for $\boldsymbol{\beta}$ is an estimated maximum likelihood estimator, nice theoretical properties hold. Specifically, in Theorem 2, we show that the proposed estimator for $\boldsymbol{\beta}$ is consistent and asymptotical normal, and the asymptotic variance is the same as the one obtained based on maximum likelihood function where the true $f(Y \mid \boldsymbol{\beta}\mathbf{h})$ is known. Thus, the proposed algorithm possesses not only computational simplicity but also estimation efficiency.

4 Theoretical Properties

We now establish the large sample properties, including the model identifiability, consistency and the asymptotic normality, as well as the efficiency of the proposed estimator for $\boldsymbol{\beta}$. Their proofs are deferred to Supplementary Materials. To establish the asymptotic properties, we need extra notations and conditions. Throughout, we let $\mathbf{h}_{i0}$ be the vector of the true factors and $\mathbf{b}_{j0}$be the true loadings, with $\mathbf{H}_0$ and $\mathbf{B}_0$ being the corresponding matrices. Write the true value of $\boldsymbol{\beta}$ as $\boldsymbol{\beta}_0$, and write $\lambda_{min}(\mathbf{M})$ and $\lambda_{max}(\mathbf{M})$ as the minimum and maximum eigenvalues of a symmetric matrix $\mathbf{M}$, respectively. Let $\Gamma = \lim_{p\to\infty} p^{-1} \sum_{j=1}^{p} \lambda_j \mathbf{b}_{j0}\mathbf{b}_{j0}^{\mathrm{T}}$ and $\Sigma_\Lambda = \lim_{p\to\infty} p^{-1}\mathbf{B}_0^{\mathrm{T}}\mathbf{B}_0$. Let $\pi_{\mathbf{h}}(\cdot)$ be the pdf of $\mathbf{h}_{i0}$, $\pi(\cdot)$ be the pdf of $\boldsymbol{\beta}_0\mathbf{h}_{i0}$, $f(\cdot,\cdot)$ be the joint pdf of $(y_i, \boldsymbol{\beta}_0\mathbf{h}_{i0})$, $f(\cdot \mid \cdot)$ be the conditional pdf of y_i given $\boldsymbol{\beta}_0\mathbf{h}_{i0}$ and $f^{(01)}(y \mid \mathbf{v}) = \partial f(y \mid \mathbf{v})/\partial \mathbf{v}$.

First, we state the identifiability in the following proposition.

Proposition 1. *If*

(A1) $(\mathbf{H}_0, \mathbf{B}_0, \boldsymbol{\beta}_0)$ *satisfy the identifiability conditions (E1) – (E3).*

(A2) *there exists a constant* M*, independent of* p *and* n*, such that* $E(||\mathbf{h}_{i0}||_2^4) \leqslant M$*,* $\sup_j ||\mathbf{b}_{j0}||_2 \leqslant M$*, and* $E(u_{ij}^8) \leqslant M$ *for all* $j = 1, \cdots, p$*;*

(A3) $p^{-1/2} \sum_{j=1}^{p} \mathbf{b}_{j0} u_{ij} \xrightarrow{d} N(0, \Gamma)$ *as* $p \to \infty$*;*

(A4) *there are two positive constant* c_1, c_2 *such that* $c_1 < \lambda_{min}(\Sigma_\Lambda) < \lambda_{max}(\Sigma_\Lambda) < c_2$

and Condition (C4) *in the following hold, then* $\mathbf{B}_0, \mathbf{H}_0$ *and* $\boldsymbol{\beta}_0$ *are unique when* $p \to \infty$.

In[3], they require boundedness of two type of norm about $\mathbf{b}_j$, which is not necessary due to the equivalence of norm. In addition, some requirements on the random variable $\mathbf{u}_i$ such as, for every i, j, $p^{-1/2}\sum_{l=1}^{p}|u_{il}u_{jl} - E(u_{il}u_{jl})|^4 \leqslant M$, are restrictive and may be infeasible in practice. In fact, we only need the conclusion $p^{-1}\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_i = O_p(1)$ which can be obtained from Conditions (A2) in Proposition 1. Finally, we need an extra condition (C4) to ensure the existence of the proposed estimators.

Then we give the conditions for the asymptotical properties of $\widehat{\boldsymbol{\beta}}$.

(C1) The univariate kernel function $K(\cdot)$ is Lipschitz, has compact support and satisfies

$$\int K(v)dv = 1, \int v^s K(v)dv = 0, 1 \leqslant s \leqslant r-1, 0 \neq \int v^r K(v)dv < \infty,$$

i.e. $K(\cdot)$ has order of r. The derivative $\dot{K}(v) = dK(v)/dv$ is Lipschitz continuous. The d-dimensional kernel function is a product of d univariate kernel functions, i.e. $\mathcal{K}_b(\mathbf{v}) = \mathcal{K}(\mathbf{v}/b)/b^d = \Pi_{l=1}^{d}K_b(v_l)\hat{=}\Pi_{l=1}^{d}K(v_l/b)/b^d$ for $\mathbf{v} = (v_1, v_2, \ldots, v_d)^{\mathrm{T}}$.

(C2) $\pi_{\mathbf{h}}(\cdot)$ and $\pi(\cdot)$ are bounded from zero and infinity. The functions $\pi(\mathbf{v})$ and $f(y, \mathbf{v})$ have $(r+1)$th order derivatives and their $(r+1)$th derivatives are locally Lipschitz continuous.

(C3) $p \to \infty$, $n^{1/2}p^{-1} \to 0$ and $wp \to \infty$ as $n \to \infty$, the bandwidths satisfy $b = o(n^{-1/(4r)})$, $b_y = o(n^{-1/(4r)})$ and $(b_y b^{d+2} n^{1/2})/\log^2(n) \to \infty$.

(C4) $E\left[f^{(01)}(y_i \mid \boldsymbol{\beta}\mathbf{h}_{i0}) \otimes \mathbf{h}_{i0}/f(y_i \mid \boldsymbol{\beta}\mathbf{h}_{i0})\right]$ is a smooth vector function of $\boldsymbol{\beta}$, has unique root and has non-singular derivative matrix at $\boldsymbol{\beta}_0$.

Estimation of the derivative $f^{(01)}(\cdot \mid \cdot)$ is required for estimating $\boldsymbol{\beta}$ and $\mathbf{H}$. Non-parametric estimators for the derivative functions converge more slowly than the function

estimate itself, so a higher-order kernel $K(\cdot)$[31], described in Condition (C1), is needed here to ensure sufficient convergence rate. Condition (C2) contains some mild requirements to simplify the mathematical derivation. Condition (C3) puts some restrictions on the bandwidth choice in relation to the dimension d and the sample size n, and on w. When p is large enough, the mild Condition $wp \to \infty$ in (C3) avoids the issues of over-fitting bias and inflated type I error due to using the information of y_i to estimate $\mathbf{h}_i$. Condition (C4) guarantees identifiability for $\boldsymbol{\beta}$ and the existence of the asymptotic variance for $\widehat{\boldsymbol{\beta}}$.

Remark 2. Condition (C3) indicates p can diverge with n at any rate that is faster than $\sqrt{n}$, including exponential rate. In our simulations, we also included the cases with $n = 300$ while $p = 500$ and 1000, which shows that the performance of proposed method gets better as p increases.

Theorem 1. *Under Conditions (C1)-(C4) and the same other Conditions in Proposition 1,* $\widehat{\boldsymbol{\beta}} \to \boldsymbol{\beta}_0$ *in probability when* $n \to \infty$.

Theorem 2. *Under Conditions (C1)-(C4) and the same other Conditions in Proposition 1,* $\sqrt{n}vecl(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \to N(0, \mathbf{T}_0^{-1})$ *in distribution when* $n \to \infty$. *Here,* $vecl(\mathbf{M})$ *is the vector formed by concatenating the columns of the right* $d \times (q-d)$ *block of the* $d \times q$ *matrix* $\mathbf{M}$, $\mathbf{T}_0 = E\left[\{\partial l_0(\boldsymbol{\beta}; y_i, \mathbf{h}_{i0})/\partial vecl(\boldsymbol{\beta})\}\{\partial l_0(\boldsymbol{\beta}; y_i, \mathbf{h}_{i0})/\partial vecl(\boldsymbol{\beta})\}^{\mathrm{T}}_{|\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right]$ *and* $l_0(\boldsymbol{\beta}; y_i, \mathbf{h}_{i0}) = vecl\{f^{(01)}(y_i \mid \boldsymbol{\beta}\mathbf{h}_{i0})\mathbf{h}_{i0}^{\mathrm{T}}/f(y_i \mid \boldsymbol{\beta}\mathbf{h}_{i0})\}$. *The asymptotic covariance matrix of the proposed estimator* $\widehat{\boldsymbol{\beta}}$ *achieves the efficient estimation variance bound hence is efficient.*

Remark 3. Our objective function (3) can be regarded as a penalized likelihood function, where the first term is the likelihood and the second term is the penalty function. Note that the penalty term is not on $\boldsymbol{\beta}$ so is not directly linked to the estimator for $\boldsymbol{\beta}$. The estimator for $\boldsymbol{\beta}$ is based on the first term, which is the likelihood function alone, so that the estimator for $\boldsymbol{\beta}$ is efficient, although the likelihood function itself depends on $\mathbf{h}_i, i = 1, \cdots, n$. We proved that as long as $\|\widehat{\mathbf{h}}_i - \mathbf{h}_{i0}\| = o_p(n^{-1/4})$, the $\sqrt{n}$-consistency and asymptotical

efficiency of $\widehat{\boldsymbol{\beta}}$ can be ensured. Under proper choice of weight w, e.g. $wp \to \infty$, we have $\widehat{\mathbf{h}}_i - \mathbf{h}_{i0} = O_p(p^{-1/2} + n^{-1})$ by Lemma 3 of Supplementary Materials. Combining with $n^{1/2}p^{-1} \to 0$ in Condition (C3), $\widehat{\mathbf{h}}_i$ does satisfy the required condition.

Remark 4. We note that the results in Theorem 2 do not depend on the bandwidths b and b_y, hence the bandwidths, as long as they are in the range specified by Condition (C3), is not crucial for the asymptotic performance of the estimate. This observation is confirmed by our simulation studies. A practical implication of this result is that our estimates are not sensitive to the bandwidths choice, which greatly simplifies the practical implementation of our method. A rough selecting method for b and b_y is enough.

5 Numerical Studies

We conduct simulations to compare the finite-sample performance of the proposed method with those of Factor Analysis and Dimension Reduction (FADR[19]) and its semiparametric version (SFADR[3]), in combination with various dimension reduction techniques, including sliced inverse regression (SIR[22]), principal Hessian directions (PHD[32]), directional regression (DR[26]), sliced average variance estimation (SAVE[33]); these estimators are labeled as FADR-SIR, FADR-PHD, FADR-DR, FADR-SAVE, SFADR-SIR, SFADR-PHD, SFADR-DR, and SFADR-SAVE. To assess the efficiency of the proposed method, we compare it with the "Oracle" estimator, where the density $f(Y|\boldsymbol{\beta}\mathbf{h})$ is known.

As outlined in Section 3.2, we also investigate the performance of the methods for choosing (q, d), which are important for the performance of the proposed method. Define the Euclidean distance between the resulting estimators $\widehat{\boldsymbol{\beta}}$ and the true values $\boldsymbol{\beta}_0$ as $\text{dist}(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}_0) = \| \widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})^{-1}\widehat{\boldsymbol{\beta}}^{\mathrm{T}} - \boldsymbol{\beta}_0({\boldsymbol{\beta}_0}^{\mathrm{T}}\boldsymbol{\beta}_0)^{-1}{\boldsymbol{\beta}_0}^{\mathrm{T}} \|_F$. We evaluate the performance of the estimators via Euclidean distance (ED), biases, standard deviation (SD), and the root of the mean squared errors (RMSE), based on 1000 replications of the data $(\mathbf{x}_i, y_i), i = 1, \cdots, n = 300$.

5.1 Simulation 1: estimation accuracy

5.1.1 Simulation settings

We give the settings for the factor model (1). First, to construct $\mathbf{B}_0$, we generate n independent p dimensional random vectors, $\mathbf{z}_i$'s, from a multivariate normal distribution with mean zero and covariance matrix $(\sigma_{ij})_{p\times p}$ with $\sigma_{ij} = 0.5^{|i-j|}$. Let $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)^{\mathrm{T}}$. Then we perform eigen-decomposition on $\mathbf{Z}\mathbf{Z}^{\mathrm{T}}$, retain the $n \times q$ orthogonal matrix $\mathbf{E}$ that spans the eigenspace corresponding to the first q largest eigenvalues, and define $\mathbf{B}_0 = \sqrt{1/6}\mathbf{Z}^{\mathrm{T}}\mathbf{E}$. It holds that $\mathbf{B}_0^{\mathrm{T}}\mathbf{B}_0$ is a diagonal matrix with decreasing diagonal entries. We next consider two versions of factor models for X: Xmodel I ensures $\mathbf{x}_i$ satisfy the linear condition[22], while Xmodel II does not. In all these factor models, $\mathbf{u}_i \sim N(0, 0.1\mathbf{I}_p)$.

Xmodel I: $\mathbf{h}_{i0}$ is from a multivariate normal distribution with mean zero and covariance matrix $(\sigma_{ij})_{q\times q}$ with $\sigma_{ij} = 0.5^{|i-j|}$.

Xmodel II: $\mathbf{h}_{i0} = (h_{i1}, \cdots, h_{iq})^{\mathrm{T}}$, where $(h_{i1}, h_{i2})^{\mathrm{T}}$ are generated from two-dimensional normal distribution with mean zero and covariance matrix $(\sigma_{ij})_{2\times 2}$ with $\sigma_{ij} = 0.5^{|i-j|}$; $h_{i3} = |h_{i1} + h_{i2}| + h_{i1}\xi_{i1}$; $h_{i4} = |h_{i1} + h_{i2}|^2 + |h_{i2}|\xi_{i2}$, where ξ_{i1}, ξ_{i2} are independently generated from the standard normal distribution; h_{i5} is generated from a Bernoulli distribution with success probability $\exp(h_{i2})/\{1 + \exp(h_{i2})\}$, and h_{i6} is a Bernoulli random variable with success probability $\Phi(h_{i2})$, where Φ is the standard normal distribution function. Then we center and normalize $\mathbf{H}_0 = (\mathbf{h}_{10}, \cdots, \mathbf{h}_{n0})^{\mathrm{T}}$ so that $\mathbf{H}_0$ satisfies the identification of $\mathbf{H}_0$ described in Section 2.

We next present the settings for the sufficient dimension-reduction model (2) for Y, termed Ymodel.

Ymodel I: $y_i = (\mathbf{h}_{i0}^{\mathrm{T}}\boldsymbol{\beta}_1)^2 + 2|\mathbf{h}_{i0}^{\mathrm{T}}\boldsymbol{\beta}_2 + 1| + 0.1(\mathbf{h}_{i0}^{\mathrm{T}}\boldsymbol{\beta}_1)^2\epsilon_i$,

Ymodel II: $y_i = \sin(\mathbf{h}_{i0}^{\mathrm{T}}\boldsymbol{\beta}_1) + 2\cos(\mathbf{h}_{i0}^{\mathrm{T}}\boldsymbol{\beta}_2) + 0.1\epsilon_i$,

Table 1: Bias, SD and RMSE of $\hat{\boldsymbol{\beta}}$ with Xmodel I, Ymodel I, $n = 300$ and $p = 500$.

		Oracle	Proposed	SF-SIR	F-SIR	SF-PHD	F-PHD	SF-DR	F-DR	SF-SAVE	F-SAVE
$\boldsymbol{\beta}_{13}$	Bias	-3e-4	3e-3	2e-3	0.28	2e-3	0.24	5e-3	0.03	-4e-3	0.05
	SD	0.07	0.08	0.08	7.76	0.14	3.02	0.09	0.76	0.13	1.57
	RMSE	0.07	0.08	0.08	7.77	0.14	3.03	0.09	0.76	0.13	1.57
$\boldsymbol{\beta}_{14}$	Bias	-3e-4	1e-4	2e-3	-0.02	-2e-3	0.20	2e-4	0.03	7e-4	0.04
	SD	0.08	0.08	0.08	12.79	0.14	4.50	0.09	0.58	0.12	1.87
	RMSE	0.08	0.08	0.08	12.79	0.14	4.51	0.09	0.58	0.12	1.87
$\boldsymbol{\beta}_{15}$	Bias	-4e-4	1e-3	4e-3	-0.28	-0.01	0.15	4e-4	0.05	4e-3	0.02
	SD	0.08	0.08	0.10	2.49	0.14	3.19	0.09	0.65	0.13	1.40
	RMSE	0.08	0.08	0.10	2.50	0.14	3.19	0.09	0.66	0.13	1.41
$\boldsymbol{\beta}_{16}$	Bias	-2e-4	-2e-3	4e-3	-1.07	-9e-3	0.21	1e-3	0.04	3e-3	0.06
	SD	0.08	0.08	0.08	6.09	0.13	3.32	0.09	0.59	0.12	1.93
	RMSE	0.08	0.08	0.08	6.18	0.13	3.32	0.09	0.60	0.12	1.93
$\boldsymbol{\beta}_{23}$	Bias	2e-4	-3e-3	2e-3	0.72	-4e-3	-0.08	2e-3	-0.19	2e-3	-0.04
	SD	0.07	0.07	0.19	3.27	0.10	1.05	0.10	3.73	0.09	0.72
	RMSE	0.07	0.07	0.19	3.35	0.10	1.05	0.10	3.74	0.09	0.72
$\boldsymbol{\beta}_{24}$	Bias	-6e-4	3e-3	9e-3	-0.11	5e-3	0.13	2e-4	0.10	-9e-4	0.05
	SD	0.06	0.07	0.17	5.08	0.10	2.08	0.11	1.80	0.09	0.67
	RMSE	0.06	0.07	0.17	5.08	0.10	2.08	0.11	1.80	0.09	0.67
$\boldsymbol{\beta}_{25}$	Bias	4e-04	-4e-3	1e-3	0.10	-4e-4	-0.09	-4e-3	-0.03	-8e-4	-0.03
	SD	0.06	0.07	0.18	1.86	0.10	1.04	0.11	2.85	0.09	0.67
	RMSE	0.06	0.07	0.18	1.86	0.10	1.05	0.11	2.85	0.09	0.67
$\boldsymbol{\beta}_{26}$	Bias	-5e-4	3e-3	9e-3	0.01	2e-4	0.14	4e-3	0.10	-1e-4	0.06
	SD	0.07	0.07	0.21	2.58	0.10	1.92	0.10	1.87	0.09	0.75
	RMSE	0.07	0.07	0.21	2.58	0.10	1.92	0.10	1.88	0.09	0.75

where $\epsilon_i \sim N(0,1), \boldsymbol{\beta}_0 = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)^T \in R^{2\times 6}$ with $\boldsymbol{\beta}_1 = (1,0,1,1,1,1)^T, \boldsymbol{\beta}_2 = (0,1,-1,1,-1,1)^T$.

5.1.2 Comparison Results

We summarize the results of the comparisons of our method with the Oracle method, and various FADR and SFADR estimators, including SFADR-SIR, FADR-SIR, SFADR-PHD, FADR-PHD, SFADR-DR, FADR-DR, SFADR-SAVE and FADR-SAVE (further abbreviated in the tables and figures as SF-SIR, F-SIR, SF-PHD, F-PHD, SF-DR, F-DR, SF-SAVE and F-SAVE, respectively). With $n = 300$ and $p = 500, 1000$ and based on 1000 replicates, Tables 1–4 present biases, SDs and RMSEs of $\widehat{\boldsymbol{\beta}}$ for (Xmodel I, Ymodel I) and (Xmodel II, Ymodel I), and Figure 2 illustrates the ED of $\widehat{\boldsymbol{\beta}}$ for (Xmodel I, Ymodel II) and (Xmodel II, Ymodel II) obtained by using these methods. The evaluation of $\beta_{11}, \beta_{12}, \beta_{21}$ and β_{22} is not provided since they are constrained by the identifiability conditions.

Table 2: Bias, SD and RMSE of $\hat{\boldsymbol{\beta}}$ with Xmodel I, Ymodel I, $n = 300$ and $p = 1000$.

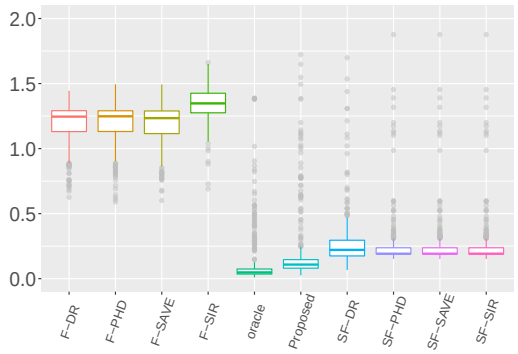
		Oracle	Proposed	SF-SIR	F-SIR	SF-PHD	F-PHD	SF-DR	F-DR	SF-SAVE	F-SAVE
$\boldsymbol{\beta}_{13}$	Bias	-2e-3	4e-3	4e-3	0.68	-5e-4	0.18	2e-3	0.05	-2e-3	0.10
	SD	0.08	0.08	0.09	4.18	0.13	2.65	0.09	0.73	0.13	0.67
	RMSE	0.08	0.08	0.09	4.24	0.13	2.65	0.09	0.73	0.13	0.68
$\boldsymbol{\beta}_{14}$	Bias	-2e-3	-1e-3	7e-4	0.64	-9e-3	0.19	-1e-3	0.02	-1e-4	0.08
	SD	0.07	0.08	0.09	7.63	0.13	2.38	0.09	0.49	0.12	0.65
	RMSE	0.07	0.08	0.09	7.66	0.13	2.39	0.09	0.49	0.12	0.66
$\boldsymbol{\beta}_{15}$	Bias	-2e-3	5e-3	7e-3	-0.17	-3e-3	0.17	3e-3	0.03	-0.01	0.11
	SD	0.08	0.08	0.09	1.21	0.14	2.72	0.09	0.63	0.12	0.74
	RMSE	0.08	0.08	0.09	1.22	0.14	2.73	0.09	0.63	0.12	0.75
$\boldsymbol{\beta}_{16}$	Bias	-7e-4	6e-4	1e-3	-1.30	-2e-3	0.18	8e-4	0.03	-5e-3	0.09
	SD	0.08	0.08	0.09	1.99	0.14	2.56	0.09	0.46	0.12	0.66
	RMSE	0.08	0.08	0.09	2.38	0.14	2.57	0.09	0.46	0.12	0.67
$\boldsymbol{\beta}_{23}$	Bias	2e-3	-1e-3	-8e-3	0.10	6e-4	-0.08	-2e-4	-0.23	2e-3	-0.04
	SD	0.07	0.07	0.10	2.46	0.09	0.86	0.10	1.94	0.08	0.60
	RMSE	0.07	0.07	0.10	2.45	0.09	0.86	0.10	1.95	0.08	0.60
$\boldsymbol{\beta}_{24}$	Bias	-2e-3	3e-3	0.01	5e-3	-9e-4	0.06	2e-3	0.21	3e-4	0.04
	SD	0.06	0.07	0.12	2.18	0.09	1.07	0.10	1.49	0.08	0.70
	RMSE	0.06	0.07	0.13	2.18	0.09	1.08	0.10	1.51	0.08	0.70
$\boldsymbol{\beta}_{25}$	Bias	2e-3	-5e-4	-0.01	0.14	4e-3	-0.06	1e-3	-0.21	-4e-3	-0.04
	SD	0.06	0.06	0.08	0.99	0.10	1.00	0.10	1.56	0.09	0.66
	RMSE	0.06	0.06	0.08	1.00	0.10	1.00	0.10	1.57	0.09	0.66
$\boldsymbol{\beta}_{26}$	Bias	-1e-3	2e-3	-2e-3	1.57	-1e-3	0.06	1e-3	0.21	-4e-4	0.04
	SD	0.06	0.06	0.13	1.69	0.10	1.00	0.10	1.44	0.08	0.82
	RMSE	0.06	0.06	0.13	2.30	0.10	1.00	0.10	1.45	0.08	0.82

Table 3: Bias, SD and RMSE of $\hat{\boldsymbol{\beta}}$ with Xmodel II, Ymodel I, $n = 300$ and $p = 500$. Note the results of F-SIR and F-PHD are not presented due to the very poor performance.

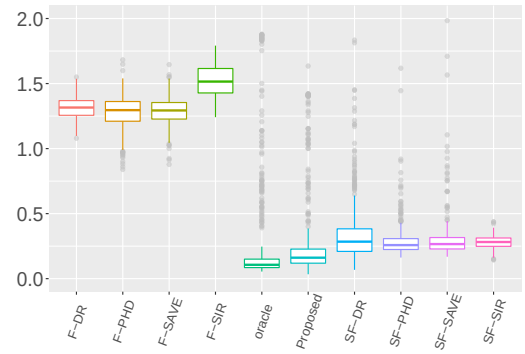
		Oracle	Proposed	SF-SIR	SF-PHD	SF-DR	F-DR	SF-SAVE	F-SAVE
$\boldsymbol{\beta}_{13}$	Bias	-6e-4	0.02	0.02	0.20	-0.09	1.23	0.09	-0.61
	SD	0.10	0.11	0.17	0.75	0.12	5.69	0.20	32.90
	RMSE	0.10	0.12	0.17	0.78	0.14	5.83	0.22	32.91
$\boldsymbol{\beta}_{14}$	Bias	1e-3	0.03	6e-3	0.12	0.03	0.51	0.02	-1.04
	SD	0.10	0.10	0.11	0.92	0.12	8.89	0.23	37.19
	RMSE	0.10	0.11	0.11	0.93	0.12	8.90	0.23	37.20
$\boldsymbol{\beta}_{15}$	Bias	1e-3	-4e-2	8e-3	-0.06	-0.15	-0.50	-0.06	-0.61
	SD	0.10	0.11	0.23	0.49	0.11	0.84	0.15	3.82
	RMSE	0.10	0.12	0.23	0.50	0.18	0.98	0.16	3.87
$\boldsymbol{\beta}_{16}$	Bias	-4e-3	-0.01	8e-3	0.04	-0.04	-0.48	-0.02	-2.10
	SD	0.09	0.11	0.14	0.88	0.12	0.89	0.16	20.91
	RMSE	0.09	0.11	0.14	0.88	0.13	1.01	0.16	21.02
$\boldsymbol{\beta}_{23}$	Bias	4e-3	0.01	-4e-4	-0.18	-0.07	-1.62	-0.06	-1.63
	SD	0.10	0.10	0.12	0.97	0.13	10.41	0.19	25.79
	RMSE	0.10	0.10	0.12	0.98	0.14	10.54	0.20	25.85
$\boldsymbol{\beta}_{24}$	Bias	-1e-3	0.03	0.01	0.08	0.04	4.74	0.03	1.32
	SD	0.10	0.10	0.57	0.81	0.12	76.27	0.19	17.22
	RMSE	0.10	0.10	0.57	0.82	0.13	76.46	0.20	17.27
$\boldsymbol{\beta}_{25}$	Bias	-2e-3	0.01	-8e-3	-0.21	-0.06	0.49	-0.07	0.34
	SD	0.09	0.11	0.46	0.91	0.12	2.11	0.15	3.59
	RMSE	0.09	0.11	0.46	0.94	0.13	2.17	0.16	3.60
$\boldsymbol{\beta}_{26}$	Bias	-8e-3	0.01	0.02	0.13	-0.11	-0.65	0.04	-0.46
	SD	0.08	0.11	0.76	1.08	0.14	8.70	0.17	7.58
	RMSE	0.08	0.11	0.76	1.09	0.18	8.72	0.18	7.60

Table 4: Bias, SD and RMSE of $\hat{\boldsymbol{\beta}}$ with Xmodel II, Ymodel I, $n = 300$ and $p = 1000$. Note the results of F-SIR and F-PHD are not presented due to the very poor performance.

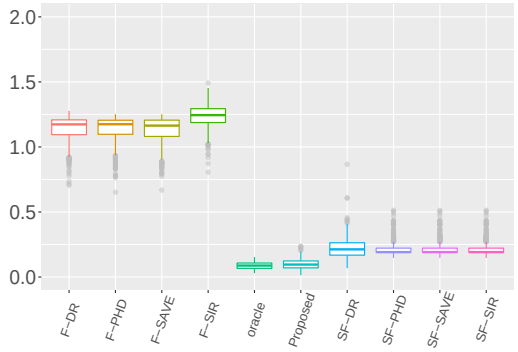
		Oracle	Proposed	SF-SIR	SF-PHD	SF-DR	F-DR	SF-SAVE	F-SAVE
β_{13}	Bias	-1e-3	0.02	0.02	0.21	-0.08	3.86	0.09	-7.90
	SD	0.09	0.11	0.16	0.71	0.11	95.12	0.17	153.55
	RMSE	0.09	0.11	0.16	0.74	0.14	95.20	0.19	153.75
β_{14}	Bias	2e-4	0.02	6e-3	0.11	0.03	-10.50	0.02	-7.16
	SD	0.10	0.10	0.11	0.63	0.11	354.04	0.20	495.61
	RMSE	0.10	0.10	0.11	0.64	0.12	354.20	0.20	495.66
β_{15}	Bias	5e-4	-0.04	6e-4	-0.05	-0.15	0.03	-0.06	-0.10
	SD	0.10	0.11	0.10	0.40	0.12	19.01	0.15	79.75
	RMSE	0.10	0.11	0.10	0.40	0.18	19.01	0.16	79.75
β_{16}	Bias	-5e-3	-6e-3	0.01	0.06	-0.04	-3.31	-0.02	-11.41
	SD	0.09	0.11	0.14	0.79	0.12	91.22	0.14	323.47
	RMSE	0.09	0.11	0.14	0.79	0.12	91.28	0.14	323.67
β_{23}	Bias	4e-3	0.02	2e-3	-0.17	-0.06	8.38	-0.06	3.30
	SD	0.10	0.10	0.11	0.73	0.12	337.00	0.16	150.71
	RMSE	0.10	0.10	0.11	0.75	0.13	337.09	0.17	150.74
β_{24}	Bias	-2e-3	0.03	9e-3	0.08	0.04	-35.27	0.01	-14.68
	SD	0.09	0.10	0.12	0.76	0.11	1256.46	0.20	301.05
	RMSE	0.09	0.10	0.12	0.76	0.12	1257.00	0.20	301.41
β_{25}	Bias	-9e-4	0.02	-0.02	-0.21	-0.06	2.33	-0.07	2.83
	SD	0.08	0.10	0.11	0.69	0.11	66.93	0.14	49.19
	RMSE	0.08	0.10	0.11	0.72	0.12	66.97	0.16	49.27
β_{26}	Bias	-8e-3	0.01	0.01	0.12	-0.10	-10.33	0.04	-5.02
	SD	0.07	0.11	0.25	0.80	0.13	323.56	0.14	172.85
	RMSE	0.07	0.11	0.25	0.81	0.17	323.73	0.14	172.92



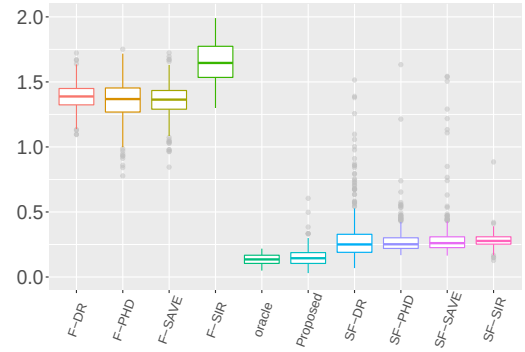
(a) $p = 500$ under Xmodel I and Ymodel II



(b) $p = 500$ under Xmodel II and Ymodel II



(c) $p = 1000$ under Xmodel I and Ymodel II



(d) $p = 1000$ under Xmodel II and Ymodel II

Figure 2: Comparison of Euclidean distance with $n = 300$. The points in grey are outliers that are greater than the third quartile plus/minus 1.5×interquartile range[34].

We observe the following. First, our estimator yields a smaller root mean squared error than the other competing approachs in all of the settings considered, regardless of the dimension of $\mathbf{X}$ and whether the linearity condition[22] on $\mathbf{H}$ is satisfied (Xmodel I) or not (Xmodel II). Relative to the Oracle estimator, the average empirical efficiency of the proposed method, FADR-SIR, FADR-PHD, FADR-DR, FADR-SAVE, SFADR-SIR, SFADR-PHD, SFADR-DR, and SFADR-SAVE is, respectively, $90.98\%, 55.63\%, 1.37\%, 25.52\%, 1.99\%,$ $56.68\%, 5.09\%, 47.48\%$ and 4.29%, indicating that the proposed estimator is nearly efficient, as stated in Theorem 2. The average empirical efficiency relative to the Oracle method is defined as $\frac{1}{(q-d)d}\sum_{j\geqslant d+1,k\leqslant d}\frac{SD_{Oracle}(\hat{\beta}_{jk})}{SD(\hat{\beta}_{jk})}$ ranging from 0 to 1, with a larger number indicating higher estimation efficiency. Second, the variation of $\widehat{\boldsymbol{\beta}}$ decreases with increasing p for all of the methods (except for FADR in Xmodel II), which is not surprising as a higher dimension of $\mathbf{x}_i$ may mean more information on the latent factors of $\mathbf{h}_i$. On the other hand, the variation for FADR increases when p increases in Xmodel II, because Xmodel II does not satisfy the linearity condition on the latent factors required by FADR. Figure 2 shows the proposed method outperforms SFADR and FADR and is close to the Oracle method in the Euclidean distance. We also note the instability of the proposed method, oracle method and SFADR methods in Figure 2. When $p = 500$, the proportions of the outliers for the Oracle, proposed and SFADR methods are approximately 5%; when $p = 1000$, the proportion for the proposed method drops to 0.6%, while the proportion remains approximately 5% for the SFADR method (SF-SIR, SF-PHD and SF-SAVE), as shown in Table S1 in Supplementary Materials. Figure 2 implies that the instability is induced by the estimation of factors, highlignting the importance of our supervised learning of factors.

Intuitively, as the objective function of the proposed method does not involve the distributions of errors in models (1) and (2), the proposed method should be robust to the errors' distribution. To verify this, we generate u_{ij} in Xmodel I or ϵ_i in Ymodel I from an exponential distribution $Exp(1)-1$ with mean zero, and consider three cases with $(n, p) = (300, 100)$:

$u_{ij} \sim Exp(1) - 1$ while $\epsilon_i \sim N(0,1)$; $u_{ij} \sim N(0,1)$ while $\epsilon_i \sim Exp(1) - 1$; and both of them are non-normal. We compare the proposed method with SF-SIR and F-SAVE, and find it outperformed the latter two as shown in Simulation 1; Figure 3 also suggests that the proposed method was robust to non-normal errors and outperformed the two other methods in estimation accuracy.

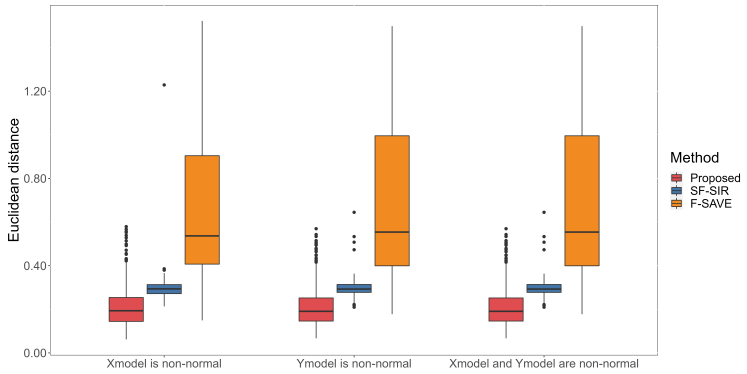


Figure 3: Comparison of estimation accuracy for the proposed method and two other competitors.

We examine the performance of using the criterion (5) described in Section 3.2 for selecting (q, d). Specifically, we calculate the frequency of (q, d) pairs selected by the criterion based on 200 repetitions in (Xmodel I, Ymodel II) and (Xmodel II, Ymodel II), respectively. We report the results in Table 5 with $(n, p) = (300, 50)$. It appears that (5) works well by identifying the true $(q, d) = (6, 2)$.

Table 5: Frequency of (q, d) selected over 200 repetitions under Ymodel II with $(n, p) = (300, 50)$.

		q=2	q=3	q=4	q=5	q=6	q=7
Xmodel I	d=1	0	0	0	0	0	0
	d=2	–	0	0	0	198	2
	d=3	–	–	0	0	0	0
Xmodel II	d=1	0	0	0	0	0	0
	d=2	–	0	0	0	189	11
	d=3	–	–	0	0	0	0

5.2 Simulation 2: data-driven simulation

Here, we conduct two data-driven simulation studies to investigate the performance of ranking SNPs and testing the significance of eQTLs based on the quantity $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_p)^T \equiv \mathbf{B}(\mathbf{B}^{\mathrm{T}}\mathbf{B})^{-1}\boldsymbol{\beta} \in R^{p\times d}$, which describes the relationship between y_i and $\mathbf{x}_i$. Particularly, the norm of $\boldsymbol{\alpha}_j$ is used to rank SNPs, and the testing $H_0 : \boldsymbol{\alpha}_j = 0$ to identify whether the variable j is significantly associated with the outcome y_i. We set $(n, p) = (300, 100)$ to mimic the real data, i.e., $p < n$.

5.2.1 Ranking SNPs

In this simulation, we consider the same data generation process as Simulation 1. Following Yang et al.[35], we use the top T consistent rate (CR), defined as $\frac{|\mathcal{S}_1^T \cap \mathcal{S}_0^T|}{T}$, to measure the performance of ranking SNPs, where $\mathcal{S}_1^T$ and $\mathcal{S}_0^T$ are the top T SNPs from a method and the truth, respectively. A larger CR implies better performance. We take $T = 10, 20, \cdots, 60$.

Figure 4 summarizes CR of the proposed method, SF-SIR, F-SAVE as well as univariate regression (UR). We obtain the ranks for UR based on the p-values: the top SNP has the least p-value. Figure 4 shows that the proposed method achieves the highest consistent rate among all cases considered, and the UR has the lowest consistent rate, indicating the importance of structural learning for high-dimensional correlated covariates.

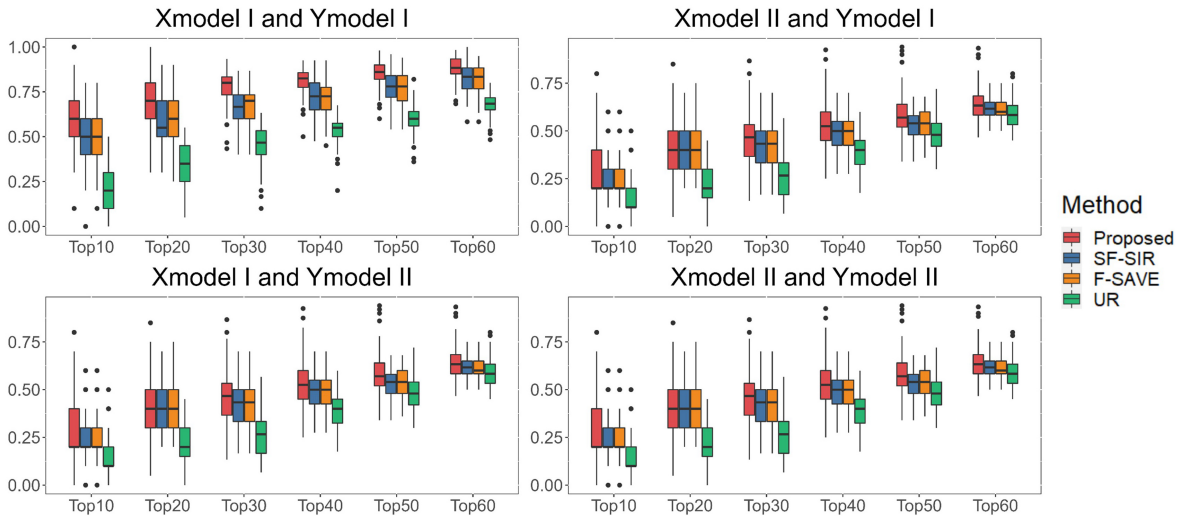


Figure 4: Comparison of the proposed method and other three methods in ranking SNPs.

5.2.2 Testing eQTLs

We generate $\mathbf{Z} = (\mathbf{z}_1^{\mathrm{T}}, \ldots, \mathbf{z}_n^{\mathrm{T}})^{\mathrm{T}}$ and $\mathbf{E}$ similarly with those in Simulation 1 except that $\mathbf{z}_i$ is generated from s-dimensional multivariate normal distribution and $\mathbf{E}$ is an $n \times s$ orthogonal matrix with $s = 10$. We then define $\mathbf{B}_1 = \mathbf{Z}^T\mathbf{E}$ and $\mathbf{B}_0 = [\mathbf{B}_1 : O_{n\times(p-s)}]$, where $O_{n\times(p-s)}$ is an $n \times (p-s)$-dimensional matrix with entries 0. We consider factor model Xmodel I with $\sigma_{ij} = 0.1^{|i-j|}$, and set two kinds of models for Y as

Ymodel III: $y_i = \sin(\mathbf{h}_{i0}^{\mathrm{T}}\boldsymbol{\beta}) + 2\cos(\mathbf{h}_{i0}^{\mathrm{T}}\boldsymbol{\beta}) + 2\epsilon_i,$

Ymodel IV: $y_i = \mathbf{h}_{i0}^{\mathrm{T}}\boldsymbol{\beta} + 2\epsilon_i,$

where $\epsilon_i \sim N(0,1)$ and $\boldsymbol{\beta} = (1, -0.5, 1, -1, 0, -1)^T$. Under this setting, $\boldsymbol{\alpha}_j \neq 0$ when $j \leqslant 10$ and $\boldsymbol{\alpha}_j = 0$ if $j > 10$. To test the hypothesis $H_0 : \boldsymbol{\alpha}_j = 0$, we first estimate the standard error of $\hat{\boldsymbol{\alpha}}$ using resampling method[36]. Concretely, we generate $V_i, i = 1, \ldots, n$ independently and identically distributed by $N(1,1)$, and obtain the estimator $\widetilde{\Omega} = (\widetilde{\mathbf{B}}, \widetilde{\mathbf{H}}, \widetilde{\boldsymbol{\beta}})$ by maximizing $\widetilde{l}(\Omega; \mathbf{y}, \mathbf{X}) = \sum_{i=1}^{n} V_i \log\{\hat{f}(y_i|\boldsymbol{\beta}\mathbf{h}_i)\} - \omega \sum_{i=1}^{n}\sum_{j=1}^{p} V_i(x_{ij} - \mathbf{h}^T\mathbf{b}_j)^2$ and then obtain $\widetilde{\boldsymbol{\alpha}} = \widetilde{\mathbf{B}}(\widetilde{\mathbf{B}}^T\widetilde{\mathbf{B}})^{-1}\widetilde{\boldsymbol{\beta}}^T$. Repeating the sampling 100 times, we then obtain the estimated

Table 6: Performance of testing the hypothesis $H_0 : \boldsymbol{\alpha}_j = 0$ for $j = 1, \ldots, d$ under level 0.1.

	TPR_j											
Model \ j	1	2	3	4	5	6	7	8	9	10	FPR	AUC
Ymodel III	.950	.714	.890	.608	.611	.765	.639	.625	.923	.943	.137	.858
Ymodel IV	.920	.584	.873	.666	.575	.807	.546	.552	.915	.979	.126	.849

TPR_j represents the rate that the nonzero $\boldsymbol{\alpha}_j$ is correctly identified; FPR represents the rate that all zero $\boldsymbol{\alpha}_j$ is incorrectly selected.

variance-covariance matrix for parameter $\boldsymbol{\alpha}$. With $s = 10$, the estimated standard error (ESE) of $\alpha_j, j = 1, \ldots, 10$ and the sampling standard deviation (SSE) for both Y models are summarized in Table S2 of Supplementary Materials based on 1000 replications, suggesting that the ESE's agree well with the corresponding SSE's. This implies that the performance of the estimated standard error based on resampling method in Jin et al.[36] is quite satisfactory. Table 6 summarizes the performance of testing results under level 0.1 using TPR, FPR and AUC criteria based on 1000 replications, where TPR_j represents the rate that the nonzero $\boldsymbol{\alpha}_j$ is correctly identified as non-zero, and FPR represents the rate that all zero $\boldsymbol{\alpha}_j, j = s+1, \ldots, p$ are incorrectly selected. Table 6 shows that the FPR is close to 0.1 and $TPR_j, j = 1, \cdots, 10$ are much larger than FPR under Ymodel III and Ymodel IV. Besides, the AUC's for both Y models are around 0.85. The results in Table 6 indicate that the associations between y_i and $\mathbf{x}_i$ can be correctly tested with high probability.

6 GTEx Data Analysis

6.1 Background and data

We apply our method to analyze the aforementioned data with $n = 278$ samples of lung tissue from the GTEx project. The response variable of interest is the expression level of gene ENSG00000225-880.4 measured by RNA-seq techniques. Located on Chromosome 1, this gene belongs to the category of lincRNA (long intergenic non-protein coding RNA) and is related to lung cancer. To improve the identification of eQTLs, we used the RNAK

normalization method, which is commonly-used in eQTL analysis[37,38], to transform the expression into a standard Gaussian. Although all the subjects were genome-wide genotyped, for more power we use the target locus approach[39,40] by focusing on the loci within 20kb in the flanking regions of the target gene. We end up with a total of 117 loci. Following the eQTL analysis in Li et al.[38], we include in our model gender, platform, three principal components of expressions of genome-wide genes, and 35 principal components of genome-wide SNPs, a total of 40 variables, to adjust for population characteristics and batch effects; see Appendix D2 in Supplementary Materials for the GTEx data pre-processing. The following delineates the application of the proposed method to identify eQTLs among these 117 candidate loci, and the comparison with the results obtained from standard GTEx analyses and from the SFADR methods (SFADR-SIR, SFADR-PHD, SFADR-DR, SFADR-SAVE) in Jiang et al.[3].

6.2 Implementations of the methods under comparison

Standard GTEx analyses are done via the univariate association regression, which is to regress the target gene expressions against genetic variants one at a time while controlling for those 40 adjusters. In contrast, our proposed method is designed to accommodate all of the SNPs in the model, while accounting for their correlations. As such, it empowers detection of eQTLs and elucidate SNPs' roles in regulating gene expressions. To implement the proposed method, we first choose the structural dimensions (q, d) using BIC as outlined in (5). As shown in Table 7, we determine an optimal structure with $(q, d) = (4, 1)$ that has the largest BIC value.

Table 7: BIC values under various structural dimensions

	q=2	q=3	q=4	q=5	q=6	q=7
d=1	-727.48	-674.29	-644.07	-720.43	-688.75	-720.02
d=2	–	-713.32	-735.82	-694.06	-685.95	-700.05

Table 8: Estimates of $\boldsymbol{\beta}$

	$\boldsymbol{\beta}_{12}$	$\boldsymbol{\beta}_{13}$	$\boldsymbol{\beta}_{14}$
Est	-0.3490	-0.6444	-0.4037
ESE	0.1785	0.1292	0.1648
p-value	0.0505	0.0000	0.0143

Under $(q, d) = (4, 1)$, we subsequently estimate $\boldsymbol{\beta}, \mathbf{B}$ and $\mathbf{H}$ by using the proposed method, and report the point estimates (Est), estimated standard errors (SE) and p-values for $\boldsymbol{\beta}$ in Table 8, where the ESE for $\widehat{\boldsymbol{\beta}}$ is calculated using resampling method based on 100 repetitions, as described in subsection 5.2.2. Compared with SFADR, the Pearson's correlation coefficient $Cor(\widehat{\mathbf{H}}\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{H}}_{SFADR}\widehat{\boldsymbol{\beta}}_{SFADR})$ is 0.0896, implying a large difference between our method and SFADR. To further assess the effects of individual SNPs on the gene expression, we estimate $\boldsymbol{\alpha}$ with $\widehat{\boldsymbol{\alpha}} = \widehat{\mathbf{B}}(\widehat{\mathbf{B}}^{\mathrm{T}}\widehat{\mathbf{B}})^{-1}\widehat{\boldsymbol{\beta}}^{\mathrm{T}} \in R^p$, of which the last 117 elements correspond to the effects of SNPs on the gene expression level in the sufficient direction. We then test null hypothesis $H_0 : \boldsymbol{\alpha}_j = 0, j = 41, \ldots, 157$ to identify the SNPs that are significantly associated with the expression levles of ENSG00000225880.4.

6.3 Comparisons of model predictiveness

We first use two-fold cross validation over 100 random splits to evaluate and compare the predictiveness of the proposed model, the SFADR models in Jiang et al.[3] and the FADR methods in Fan et al.[19]. The cross-validated prediction errors averaging over 100 random splits are 1.0086, 1.0257, 1.0511 and 1.0464 for the SFADR-SIR, SFADR-PHD, SFADR-DR and SFADR-SAVE methods, respectively, and are 1.2420, 1.2916, 1.3215 and 1.3106 for the FADR-SIR, FADR-PHD, FADR-DR and FADR-SAVE methods, respectively. In contrast, the average prediction error of our method is 0.9714, the smallest among all of the methods. As expected, FADR has the poorest model predictiveness due to its restrictive assumptions (i.e., linearity conditions on factors), whereas, by relaxing this assumption and using more flexible semi-parametric approaches, both SFADR and the proposed method improve model

fitness. The supervised learning in the proposed method further enhances model fitness compared to the SFADR approaches. Since the univariate regression approach in GTEx analysis uses separate pair-wise regressions and for fairness, we have opted not to compare its model adequacy with the other approaches.

6.4 Comparisons of the identified SNPs

Various methods have identified different numbers of eQTLs after the Bonferroni correction; the univariate regression approach (UR), the proposed supervised factoring approach, and the unsupervised SFADR methods, respectively, detect 76, 54, and 27 SNPs that are significantly associated with the expression levels of ENSG00000225880.4, and the SNPs identified by SFADR methods were all contained in the SNPs identified by UR; see Figure 1(a) for a Venn diagram of these identified eQTLs. Among the SNPs identified by the proposed method or by the univariate analyses, 36 SNPs are detected by both, 40 are uniquely identified by the univariate analyses, and 18 are uniquely identified by the proposed method. Moreover, the unsupervised factoring SFADR methods identify much fewer SNPs (only 27), 85% of which are also identified by the proposed method.

6.5 Validations of the identified eQTLs

As genetic variants nearing the target gene are likely to be correlated, some eQTLs might be falsely identified due to their high correlations with the true functional variants. To validate the identified eQTLs from the proposed method, and compare them with those identified by UR, we turn to the newly developed FUN-LDA scores[24,25]. FUN-LDA scores integrate epigenetic annotations from several large scale epigenomics projects such as ENCODE and Roadmap Epigenomics to predict the likelihood of an individual SNP to be a true functional variant in specific cell types and tissues; eQTLs with high FUN-LDA scores have a

high chance to be actual functional variants, and those with near-zero FUN-LDA scores may be false positives; see Figure 1(b) for the FUN-LDA scores of all the 117 candidate SNPs. As expected, near ENSG00000225880.4, identified by both the proposed approach and univariate regression is a cluster of SNPs with high FUN-LDA scores. Outside the range of ENSG00000225880.4, there are 3 SNPs with FUN-LDA scores larger than 0.3. The proposed method is able to identify all these three, while univariate regression and SFDAR can only locate two and one of them, respectively (Figure 1(b) and Figure S1 in Supplementary Materials). Out of 40 SNPs identified only by UR, 8.8% of them have high functional scores (i.e. FUN-LDA$>$ 0.1) in lung tissues; among the 18 SNPs uniquely identified by the proposed method, 16.7% of them have functional scores larger than 0.1. Taken altogether, we conclude that our joint analytical model may be more suitable for identifying functional variants than the traditional univariate analysis.

To demonstrate the proposed method's capability of identifying the eQTLs in the case of $n < p$, we only use the half of samples, the first 139 out of 278, to detect the eQTLs by comparing the proposed method with UR and SFADR. The proposed method, UR and SFADR, respectively, identified 42, 47 and 11 SNPs that were significantly associated with the expression levels of target gene; see Figure S2(a) in the Supplementary Materials for a Venn diagram of these identified eQTLs. Furthermore, we count the number of the identified SNPs whose FUN-LDA scores are greater than 0.3 or 0.1. Among the 117 SNPs, there are 5 and 12 SNPs with FUN-LDA scores greater than 0.3 and 0.1, respectively. The proposed method, UR and SFADR, respectively, identified 3, 1 and 1 SNPs with FUN-LDA scores greater than 0.3; and 6, 4 and 1 SNPs with FUN-LDA scores greater than 0.1. See Figure S2(b)&(c) in the Supplementary Materials for more details. When comparing with the proposed method using the full data, we find that out of the three SNPs whose FUN-LDA scores are greater than 0.3, there are two overlapped SNPs (rs3131967 and rs2905042), and out of the six SNPs whose scores are greater than 0.1, there are five overlapped SNPs. This

fact indicates the robustness of the proposed method for GTEx data analysis.

6.6 Cluster eQTLs using the estimated factor loadings

Presumably, our proposed method can gain power for eQTL detection by utilizing the similar factor loadings of the SNPs. To see this, we cluster the identified eQTLs by our proposed method based on their factor loadings, determine the number of clusters by minimizing the DB index[41] as shown in Figure S2(a) in Supplementary Materials, and end up with 4 well separated clusters as shown in Figure S2(b) in Supplementary Materials. Figure 5(a) further plots the locations of the 25 SNPs in cluster 2, along with their functional scores in lung tissues. In particular, the top 2 SNPs in the cluster are SNPs rs2905042 and rs2286139 with functional scores greater than 0.2. Figure 5(b) plots the distributions of the functional scores of the 25 SNPs in cluster 2 across 36 tissue types, where the darkest bars represent the proportion of SNPs with functional scores larger than 0.1 (likely functional variants), and the lightest bars represent the proportion of those with scores smaller than 0.001. The two top tissues with the largest proportions of true functional SNPs in this cluster are skeletal muscles, suggesting functional SNPs in cluster 2 are more likely to be expressed in muscle tissues. The two top SNPs rs2905042 and rs2286139 have moderate functional scores in lung tissues, and high scores across several muscle type tissues, including skeletal, Psoas and left ventricle muscles. However, with a weak signal, rs2905042 is not identified by UR, but by borrowing information from the shared factor loadings with rs2286139, the proposed supervised factoring method is able to identify both of them as eQTLs. On the other hand, the SFADR does not detect either rs2905042 or rs2286139. We also present a similar finding of the 10 SNPs in cluster 4; see Figure S3 in Supplementary Materials.

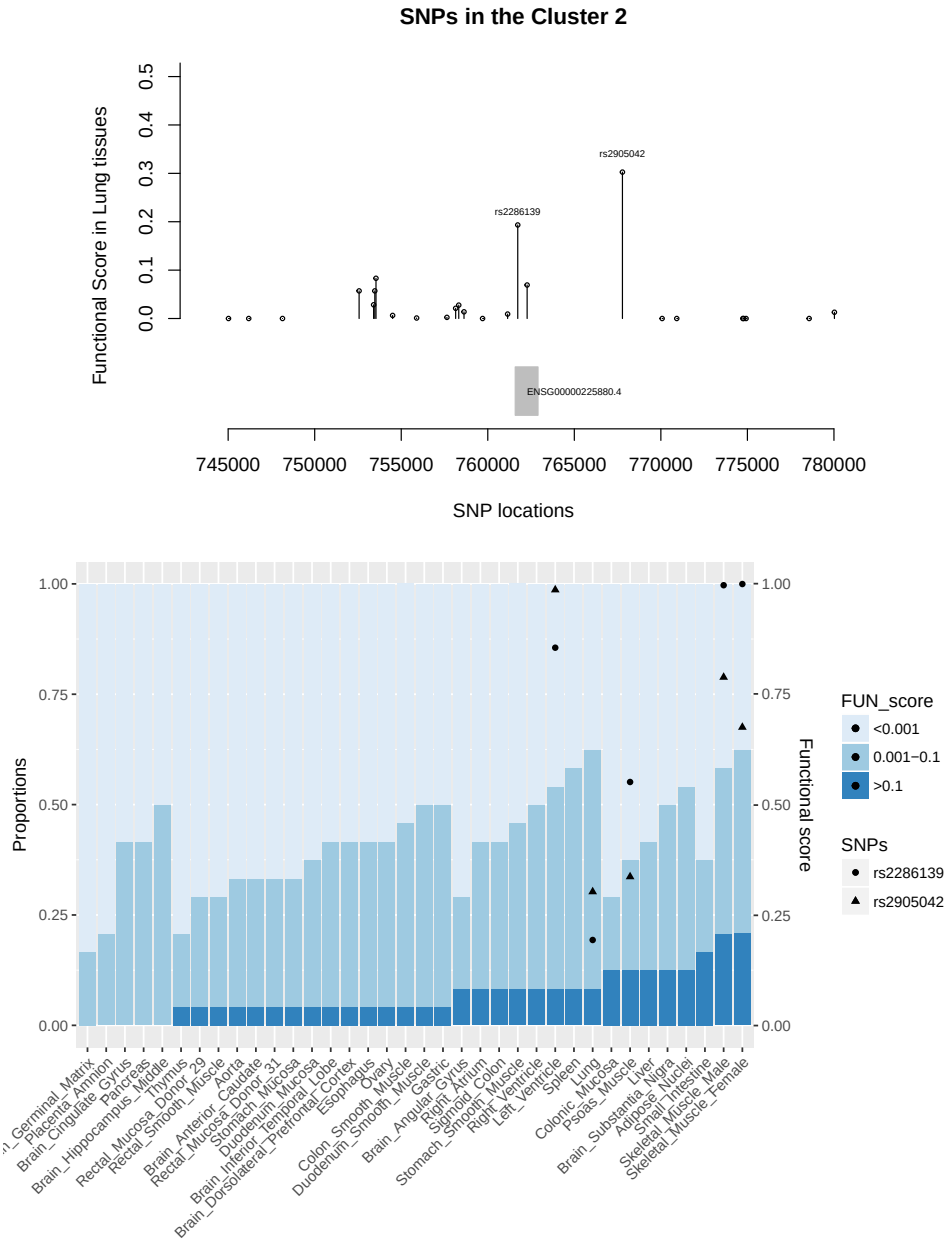


Figure 5: *Top*: Locations of the SNPs in cluster 2 and their functional scores in lung tissues. *Bottom*: Distribution of functional scores of the SNPs in cluster 2 across multiple tissues. The darkest blue bars represent the proportions of SNPs with functional scores larger than 0.1, the lightest bars represent the proportions with scores smaller than 0.001, and the median blue bars represent the proportions of those with scores between 0.001 and 0.1. The solid circles are the cross-tissue functional scores of rs2286139, and the solid triangles are those of rs2905042.

7 Discussion

We propose a supervised structural dimensional reduction method for semiparametric regression models with high-dimensional covariates, which seamlessly integrates factor analysis and sufficient dimension reduction under a penalized likelihood framework. There are several merits. First, by making full use of the information about correlations among covariates and relationships between responses and covariates, the method can handle high dimensional correlated covariates while embracing the blessing of high dimensionality. Second, the method is flexible enough to handle unspecified distributions of responses and forecasting functions, while relaxing the linearity condition on the latent factors in Fan et al.[19]. Moreover, to overcome the computational challenges, we have proposed an efficient iterative algorithm which benefits from closed-form solutions at each iteration. Last but not least, our method yields important findings from a GTEx study by identifying new SNPs that may regulate the expression of a lung cancer related gene.

Acknowledgements

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS.

Conflict Interest

The authors declare no potential conflict of interests.

Data Availability Statement

The GTEx data used for the analyses described in this manuscript were obtained from dbGaP accession number phs000424.v6.p1.

References

[1] Nica Alexandra C, Dermitzakis Emmanouil T. Expression quantitative trait loci: present and future *Phil. Trans. R. Soc. B.* 2013;368:20120362.

[2] Carithers Latarsha J, Moore Helen M. The genotype-tissue expression (GTEx) project 2015.

[3] Jiang F, Ma Y, Wei Y. Sufficient direction factor model and its application to gene expression quantitative trait loci discovery *Biometrika.* 2019;106:417-432.

[4] Porcu Eleonora, Rüeger Sina, Lepik Kaido, Santoni Federico A, Reymond Alexandre, Kutalik Zoltán. Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits *Nature communications.* 2019;10:1–12.

[5] Ardlie Kristin G., DeLuca The GTEx. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans *Science.* 2015;348:648–660.

[6] Frank Lldiko E, Friedman Jerome H. A Statistical View of Some Chemometrics Regression Tools *Technometrics.* 1993;35:109–135.

[7] Tibshirani R. J.. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society.* 1996;58:267-288.

[8] Fan Jianqing, Li Runze. Variable selection via nonconvave penalized likelihood and its oracle properties *Journal of the American Statistical Association.* 2001;96:1348–1360.

[9] Zou Hui, Hastie Trevor. Addendum: regularization and variable selection via the elastic net *Journal of the Royal Statistical Society: Series B (Statistical Methodology).* 2005;67:768–768.

[10] Zou Hui. The Adaptive Lasso and Its Oracle Properties *Journal of the American Statistical Association.* 2006;101:1418–1429.

[11] Candes Emmanuel, Tao Terence. The Dantzig Selector: Statistical Estimation When p Is Much Larger than n *The Annals of Statistics.* 2007;35:2313–2351.

[12] Fan Jianqing, Lv Jinchi. Sure independence screening for ultrahigh dimensional feature space *Journal of the Royal Statistical Society.* 2008;70:849–911.

[13] Fan Jianqing, Samworth Richard, Wu Yichao. Ultrahigh dimensional feature selection: beyond the linear model *Journal of Machine Learning Research Jmlr.* 2009;10:2013.

[14] Zhao S. D., Li Y.. Principled sure independence screening for Cox models with ultra-high-dimensional covariates *Journal of Multivariate Analysis.* 2012;105:397.

[15] Fan Jianqing, Feng Yang, Song Rui. Nonparametric Independence Screening in Sparse Ultra-High Dimensional Additive Models *Journal of the American Statistical Association.* 2011;106:544.

[16] Li Runze, Zhong Wei, Zhu Liping. Feature Screening via Distance Correlation Learning *Journal of the American Statistical Association.* 2012;107:1129–1139.

[17] Ma Y, Li Y, Lin H. Concordance measure-based feature screening and variable selection *Statistica Sinica.* 2017;27:1967–1985.

[18] Wu Cen, Zhang Qingzhao, Jiang Yu, Ma Shuangge. Robust network-based analysis of the associations between (epi) genetic measurements *Journal of multivariate analysis.* 2018;168:119–130.

[19] Fan Jianqing, Xue Lingzhou, Yao Jiawei. Sufficient forecasting using factor models *Journal of Econometrics.* 2017.

[20] Bengio Y.. Learning Deep Architectures for AI *Foundations and Trends in Machine Learning.* 2009;2:1–127.

[21] Bengio Y., Courville A., Vincent P.. Representation Learning: A Review and New Perspectives *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 2013;35:1798–1828.

[22] Li Ker Chau. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association.* 1991;86:316–327.

[23] Tu Yundong, Lee Tae-Hwy. Forecasting using supervised factor models *Journal of Management Science and Engineering.* 2019.

[24] Ionita-Laza Iuliana, McCallum Kenneth, Xu Bin, Buxbaum Joseph D. A spectral approach integrating functional genomic annotations for coding and noncoding variants *Nature genetics.* 2016;48:214.

[25] Backenroth Daniel, He Zihuai, Kiryluk Krzysztof, et al. FUN-LDA: A latent Dirichlet allocation model for predicting tissue-specific functional effects of noncoding variation *bioRxiv.* 2017:069229.

[26] Li Bing, Wang Shaoli. On Directional Regression for Dimension Reduction *Journal of the American Statistical Association.* 2007;102:997–1008.

[27] Hall Peter, Sheather Simon J, Jones MC, Marron James Stephen. On optimal data-based bandwidth selection in kernel density estimation *Biometrika.* 1991;78:263–269.

[28] Wang Hansheng, Li Runze, Tsai Chih-Ling. Tuning parameter selectors for the smoothly clipped absolute deviation method *Biometrika.* 2007;94:553–568.

[29] Lin Huazhen, Zhou Ling, Peng Heng, Zhou Xiao-Hua. Selection and combination of biomarkers using ROC method for disease classification and prediction *Canadian Journal of Statistics.* 2011;39:324–343.

[30] Ye Jianming. On measuring and correcting the effects of data mining and model selection *Journal of the American Statistical Association.* 1998;93:120–131.

[31] Müller H.. Smooth Optimum Kernel Estimators of Densities, Regression Curves and Modes *Annals of Statistics.* 1984;12:766-774.

[32] Li Ker Chau. On Principal Hessian Directions for Data Visualization and Dimension Reduction: Another Application of Stein's Lemma *Journal of the American Statistical Association.* 1992;87:1025–1039.

[33] Cook R. D, Weisberg S.. Comments on "Sliced inverse regression for dimension reduction" by K. C. Li *Journal of the American Statistical Association.* 1991;86:28-33.

[34] Tukey John W, others . *Exploratory data analysis*;2. Reading, MA 1977.

[35] Yang Can, Wang Lin, Zhang Shuqin, Zhao Hongyu. Accounting for non-genetic factors by low-rank representation and sparse regression for eQTL mapping *Bioinformatics.* 2013;29:1026–1034.

[36] Jin Zhezhen, Ying Zhiliang, Wei LJ. A simple resampling method by perturbing the minimand *Biometrika.* 2001;88:381–390.

[37] Yang Jiajun, Wang Dongyang, Yang Yanbo, et al. A systematic comparison of normalization methods for eQTL analysis *Briefings in Bioinformatics.* 2021;22:bbab193.

[38] Li Jiang, Xue Yawen, Amin Muhammad Talal, et al. ncRNA-eQTL: a database to systematically evaluate the effects of SNPs on non-coding RNA expression across cancer types *Nucleic acids research.* 2020;48:D956–D963.

[39] Sprowles AE, Stephens MR, Clipperton NW, May BP. Fishing for SNPs: a targeted locus approach for single nucleotide polymorphism discovery in rainbow trout *Transactions of the American Fisheries Society.* 2006;135:1698–1721.

[40] Ahrens Collin W, Rymer Paul D, Stow Adam, et al. The search for loci under selection: trends, biases and progress *Molecular ecology.* 2018;27:1342–1356.

[41] Davies David L, Bouldin Donald W. A cluster separation measure *IEEE transactions on pattern analysis and machine intelligence.* 1979:224–227.