http://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Articles in Advance, pp. 1–14
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Methods

Nonsparse Learning with Latent Variables

Zemin Zheng,[a] Jinchi Lv,[b] Wei Lin[c]

[a] International Institute of Finance, School of Management, University of Science and Technology of China, Hefei 230026, China; [b] Marshall School of Business, University of Southern California, Los Angeles, California 90089; [c] School of Mathematical Sciences and Center for Statistical Science, Peking University, Beijing 100871, China

Contact: zhengzm@ustc.edu.cn, https://orcid.org/0000-0002-0240-9411 (ZZ); jinchilv@marshall.usc.edu, https://orcid.org/0000-0002-5881-9591 (JL); weilin@math.pku.edu.cn, https://orcid.org/0000-0002-7598-6199 (WL)

Received: February 23, 2019
Revised: July 22, 2019
Accepted: January 2, 2020
Published Online in Articles in Advance: December 23, 2020

Subject Classification: statistics: data analysis, estimation, regression
Area of Review: High-Dimensional Learning

https://doi.org/10.1287/opre.2020.2005

Abstract. As a popular tool for producing meaningful and interpretable models, large-scale sparse learning works efficiently in many optimization applications when the underlying structures are indeed or close to sparse. However, naively applying the existing regularization methods can result in misleading outcomes because of model misspecification. In this paper, we consider nonsparse learning under the factors plus sparsity structure, which yields a joint modeling of sparse individual effects and common latent factors. A new methodology of nonsparse learning with latent variables (NSL) is proposed for joint estimation of the effects of two groups of features, one for individual effects and the other associated with the latent substructures, when the nonsparse effects are captured by the leading population principal component score vectors. We derive the convergence rates of both sample principal components and their score vectors that hold for a wide class of distributions. With the properly estimated latent variables, properties including model selection consistency and oracle inequalities under various prediction and estimation losses are established. Our new methodology and results are evidenced by simulation and real-data examples.

Funding: This work was supported by the Beijing Natural Science Foundation [Grant Z190001], the National Key R&D Program of China [Grant 2016YFC0207703], the National Science Foundation [Grant DMS-1953356], the National Natural Science Foundation of China [Grants 72071187, 11601501, 11671018, 11671374, 71532001, 71731010, and 71921001], the Natural Science Foundation of Anhui Province [Grant 1708085QA02], Fundamental Research Funds for the Central Universities [Grant WK2040160028], the Adobe Data Science Research Award, the Simons Foundation, and the Beijing Academy of Artificial Intelligence.
Supplemental Material: The e-companion is available at https://doi.org/10.1287/opre.2020.2005.

Keywords: high dimensionality • nonsparse coefficient vectors • factors plus sparsity structure • principal component analysis • spiked covariance • model selection

1. Introduction

Advances of information technologies have made high-dimensional data increasingly frequent not only in the domains of machine learning and biology but also in economics (Belloni et al. 2017, Uematsu and Tanaka 2019), marketing (Paulson et al. 2018), and numerous operations research and engineering optimization applications (Xu et al. 2016). In the high-dimensional regime, the number of available samples can be less than the dimensionality of the problem, so the optimization formulations will contain many more variables and constraints than in fact are needed to obtain a feasible solution. The key assumption that enables high-dimensional statistical inference is that the regression function lies in a low-dimensional manifold (Hastie et al. 2009, Fan and Lv 2010, Bühlmann and van de Geer 2011). Based on this sparsity assumption, a long list of regularization methods has been developed to generate meaningful and interpretable models, including those of Tibshirani (1996), Fan and Li (2001), Zou and Hastie (2005), Candès and Tao (2007), Belloni et al. (2011), Sun and Zhang (2012), and Chen et al. (2016), among many others. Algorithms and theoretical guarantees were also established for various regularization methods. See, for example, Zhao and Yu (2006), Radchenko and James (2008), Bickel et al. (2009), Tang and Leng (2010), Fan et al. (2012), Candès et al. (2018), and Belloni et al. (2018).

Although large-scale sparse learning works efficiently when the underlying structures are indeed or close to sparse, naively applying the existing regularization methods can result in misleading outcomes because of model misspecification (White 1982, Lv and Liu 2014, Hsu et al. 2019). In particular, it was imposed in most high-dimensional inference methods that the coefficient vectors are sparse, which has been questioned in real applications. For instance, Boyle et al. (2017) suggested the omnigenic model, in which the genes associated with complex traits tend to be spread across most of the genome. Similarly, it was conjectured earlier by Pritchard (2001) that instead of being sparse, the causal variants responsible for a trait can be distributed. Under such cases, making a correct statistical inference is an important yet challenging task. Though it is generally impossible to accurately estimate large numbers of nonzero parameters with relatively low sample size, nonsparse learning may be achieved by considering a natural extension of the sparse scenario—that is, the factors plus sparsity structure. Specifically, we assume the coefficient vector of predictors to be sparse after taking out the impacts of certain unobservable factors, which yields a joint modeling of sparse individual effects and common latent factors. A similar idea was exploited by Fan et al. (2013) using the low-rank-plus-sparse representation for large covariance estimation, where a sparse error covariance structure is imposed after extracting common but unobservable factors.

To characterize the impacts of latent variables, various methods have been proposed under different model settings. For instance, the latent and observed variables were assumed to be jointly Gaussian by Chandrasekaran et al. (2012) for graphical model selection. To control for confounding in genetic genomics studies, Lin et al. (2015) used genetic variants as instrumental variables. Pan et al. (2015) characterized latent variables by confirmatory factor analysis (CFA) in survival analysis and estimated them using the expectation–maximization algorithm. Despite the growing literature, relatively few studies deal with latent variables in high dimensions. In this paper, we focus on high-dimensional linear regression incorporating two groups of features besides the response variable—that is, predictors with individual effects and covariates associated with the latent substructures. The numbers of both predictors and potential latent variables can be large, where the latent variables are nonsparse linear combinations of the covariates. To the best of our knowledge, this is a new contribution to the case of high-dimensional latent variables. Our analysis also allows for a special case that the two groups of features are identical, meaning that the latent variables are associated with the original predictors.

We would like to provide a possible methodology of nonsparse learning when the nonsparse effects of the covariates can be captured by their leading population principal component score vectors, which are unobservable because of the unknown population covariance matrix. The main reasons are as follows. Practically, principal components evaluate orthogonal directions that reflect maximal variations in the data, thus often employed as surrogate variables to estimate the unobservable factors in many contemporary applications such as genome-wide expression studies (Leek and Storey 2007). In addition, the leading principal components are typically extracted to adjust for human genetic variations across population substructures (Menozzi et al. 1978, Cavalli-Sforza et al. 1993) or stratification (Price et al. 2006). From a theoretical point of view, principal components yield the maximum likelihood estimates of unobservable factors when the factors are uncorrelated with each other, even subject to certain measurement errors (Mardia et al. 1979). Moreover, the effects of the covariates are mainly worked through their leading population principal component score vectors when the remaining eigenvalues decay rapidly.

The major contributions of this paper are threefold. First, we propose nonsparse learning with latent variables based on the aforementioned factors plus sparsity structure to simultaneously recover the significant predictors and latent factors as well as their effects. Exploring population principal components as common latent variables will be helpful in attenuating collinearity and facilitating dimension reduction. Second, to estimate population principal components, we use the sample counterparts and provide the convergence rates of both sample principal components and their score vectors that hold for a wide class of distributions. The convergence property of sample score vectors is critical to the estimation accuracy of latent variables. This is, however, much less studied in the literature compared with the principal components, and our work is among the first attempts in the high-dimensional case. Third, we characterize the model identifiability condition and show that the proposed methodology is applicable to general families with properly estimated latent variables. In particular, under some regularity conditions, NSL via the threshold regression is proved to enjoy model selection consistency and oracle inequalities under various prediction and estimation losses.

The rest of this paper is organized as follows. Section 2 presents the new methodology of nonsparse learning with latent variables. We establish asymptotic properties of sample principal components and their score vectors in high dimensions, as well as theoretical properties of the proposed methodology via the threshold regression in Section 3. Simulated and real-data examples are provided in Section 4. Section 5 discusses extensions and possible future work. All the proofs of the main results and additional technical details are included in the e-companion to this paper.

2. Nonsparse Learning with Latent Variables

2.1. Model Setting

Denote by $\mathbf{y} = (y_1, \ldots, y_n)^T$ the n-dimensional response vector, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$ the $n \times p$ random design matrix with p predictors, and $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_q)$ the $n \times q$ random matrix with q features. Assume that the rows of $\mathbf{X}$ are independent with covariance matrix $\mathbf{\Sigma}_X$ and that the rows of $\mathbf{W}$ are independent with mean zero and covariance matrix $\mathbf{\Sigma}_W$. We consider the following high-dimensional linear regression model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \mathbf{W}\boldsymbol{\eta}_0 + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{\beta}_0 = (\beta_{0,1}, \ldots, \beta_{0,p})^T$ and $\boldsymbol{\eta}_0 = (\eta_{0,1}, \ldots, \eta_{0,q})^T$ are, respectively, the regression coefficient vectors of the predictors and the additional features, and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$ is an n-dimensional error vector independent of $\mathbf{X}$ and $\mathbf{W}$.

The Gaussianity of the random noises is imposed for simplicity, and our technical arguments still apply as long as the error tail probability bound decays exponentially. Different from most of the existing literature, the regression coefficients $\boldsymbol{\eta}_0$ for covariates $\mathbf{W}$ can be nonsparse, whereas the coefficient vector $\boldsymbol{\beta}_0$ for predictors is assumed to be sparse with many zero components after adjusting for the impacts of additional features. Therefore, Model (1) is a mixture of sparse and nonsparse effects. Both the dimensionality p and the number of features q are allowed to grow nonpolynomially fast with the sample size n.

As discussed in Section 1, to make the nonsparse learning possible, we impose the assumption that the impacts of covariates $\mathbf{W}$ are captured by their K leading population principal component score vectors $\mathbf{f}_i = \mathbf{W}\mathbf{u}_i$ for $1 \le i \le K$, where $\{\mathbf{u}_i\}_{i=1}^K$ are the top-K principal components of the covariance matrix Σ_W. That is, the coefficient vector $\boldsymbol{\eta}_0$ lies in the span of the top K population principal components, and thus, $\boldsymbol{\eta}_0 = \mathbf{U}_0\boldsymbol{\gamma}_0$ for some coefficient vector $\boldsymbol{\gamma}_0$ and $\mathbf{U}_0 = (\mathbf{u}_1, \ldots, \mathbf{u}_K)$. In fact, when the covariance matrix Σ_W adopts a spiked structure (to be discussed in Section 3.1), the part of $\boldsymbol{\eta}_0$ orthogonal to the span of the leading population principal components will play a relatively small role in prediction in view of $\mathbf{W}\boldsymbol{\eta}_0 = \widehat{\mathbf{V}}\widehat{\mathbf{D}}\widehat{\mathbf{U}}^T\boldsymbol{\eta}_0$, where $\widehat{\mathbf{V}}\widehat{\mathbf{D}}\widehat{\mathbf{U}}^T$ is the singular value decomposition of $\mathbf{W}$.

Denote by $\mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_K)$ the $n \times K$ matrix consisting of the K potential latent variables and by $\boldsymbol{\gamma}_0 = (\gamma_{0,1}, \ldots, \gamma_{0,K})^T$ their true regression coefficient vector. Then Model (1) can be rewritten as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \mathbf{F}\boldsymbol{\gamma}_0 + \varepsilon. \tag{2}$$

It is worth pointing out that the latent variables $\mathbf{F}$ are unobservable to us because of the unknown vectors $\mathbf{u}_i$, which makes our work distinct from most existing studies. For the identifiability of population principal components in high dimensions, K will be the number of significant eigenvalues in the spiked covariance structure of Σ_W, and we allow it to diverge with the sample size.

Model (2) is applicable to two different situations. The first is that we aim at recovering the relationship between predictors $\mathbf{X}$ and response $\mathbf{y}$, whereas the features $\mathbf{W}$ are treated as confounding variables for making correct inferences on the effects of $\mathbf{X}$, such as the gene expression studies with sources of heterogeneity (Leek and Storey 2007). Then the latent variables $\mathbf{f}_i$ are not required to be associated with different eigenvalues as long as their joint impacts $\mathbf{F}\boldsymbol{\gamma}_0$ can be estimated. The other situation is that we are interested in exploring the effects of both $\mathbf{X}$ and $\mathbf{F}$ such that the latent variables $\mathbf{f}_i$ should be identifiable. This occurs in applications when the latent variables are also meaningful. For instance, the principal components of genes can be biologically interpretable as representing independent regulatory programs or processes (referred to as *eigengenes*) from their expression patterns (Alter et al. 2000, Bair et al. 2006).

In this paper, we mainly focus on the second situation because the latent variables in our motivating application can also be biologically important. Specifically, both nutrient intake and human gut microbiome composition are believed to be important in the analysis of body mass index (BMI), and they share strong associations (Chen and Li 2013). To alleviate the strong correlations and facilitate the analysis of possibly nonsparse effects, we take nutrient intake as a predictor and adjust for confounding variables by incorporating the principal components of gut microbiome composition because principal components of human gut microbiome were found to reveal different enterotypes that affect energy extraction from the diet (Arumugam et al. 2011). The results of this real-data analysis will be presented in Section 4.2.

In general applications, which variables should be chosen as $\mathbf{X}$ and which should be chosen as $\mathbf{W}$ depend on the domain knowledge and research interests. Overall, predictors $\mathbf{X}$ stand for features with individual effects, whereas observable covariates $\mathbf{W}$ are covariates that reflect the confounding substructures. Our analysis also allows for a special case that $\mathbf{W}$ is a part of $\mathbf{X}$, meaning that the latent variables are nonsparse linear combinations of the original predictors. The identifiability of Model (2) will be discussed in Section 3.3 after Condition 4.

2.2. Estimation Procedure by NSL

With unobservable latent factors $\mathbf{F}$, it is challenging to consistently estimate and recover the support of the regression coefficient vector $\boldsymbol{\beta}_0$ for observable predictors and the coefficients $\boldsymbol{\gamma}_0$ for latent variables. We partially overcome this difficulty by assuming that the factors appear in an unknown linear form of the covariates $\mathbf{W}$. Then $\mathbf{F}$ can be estimated by the sample principal component scores of matrix $\mathbf{W}$. Because the rows of $\mathbf{W}$ have mean zero, the sample covariance matrix $\mathbf{S} = n^{-1}\mathbf{W}^T\mathbf{W}$ is an unbiased estimate of $\boldsymbol{\Sigma}_W$ with top K principal components $\{\widehat{\mathbf{u}}_i\}_{i=1}^K$. So the estimated latent variables are $\widehat{\mathbf{F}} = (\widehat{\mathbf{f}}_1, \ldots, \widehat{\mathbf{f}}_K)$ with $\widehat{\mathbf{f}}_i = \mathbf{W}\widehat{\mathbf{u}}_i$ for $1 \le i \le K$. To ensure identifiability, both $\mathbf{f}_i$ and $\widehat{\mathbf{f}}_i$ are rescaled to have a common L_2-norm $n^{1/2}$, matching that of the constant predictor $\mathbf{1}$ for the intercept. For future prediction, we can transform the coefficient vector $\boldsymbol{\gamma}$ back by multiplying the scalars $n^{1/2}\|\mathbf{W}\widehat{\mathbf{u}}_i\|_2^{-1}$. The notation $\|\cdot\|_q$ denotes the L_q-norm of a given vector for $q \in [0, \infty]$.

To produce a joint estimate for the true coefficient vectors $\boldsymbol{\beta}_0$ and $\boldsymbol{\gamma}_0$, we suggest NSL, which minimizes

$$Q\left\{\left(\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T\right)^T\right\} = (2n)^{-1}\left\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \widehat{\mathbf{F}}\boldsymbol{\gamma}\right\|_2^2 + \left\|p_\lambda\left\{\left(\boldsymbol{\beta}_*^T, \boldsymbol{\gamma}^T\right)^T\right\}\right\|_1, \quad (3)$$

the penalized residual sum of squares with penalty function $p_\lambda(\cdot)$. Here $\boldsymbol{\beta}_* = (\beta_{*,1}, \ldots, \beta_{*,p})^T$ is the Hadamard (component-wise) product of two p-dimensional vectors $(n^{-1/2}\|\mathbf{x}_k\|_2)_{1\le k\le p}$ and $\boldsymbol{\beta}$. It corresponds to the design matrix with each column rescaled to have a common L_2-norm $n^{1/2}$. The penalty function $p_\lambda(t)$ is defined on $t \in [0, \infty)$, indexed by $\lambda \ge 0$, and assumed to be increasing in both λ and t with $p_\lambda(0) = 0$. We use a compact notation for

$$p_\lambda\left\{\left(\boldsymbol{\beta}_*^T, \boldsymbol{\gamma}^T\right)^T\right\} = \left\{p_\lambda\left(|\beta_{*,1}|\right), \ldots, p_\lambda\left(|\beta_{*,p}|\right), p_\lambda\left(|\gamma_1|\right), \ldots, p_\lambda\left(|\gamma_K|\right)\right\}^T.$$

The proposed methodology in (3) enables the possibility of simultaneously estimating $\boldsymbol{\beta}_0$ and $\boldsymbol{\gamma}_0$, identifying the significant observable predictors and latent factors altogether. However, it is still difficult to obtain accurate estimates because the confounding factors $\mathbf{F}$ are replaced by the estimate $\widehat{\mathbf{F}}$, and the correlations between the observable predictors and latent variables can aggravate the difficulty. To prevent the estimation errors being further magnified in prediction, we consider $\boldsymbol{\gamma}$ in an L_∞ ball $\mathbb{B}_\rho = \{\boldsymbol{\gamma} \in \mathbb{R}^K : \|\boldsymbol{\gamma}\|_\infty \le \rho\}$, where any component of $\boldsymbol{\gamma}$ is assumed to be no larger than ρ in magnitude. We allow ρ to diverge slowly such that it will not deteriorate the overall prediction accuracy.

2.3. Comparisons with Existing Methods

The proposed methodology can be regarded as a realization of the aforementioned low-rank plus sparse representation (Fan et al. 2013) in the high-dimensional linear regression setting, but there are significant differences lying behind them. First, the latent variables in our setup are not necessarily a part of the original predictors but can stem from any sources related to the underlying features. Second, unlike the typical assumption in factor analysis that the factors and the remaining part are uncorrelated, we allow latent variables to share correlations with the observable predictors. In the extreme case, the latent variables can be nonsparse linear combinations of the predictors. Third, latent variables are employed to recover the information beyond the sparse effects of predictors, and thus we do not modify or assume simplified correlations between the original predictors even after accounting for the latent substructures.

Another method proposed by Kneip and Sarda (2011) also incorporated principal components as extra predictors in penalized regression, but it applies to a single group of features. Even if the two groups of features $\mathbf{X}$ and $\mathbf{W}$ are identical, the method differs from ours in the following aspects. First of all, based on the framework of factor analysis, the observed predictors in Kneip and Sarda (2011) were mixtures of individual features and common factors, both of which were unobservable. In view of this, we aim at different scopes of applications. Moreover, Kneip and Sarda (2011) suggested sparse regression on the projected model, where individual features were recovered as residuals of projecting the observed predictors on the factors. In contrast, we keep the original predictors such that they will not be contaminated when the estimated latent variables are irrelevant. Last but not least, benefiting from factor analysis, the individual features in Kneip and Sarda (2011) were uncorrelated with each other and also shared no correlation with the factors. But we do not impose such assumptions, as explained earlier.

The proposed methodology is also closely related to principal component regression (PCR). PCR suggests regressing the response vector on a subset of principal components instead of all explanatory variables, and comprehensive properties have been established in the literature for its importance in reducing collinearity and enabling prediction in high dimensions. For instance, Cook (2007) explored situations where the response can be regressed on the leading principal components of predictors with little loss of information. Probabilistic explanation was provided by Artemiou and Li (2009) to support the phenomenon that the response is often highly correlated with the leading principal components. Our new methodology takes advantage of the strengths of principal components to extract the most relevant information from additional sources and adjust for confounding and nonsparse effects, whereas model interpretability is also retained by exploring the individual effects of observable predictors.

In addition to the aforementioned literature, there are two recent lines of work addressing nonsparse learning in high dimensions. Essential regression (Bing et al. 2019, 2020) is a new variant of factor regression models where both the response and covariates depend linearly on unobserved low-dimensional factors. Our model assumptions are quite different from those of Bing et al. (2019, 2020) in that we also allow the presence of covariates with sparse individual effects. Another line of work including that of Bradic et al. (2020) and Zhu and Bradic (2018) aims at hypothesis testing under nonsparse linear structures. Test statistics were constructed through restructured regression or marginal correlations, but the original regression coefficients were not estimated.

3. Theoretical Properties

We will first establish the convergence properties of sample principal components and their score vectors for a wide class of distributions under the spiked covariance structure. With the aid of these convergence properties, additional properties including model selection consistency and oracle inequalities will be proved for the proposed methodology via the threshold regression using hard thresholding.

3.1. Spiked Covariance Model

High-dimensional principal component analysis (PCA) particularly in the context of spiked covariance model, introduced by Johnstone (2001), has been studied by Paul (2007), Jung and Marron (2009), Shen et al. (2016), and Wang and Fan (2017), among many others. This model assumes that the first few eigenvalues of the population covariance matrix deviate from one, whereas the rest are equal to one. Although sample principal components are generally inconsistent without strong conditions when the number of covariates is comparable to or larger than the sample size (Johnstone and Lu 2009), with the aid of a spiked covariance structure, consistency of sample principal components was established in the literature under different high-dimensional settings. For instance, in the high-dimension, low-sample-size context, Jung and Marron (2009) proved the consistency of sample principal components for spiked eigenvalues. When both the dimensionality and sample size are diverging, phase transition of sample principal components was studied by Paul (2007) and Shen et al. (2016) for multivariate Gaussian observations. The asymptotic distributions of spiked principal components were established by Wang and Fan (2017) for sub-Gaussian distributions with a finite number of distinguishable spiked eigenvalues.

In this section, we adopt the generalized version of spiked covariance model studied by Jung and Marron (2009) for the covariance structure of covariate matrix $\mathbf{W}$, where the population covariance matrix $\boldsymbol{\Sigma}_W$ is assumed to contain K spiked eigenvalues that can be divided into m groups. The eigenvalues grow at the same rate within each group, whereas the orders of magnitude of the m groups are different from each other. To be specific, there are positive constants $\alpha_1 > \alpha_2 > \cdots > \alpha_m > 1$ such that the eigenvalues in the lth group grow at the rate of q^{α_l}, $1 \le l \le m$, where q is the dimensionality or number of covariates in $\mathbf{W}$. The constants α_l are larger than one because otherwise the sample eigenvectors can be strongly inconsistent (Jung and Marron 2009). Denote the group sizes by positive integers $k_1, \ldots, k_m$ satisfying $\sum_{l=1}^{m} k_l = K < n$. Set $k_{m+1} = q - K$, which is the number of nonspiked eigenvalues. Then the set of indices for the lth group of eigenvalues is

$$J_l = \left\{1 + \sum_{j=1}^{l-1} k_j, \ldots, k_l + \sum_{j=1}^{l-1} k_j\right\}, \quad l = 1, \ldots, m+1. \tag{4}$$

Although this eigenstructure looks almost the same as that in Jung and Marron (2009), the key difference lies in the magnitudes of the sample size n and the number of spiked eigenvalues K, both of which are allowed to diverge in our setup instead of being fixed. This makes the original convergence analysis of sample eigenvalues and eigenvectors invalid because the number of entries in the dual matrix $\mathbf{S}_D = n^{-1}\mathbf{W}\mathbf{W}^T$ is no longer finite. We will overcome this difficulty by conducting a delicate analysis on the deviation bound of the entries such that the corresponding matrices converge in Frobenius norm. Our theoretical results are applicable to a wide class of distributions, including sub-Gaussian distributions. For multivariate Gaussian or sub-Gaussian observations with a finite number of spiked eigenvalues, the phase transition of PCA consistency was studied by, for instance, Shen et al. (2016) and Wang and Fan (2017). Nevertheless, the convergence property of sample principal component score vectors was not provided in the aforementioned references and needs further investigation.

Assume that the eigendecomposition of the population covariance matrix $\boldsymbol{\Sigma}_W$ is given by $\boldsymbol{\Sigma}_W = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$, where $\boldsymbol{\Lambda}$ is a diagonal matrix of eigenvalues $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_q \ge 0$, and $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_q)$ is an orthogonal matrix consisting of the population principal components. Analogously, the eigendecomposition of $\mathbf{S} = \widehat{\mathbf{U}}\widehat{\boldsymbol{\Lambda}}\widehat{\mathbf{U}}^T$ provides the diagonal matrix $\widehat{\boldsymbol{\Lambda}}$ of sample eigenvalues $\widehat{\lambda}_1 \ge \widehat{\lambda}_2 \ge \cdots \ge \widehat{\lambda}_q \ge 0$ and the orthogonal matrix $\widehat{\mathbf{U}} = (\widehat{\mathbf{u}}_1, \ldots, \widehat{\mathbf{u}}_q)$ consisting of sample principal components. We always assume that the sample principal components take the correct directions such that the angles between sample and population principal components are no more than a right angle.

Our main focus is the high-dimensional setting where the number of covariates q is no less than the sample size n. Denote by

$$\mathbf{Z} = \boldsymbol{\Lambda}^{-1/2}\mathbf{U}^T\mathbf{W}^T \tag{5}$$

the sphered data matrix. It is clear that the columns of $\mathbf{Z}$ are independent and identically distributed (i.i.d.) with mean zero and covariance matrix $\mathbf{I}_q$. To build our theory, we will impose a tail probability bound on the entry of $\mathbf{Z}$ and make use of the n-dimensional dual matrix $\mathbf{S}_D = n^{-1}\mathbf{Z}^T\mathbf{\Lambda}\mathbf{Z}$, which shares the same nonzero eigenvalues with $\mathbf{S}$.

3.2. Threshold Regression Using Hard Thresholding

As discussed in Section 1, there is a large spectrum of regularization methods for sparse learning in high dimensions. It has been demonstrated by Fan and Lv (2013) that the popular L_1-regularization of least absolute shrinkage and selection operator (Lasso) and concave methods can be asymptotically equivalent in thresholded parameter space for polynomially growing dimensionality, meaning that they share the same convergence rates in the oracle inequalities. For exponentially growing dimensionality, concave methods can also be asymptotically equivalent and have faster convergence rates than Lasso. Therefore, we will show theoretical properties of the proposed methodology via a specific concave regularization method, the threshold regression using hard thresholding (Zheng et al. 2014). It uses either the hard-thresholding penalty $p_{H,\lambda}(t) = \frac{1}{2}[\lambda^2 - (\lambda - t)_+^2]$ or the L_0-penalty $p_{H_0,\lambda}(t) = 2^{-1}\lambda^2 1_{\{t\neq 0\}}$ in the penalized least squares (3), both of which enjoy the hard-thresholding property (Zheng et al. 2014, lemma 1) that facilitates sparse modeling and consistent estimation.

A key concept for characterizing model identifiability in Zheng et al. (2014) is the robust spark $rspark_c(\mathbf{X})$ of a given $n \times p$ design matrix $\mathbf{X}$ with bound c, defined as the smallest possible number τ such that there exists a submatrix consisting of τ columns from $n^{-1/2}\widetilde{\mathbf{X}}$ with a singular value less than the given positive constant c, where $\widetilde{\mathbf{X}}$ is obtained by rescaling the columns of $\mathbf{X}$ to have a common L_2-norm $n^{1/2}$. The bound on the magnitude of $rspark_c(\mathbf{X})$ was established by Fan and Lv (2013) for Gaussian design matrices and further studied by Lv (2013) for more general random design matrices. Under mild conditions, $M = \tilde{c}n/(\log p)$ with some positive constant $\tilde{c}$ will provide a lower bound on $rspark_c(\mathbf{X}, \mathbf{F})$ for the augmented design matrix (see Condition 4 in Section 3.3 for details). Following Fan and Lv (2013) and Zheng et al. (2014), we consider the regularized estimator on the union of coordinate subspaces $\mathbb{S}_{M/2} = \{(\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T)^T \in \mathbb{R}^{p+K} : \|(\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T)^T\|_0 < M/2\}$ to ensure model identifiability and reduce estimation instability. So the joint estimator $(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T$ is defined as the global minimizer of the penalized least squares (3) constrained on space $\mathbb{S}_{M/2}$.

3.3. Technical Conditions

Here we list a few technical conditions and discuss their relevance. Let $\Delta = \min_{1\le l\le m-1}(\alpha_l - \alpha_{l+1})$. Then q^{Δ} reflects the minimum gap between the magnitudes of spiked eigenvalues in two successive groups. The first two conditions are imposed for Theorem 1, whereas the rest are needed in Theorem 2, to be presented in Section 3.4.

Condition 1. *There exist positive constants c_i and C such that uniformly over $i \in J_l$, $1 \le l \le m$,*

$$\lambda_i/q^{\alpha_l} = c_i + O(q^{-\Delta}) \quad \text{with} \quad c_i \le C,$$

and $\lambda_j \le C$ for any $j \in J_{m+1}$.

Condition 2. (a) *There exists some positive $\alpha < \min\{\Delta, \alpha_m - 1\}$ such that uniformly over $1 \le i \le n$ and $1 \le j \le q$, the (j, i)th entry z_{ji} of the sphered data matrix $\mathbf{Z}$ defined in (5) satisfies*

$$P\left(z_{ji}^2 > K^{-1}q^{\alpha}\right) = o\left(q^{-1}n^{-1}\right).$$

(b) *For any $1 \le l \le m$, $\|n^{-1}\mathbf{Z}_l\mathbf{Z}_l^T - I_{k_l}\|_\infty = o_p(k_l^{-1})$, where $\mathbf{Z}_l$ is a submatrix of $\mathbf{Z}$ consisting of the rows with indices in J_l.*

Condition 3. *Uniformly over j, $1 \le j \le K$, the angle ω_{jj} between the jth estimated latent vector $\widehat{\mathbf{f}}_j$ and its population counterpart $\mathbf{f}_j$ satisfies $\cos(\omega_{jj}) \ge 1 - c_2^2 \log n/8K^2\rho^2 n$ with probability $1 - \theta_1$ that converges to one as $n \to \infty$.*

Condition 4. *The inequality $\|n^{-1/2}(\mathbf{X}, \mathbf{F})\boldsymbol{\delta}\|_2 \ge c\|\boldsymbol{\delta}\|_2$ holds for any $\boldsymbol{\delta}$ satisfying $\|\boldsymbol{\delta}\|_0 < M$ with probability $1 - \theta_2$ approaching one as $n \to \infty$.*

Condition 5. *There exists some positive constant L such that*

$$P\left(\cap_{j=1}^{p}\{L^{-1} \le \|\mathbf{x}_j\|_2/\sqrt{n} \le L\}\right) = 1 - \theta_3,$$

where θ_3 converges to zero as $n \to \infty$.

Condition 6. *Denote by $s = \|\boldsymbol{\beta}_0\|_0 + \|\boldsymbol{\gamma}_0\|_0$ the number of overall significant predictors and by $b_0 = \min_{j\in \mathrm{supp}(\boldsymbol{\beta}_0)}(|\beta_{0,j}|) \wedge \min_{j\in\mathrm{supp}(\boldsymbol{\gamma}_0)}(|\gamma_{0,j}|)$ the overall minimum signal strength. It holds that $s < M/2$ and*

$$b_0 > \left[\left(\sqrt{2}c_1^{-1}\right) \vee 1\right]c_1^{-1}c_2 L\sqrt{(2s+1)(\log p)/n}$$

for some positive constants c_1 defined in Proposition 1 in Section 3.4 and $c_2 > 2\sqrt{2}\sigma$.

Condition 1 requires that the orders of magnitude of spiked eigenvalues in each group be the same, whereas their limits can be different depending on the constants c_i. This is weaker than the conditions usually imposed in the literature, such as in Shen et al. (2016), where the spiked eigenvalues in each group share exactly the same limit. Nevertheless, we will prove the consistency of spiked sample eigenvalues under very mild conditions. To distinguish the eigenvalues in different groups, convergence to the corresponding limit is assumed to be at a rate of $O(q^{-\Delta})$. As the number of spiked eigenvalues diverges with q, we impose a constant upper bound C on c_i for simplicity, and our technical argument still applies when C diverges slowly with q. Without loss of generality, the upper bound C also controls the nonspiked eigenvalues.

As pointed out earlier, the columns of the sphered data matrix $\mathbf{Z}$ are i.i.d. with mean zero and covariance matrix $\mathbf{I}_p$. Then part (a) of Condition 2 holds as long as the entries in any column of $\mathbf{Z}$ satisfy the tail probability bound. Moreover, it is clear that this tail bound decays polynomially, so it holds for a wide class of distributions including sub-Gaussian distributions. With this tail bound, the larger sample eigenvalues would dominate the sum of all eigenvalues in the smaller groups regardless of the randomness. Furthermore, by definition, we know that the columns of $\mathbf{Z}_l$ are i.i.d. with mean zero and covariance matrix $\mathbf{I}_{k_l}$ such that $n^{-1}\mathbf{Z}_l\mathbf{Z}_l^T \to \mathbf{I}_{k_l}$ entrywise as $n \to \infty$. Hence, part (b) of Condition 2 is a very mild assumption to deal with the possibly diverging group sizes k_l.

Condition 3 imposes a convergence rate of $\log n/(K^2\rho^2 n)$ for the estimation accuracy of confounding factors, so the estimation errors in $\widehat{\mathbf{F}}$ will not deteriorate the overall estimation and prediction powers. This rate is easy to satisfy in view of the results in Theorem 1 in Section 3.4 because the sample principal component score vectors are shown to converge to the population counterparts in polynomial orders of q, which is typically larger than n in high-dimensional settings.

Condition 4 assumes the robust spark of matrix $(\mathbf{X}, \mathbf{F})$ with bound c to be at least $M = \tilde{c}n/(\log p)$ with significant probability. This is the key for characterizing the model identifiability in our conditional sparsity structure and also controls the correlations between the observable predictors $\mathbf{X}$ and latent factors $\mathbf{F}$. Consider a special case where $\mathbf{F}$ consists of nonsparse linear combinations of the original predictors $\mathbf{X}$. Then Model (2) cannot be identified if we allow for nonsparse regression coefficients. However, if we constrain the model size by a certain sparsity level, such as $rspark_c(\mathbf{X}, \mathbf{F})$, the model will become identifiable because $\mathbf{F}$ cannot be represented by sparse linear combinations of $\mathbf{X}$. Using the same idea, if we impose conditions such as the minimum eigenvalue for the covariance matrix of any M_1 features in $(\mathbf{X}, \mathbf{F})$ being bounded from below, where $M_1 = \tilde{c}_1 n/(\log p)$ with $\tilde{c}_1 > \tilde{c}$ denotes the sparsity level, then theorem 2 of Lv (2013) ensures that the robust spark of any submatrix consisting of less than M_1 columns of $(\mathbf{X}, \mathbf{F})$ will be no less than $M = \tilde{c}n/(\log p)$. This holds for general distributions with tail probability decaying exponentially fast with the sample size n and the constant $\tilde{c}$ depending only on c. This justifies the inequality in Condition 4.

Although no distributional assumptions are imposed on the random design matrix $\mathbf{X}$, Condition 5 places a mild constraint that the L_2-norm of any column vector of $\mathbf{X}$ divided by its common scale $n^{1/2}$ is bounded with significant probability. This can be satisfied by many distributions and is needed due to the rescaling of $\boldsymbol{\beta}_*$ in (3). Condition 6 is similar to that of Zheng et al. (2014) for deriving the global properties via the threshold regression. The first part puts a sparsity constraint on the true model size s for model identifiability, as discussed after Condition 4, whereas the second part gives a lower bound $O\{[s(\log p)/n]^{1/2}\}$ on the minimum signal strength to distinguish the significant predictors from the others.

3.4. Main Results

We provide two main theorems in this section. The first is concerned with the asymptotic properties of sample principal components and their score vectors, which serves as a bridge for establishing the global properties in the second theorem.

A sample principal component is said to be consistent with its population counterpart if the angle between them converges to zero asymptotically. However, when several population eigenvalues belong to the same group, the corresponding principal components may not be distinguishable. In this case, subspace consistency is essential to characterizing the asymptotic properties (Jung and Marron 2009). Denote $\theta_{il} = Angle(\widehat{\mathbf{u}}_i, span\{\mathbf{u}_j : j \in J_l\})$ for $i \in J_l$, $1 \le l \le m$, which is the angle between the ith sample principal component and the subspace spanned by population principal components in the corresponding spiked group. The following theorem presents the convergence rates of sample principal components in terms of angles under the aforementioned generalized spiked covariance model. Moreover, for the identifiability of latent factors, we assume each group size to be one for the spiked eigenvalues when studying the principal component score vectors. That is, $k_l = 1$ for $1 \le l \le m$, implying that $K = m$.

Theorem 1 (Convergence Rates). *Under Conditions 1 and 2, with probability approaching one, the following statements hold:*

a. *Uniformly over $i \in J_l$, $1 \le l \le m$, $\theta_{il} = Angle(\widehat{\mathbf{u}}_i, span\{\mathbf{u}_j : j \in J_l\})$ is no more than*

$$\arccos\left(\left[1 - \sum_{t=1}^{l-1}\left[\prod_{i=t+1}^{l-1}(1+k_i)\right]O\{k_t A(t)\} - O\{A(l)\}\right]^{1/2}\right), \tag{6}$$

where $A(t) = (\sum_{l=t+1}^{m} k_l q^{\alpha_l} + k_{m+1})K^{-1}q^{\alpha-\alpha_t}$, and we define $\sum_{t=i}^{j} s_t = 0$ and $\prod_{t=i}^{j} s_t = 1$ if $j < i$ for any sequence $\{s_t\}$.

b. *If each group of spiked eigenvalues has size one, then uniformly over $1 \le i \le K$, $\omega_{ii} = Angle(\mathbf{W}\widehat{\mathbf{u}}_i, \mathbf{W}\mathbf{u}_i)$ is no more than*

$$\arccos\left(\left[1 - \sum_{t=1}^{i-1} 2^{i-t-1}O\{A(t)\} - O\{A(i)\}\right]^{1/2}\right).$$

Part (a) of Theorem 1 provides the uniform convergence rates of sample principal components to the corresponding subspaces for a general spiked covariance structure with possibly tiered eigenvalues under mild conditions. It holds even if the principal components are not separable, so the results also apply to the first kind of applications of Model (2) discussed in Section 2.1. Because the convergence rates of θ_{il}^2 to zero and $\cos^2(\theta_{il})$ to one are the same by L'Hôpital's rule, both of them are $\sum_{t=1}^{l-1}[\prod_{i=t+1}^{l-1}(1+k_i)]O\{k_t A(t)\} + O\{A(l)\}$ in view of (6). Thus, when the group sizes k_l are relatively small, the convergence rates are determined by $A(t)$, which decays polynomially with q and converges to zero fairly fast. This shows the "blessing of dimensionality" under the spiked covariance structure because the larger q gives faster convergence rates. Furthermore, it is clear that when the gaps between the magnitudes of different spiked groups are large, $A(t)$ decays quickly with q to accelerate the convergence of sample principal components.

The uniform convergence rates of sample principal component score vectors are given in part (b) of Theorem 1 when each group contains only one spiked eigenvalue such that the latent factors are separable. In fact, the proof of Theorem 1 shows that the sample score vectors converge at least as fast as the sample principal components. Then the results in part (b) are essentially the convergence rates in part (a) with $k_l = 1$. Because the number of spiked eigenvalues K is much smaller than q, the sample principal component score vectors will converge to the population counterparts polynomially with q. The convergence property of sample score vectors is critical to our purpose of nonsparse learning because it offers the estimation accuracy of latent variables, which is much less well studied in the literature. To the best of our knowledge, our work is a first attempt in high dimensions.

The established asymptotic property of sample principal component score vectors justifies the estimation accuracy assumption in Condition 3. Together with Condition 4, this leads to the following proposition.

Proposition 1. *Under Conditions 3 and 4, the inequality*

$$\|n^{-1/2}\left(\mathbf{X}, \widehat{\mathbf{F}}\right)\boldsymbol{\delta}\|_2 \ge c_1\|\boldsymbol{\delta}\|_2$$

holds for some positive constant c_1 and any $\boldsymbol{\delta}$ satisfying $\|\boldsymbol{\delta}\|_0 < M$ with probability at least $1 - \theta_1 - \theta_2$.

From the proof of Proposition 1, we see that the constant c_1 is smaller than but can be very close to c when n is relatively large. Therefore, Proposition 1 shows that the robust spark of the augmented design matrix $(\mathbf{X}, \widehat{\mathbf{F}})$ will be close to that of $(\mathbf{X}, \mathbf{F})$ when $\mathbf{F}$ is accurately estimated by $\widehat{\mathbf{F}}$. We are now ready to present theoretical properties for the proposed methodology.

Theorem 2 (Global properties). *Assume that Conditions 3–6 hold and*

$$c_1^{-1}c_2\sqrt{(2s+1)(\log p)/n} < \lambda < L^{-1}b_0\left[1 \wedge \left(c_1/\sqrt{2}\right)\right].$$

Then, for both the hard-thresholding penalty $p_{H,\lambda}(t)$ and L_0-penalty $p_{H_0,\lambda}(t)$, with probability at least $1 - 4\sigma(2/\pi)^{1/2} c_2^{-1}(\log p)^{-1/2}p^{1-\frac{c_2^2}{8\sigma^2}} - 2\sigma(2/\pi)^{1/2}c_2^{-1}s(\log n)^{-1/2}\cdot n^{-\frac{c_2^2}{8\sigma^2}} - \theta_1 - \theta_2 - \theta_3$, the regularized estimator $(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T$ satisfies that

a. The model selection consistency $\mathrm{supp}\{(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T\} = \mathrm{supp}\{(\boldsymbol{\beta}_0^T, \boldsymbol{\gamma}_0^T)^T\}$, *where* supp *denotes the support of a vector;*

b. The prediction error bound $n^{-1/2}\|(\mathbf{X}, \widehat{\mathbf{F}})(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T - (\mathbf{X}, \mathbf{F})(\boldsymbol{\beta}_0^T, \boldsymbol{\gamma}_0^T)^T\|_2 \le (c_2/2 + 2c_2c_1^{-1}\sqrt{s})\sqrt{(\log n)/n}$;

c. The oracle inequalities $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_q \le 2c_1^{-2}c_2Ls^{1/q}\sqrt{(\log n)/n}$, $\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_q \le 2c_1^{-2}c_2s^{1/q}\sqrt{(\log n)/n}$ *for $q \in [1, 2]$. The upper bounds with $q = 2$ also hold for $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_\infty$ and $\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_\infty$.*

The model selection consistency in Theorem 2(a) shows that we can recover both the significant observable predictors and the latent variables, so the whole model would be identified by combining these two parts even if it contains nonsparse coefficients. The prediction loss of the joint estimator is shown to be within a logarithmic factor $(\log n)^{1/2}$ of that of the oracle estimator when the regularization parameter λ is properly chosen, which is similar to the result in Zheng et al. (2014). This means that the prediction accuracy is maintained regardless of the hidden effects as long as the latent factors are properly estimated. The extra term $(c_2/2)\sqrt{(\log n)/n}$ in the prediction bound reflects the price we pay in estimating the confounding factors. Furthermore, the oracle inequalities for both $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\gamma}}$ under L_q-estimation losses with $q \in [1,2] \cup \{\infty\}$ are also established in Theorem 2(c). Although the estimation accuracy for the nonsparse coefficients $\mathbf{U}\boldsymbol{\gamma}_0$ of $\mathbf{W}$ are obtainable, we omit the results here because their roles in inferring the individual effects and prediction are equivalent to those of the latent variables.

The proposed methodology of NSL under the conditional sparsity structure is not restrictive to the potential family of population principal components. It is more broadly applicable to any latent family provided that the estimation accuracy of latent factors in Condition 3 and the correlations between the observable predictors and latent factors characterized by the robust spark in Condition 4 hold similarly. The population principal component provides a common and concrete example to extract the latent variables from additional covariates. A significant advantage of this methodology is that even if the estimated latent factors are irrelevant, they rarely affect the variable selection and effect estimation of the original predictors because the number of potential latent variables is generally a small proportion of that of the predictors. This also implies that a relatively large K can be chosen when we are not sure about how many latent variables indeed exist. This is a key difference between our methodology and those based on factor analysis, which renders our methodology useful for combining additional sources.

4. Numerical Studies

In this section, we investigate the finite sample performance of NSL via three regularization methods of Lasso (Tibshirani 1996), smoothly clipped absolute deviation (SCAD; Fan and Li 2001), and the threshold regression using hard thresholding (Hard; Zheng et al. 2014). All three methods are implemented through the independent component analysis algorithm (Fan and Lv 2011) because coordinate optimization enjoys scalability for large-scale problems. The oracle procedure (Oracle) that knew the true model in advance is also conducted as a benchmark.

We will explore two different models, where Model M_1 involves only observable predictors and Model M_2 incorporates estimated latent variables as extra predictors. The case of linear regression Model (2) with the confounding factor as nonsparse combination of the existing predictors is considered in the first example, whereas in the second example multiple latent factors stem from additional observable covariates, and the error vector is relatively heavy tailed with a t-distribution.

4.1. Simulation Examples

4.1.1. Simulation Example 1. In the first simulation example, we consider a special case of linear regression Model (2) with potential latent factors $\mathbf{F}$ coming from the existing observable predictors—that is, $\mathbf{W} = \mathbf{X}$. Then $\mathbf{F}\boldsymbol{\gamma}$ represents the nonsparse effects of the predictors $\mathbf{X}$, and it will be interesting to check the impacts of latent variables when they are dense linear combinations of the existing predictors. The sample size n was chosen to be 100 with true regression coefficient vectors $\boldsymbol{\beta}_0 = (\mathbf{v}^T, \ldots, \mathbf{v}^T, \mathbf{0})^T$, $\boldsymbol{\gamma}_0 = (0.5, \mathbf{0})^T$, and Gaussian error vector $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$, where $\mathbf{v} = (0.6, 0, 0, -0.6, 0, 0)^T$ is repeated k times, and $\boldsymbol{\gamma}_0$ is a K-dimensional vector with one nonzero component 0.5, denoting the effect of the significant confounding factor. We generated 200 data sets and adopted the setting of $(p, k, K, \sigma) = (1{,}000, 3, 10, 0.4)$ such that there are 6 nonzero components with magnitude 0.6 in the true coefficient vector $\boldsymbol{\beta}_0$ and 10 potential latent variables.

The key point in the design of this simulation study is to construct a population covariance matrix $\boldsymbol{\Sigma}$ with spiked structure. Therefore, for each data set, the rows of the $n \times p$ design matrix $\mathbf{X}$ were sampled as i.i.d. copies from a multivariate normal distribution $N(\mathbf{0}, \Sigma)$ with $\Sigma = 1/2(\Sigma_1 + \Sigma_2)$, where $\Sigma_1 = (0.5^{|i-j|})_{1 \le i,j \le p}$ and $\Sigma_2 = 0.5\mathbf{I}_p + 0.5\mathbf{1}\mathbf{1}^T$. The choice of Σ_1 allows for correlation between the predictors at the population level, and Σ_2 has an eigenstructure such that the spiked eigenvalue is comparable with p. Based on the construction of Σ_1 and Σ_2, it is easy to check that Σ has the largest eigenvalue 251.75, and the others are all below 1.75. For regularization methods, Model M_2 involved the top K sample principal components as estimated latent variables, whereas Oracle used the true confounding factor instead of the estimated one. We applied Lasso, SCAD, and Hard for both M_1 and M_2 to produce a sequence of sparse models and selected the regularization parameter λ by minimizing the prediction error calculated based on an independent validation set for fair comparison of all methods.

To compare the performance of the aforementioned methods under two different models, we consider several performance measures. The first measure is the *prediction error* (PE), defined as $E(Y - \mathbf{x}^T\widehat{\boldsymbol{\beta}})^2$ in Model M_1 and as $E(Y - \mathbf{x}^T\widehat{\boldsymbol{\beta}} - \widehat{\mathbf{f}}^T\widehat{\boldsymbol{\gamma}})^2$ in Model M_2, where $\widehat{\boldsymbol{\beta}}$ or $(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T$ are the estimated coefficients in the corresponding models, $(\mathbf{x}^T, Y)$ is an independent test sample of size 10,000, and $\widehat{\mathbf{f}}$ is the sample principal component score vector. For Oracle, $\widehat{\mathbf{f}}$ is replaced by the true confounding factor $\mathbf{f}$. The second to fourth measures are the L_q-estimation losses of $\boldsymbol{\beta}_0$—that is, $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_q$ with $q = 2, 1$, and ∞, respectively. The fifth and sixth measures are the false positives (FP), falsely selected noise predictors, and false negatives (FN), missed true predictors with respect to $\boldsymbol{\beta}_0$. The seventh measure is the model selection consistency (MSC) calculated as the frequency of selecting exactly the relevant variables. We also reported the estimated error standard deviation $\widehat{\sigma}$ by all methods in both models. The results are summarized in Table 1. For the selection and effect estimation of latent variables in Model M_2, we display in Table 2 the measures similar to those defined in Table 1 but with respect to $\boldsymbol{\gamma}_0$. They are L_q-estimation losses $\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_q$ with $q = 2, 1$, and ∞, FP_γ, FN_γ, and MSC_γ.

In view of Table 1, it is clear that compared with Model M_2, the performance measures in variable selection, estimation, and prediction all deteriorated seriously in Model M_1, where most of important predictors were missed, and both the estimation and prediction errors were quite large. We want to emphasize that in this first example, the latent variables are linear combinations of the observable predictors initially included in the model, which means that the nonsparse effects would not be captured without the help of estimated confounding factors. By contrast, the prediction and estimation errors of all regularization methods were reasonably small in the latent variable–augmented Model M_2. It is worth noticing that the performance of Hard was comparable to that of Oracle regardless of the estimation errors of latent features, which is in line with the theoretical results in Theorem 2. Furthermore, we can see from Table 2 that all methods with the estimated latent variables correctly identified the true confounding factor and accurately recovered its effect.

Table 1. Means and Standard Errors (in Parentheses) of Different Performance Measures by All Methods over 200 Simulations in Section 4.1.1

Model	Measure	Lasso	SCAD	Hard	Oracle
M_1	PE	65.27 (1.35)	65.29 (1.40)	68.80 (6.45)	—
	L_2-loss	1.61 (0.24)	1.61 (0.25)	2.25 (1.07)	—
	L_1-loss	4.69 (1.84)	4.70 (1.88)	5.03 (2.31)	—
	L_∞-loss	0.65 (0.13)	0.65 (0.15)	1.48 (1.10)	—
	FP	4.45 (7.15)	4.45 (7.16)	0.51 (0.90)	—
	FN	5.93 (0.26)	5.93 (0.26)	5.98 (0.16)	—
	MSC	0 (0)	0 (0)	0 (0)	—
	$\widehat{\sigma}$	7.88 (0.57)	7.88 (0.57)	7.78 (0.60)	—
M_2	PE	0.39 (0.16)	0.19 (0.01)	0.19 (0.01)	0.17 (0.01)
	L_2-loss	0.43 (0.13)	0.13 (0.03)	0.10 (0.03)	0.10 (0.03)
	L_1-loss	1.52 (0.40)	0.44 (0.07)	0.21 (0.13)	0.21 (0.06)
	L_∞-loss	0.23 (0.07)	0.07 (0.02)	0.07 (0.02)	0.07 (0.02)
	FP	28.79 (6.52)	15.99 (5.63)	0.02 (0.28)	0 (0)
	FN	0.02 (0.16)	0 (0)	0 (0)	0 (0)
	MSC	0 (0)	0 (0)	1.00 (0.07)	1 (0)
	$\widehat{\sigma}$	0.47 (0.06)	0.38 (0.03)	0.41 (0.03)	0.40 (0.03)

Note. M_1, model with only observable predictors; M_2, model includes estimated latent variables; SCAD, smoothly clipped absolute deviation; PE, prediction error; FP, false positives; FN, false negatives; MSC, model selection consistency.

Table 2. Means and Standard Errors (in Parentheses) of Different Performance Measures for Regression Coefficients of Confounding Factors by All Methods over 200 Simulations in Section 4.1.1

Measure	Lasso	SCAD	Hard	Oracle
L_2-loss	0.02 (0.00)	0.01 (0.00)	0.01 (0.00)	0.00 (0.00)
L_1-loss	0.02 (0.01)	0.01 (0.00)	0.01 (0.00)	0.00 (0.00)
L_∞-loss	0.02 (0.00)	0.01 (0.00)	0.01 (0.00)	0.00 (0.00)
FP_γ	0.29 (0.55)	0.21 (0.43)	0 (0)	0 (0)
FN_γ	0 (0)	0 (0)	0 (0)	0 (0)
MSC_γ	0.73 (0.45)	0.79 (0.41)	1 (0)	1 (0)

Notes. The notation 0.00 denotes a number less than 0.005. SCAD, smoothly clipped absolute deviation.

4.1.2. Simulation Example 2. Now we consider a more general case where the latent variables stem from a group of observable covariates instead of the original predictors. Moreover, we also want to see whether similar results hold when more significant confounding factors are involved and the errors become relatively heavy tailed. Thus, there are three main changes in the setting of this second example. First, the predictors $\mathbf{X}$ and observable covariates $\mathbf{W}$ are different, as well as their covariance structures, which will be specified later. Second, there are two significant latent variables and the K-dimensional true coefficient vector $\boldsymbol{\gamma}_0 = (0.5, -0.5, \mathbf{0})^T$. Third, the error vector $\boldsymbol{\varepsilon} = \sigma\boldsymbol{\eta}$, where the components of the n-dimensional random vector $\boldsymbol{\eta}$ are independent and follow the t-distribution with $df = 10$ degrees of freedom. The settings of $\boldsymbol{\beta}_0$ and (n, p, K, σ) are the same as in the first simulation example in Section 4.1.1, whereas the dimensionality q of covariates $\mathbf{W}$ equals 1,000, which is also large.

For the covariance structure of $\mathbf{X}$, we set $\Sigma_X = (0.5^{|i-j|})_{1\le i,j\le p}$ to allow for correlation at the population level. By contrast, in order to estimate the principal components in high dimensions, the population covariance matrix of $\mathbf{W}$ should have multiple spiked eigenvalues. Thus, we constructed it using the block diagonal structure such that

$$\Sigma_W = \begin{pmatrix} \Sigma_{11} & \mathbf{0} \\ \mathbf{0} & \Sigma_{22} \end{pmatrix},$$

where $\Sigma_{11} = 3/4(\Sigma_1 + \Sigma_2)_{1\le i,j\le 200}$ and $\Sigma_{22} = 1/2\ (\Sigma_1 + \Sigma_2)_{1\le i,j\le 800}$ with the definitions of Σ_1 and Σ_2 similar to those in Section 4.1.1 except for different dimensions. Under such construction, the two largest eigenvalues of Σ_W are 201.75 and 77.61, respectively, whereas the others are less than 2.63. Based on the aforementioned covariance structures, for each data set, the rows of $\mathbf{X}$

and **W** were sampled as i.i.d. copies from the corresponding multivariate normal distribution.

We included the top K sample principal components in Model M_2 as potential latent factors and compared the performance of Lasso, SCAD, Hard, and Oracle by the same performance measures as defined in Section 4.1.1. The results are summarized in Tables 3 and 4. From Table 3, it is clear that the methods that relied only on the observable predictors still suffered a lot under this more difficult setting, where all true predictors were missed, prediction errors were large, and the error standard deviation (SD) was poorly estimated. In contrast, the new NSL methodology via Lasso, SCAD, and Hard was able to tackle the issues associated with variable selection, coefficient estimation, prediction, and error SD estimation. With the latent variable–augmented Model M_2, Hard almost recovered the exact underlying model. Similar to the first example, in view of Table 4, all methods correctly identified the significant confounding factors and estimated their effects accurately. However, compared with Tables 1 and 2, most of the performance measures deteriorated in this second example. This is mainly due to the relatively heavy-tailed random errors, as well as the difficulty in estimating multiple high-dimensional principal components.

4.2. Application to Nutrient Intake and Human Gut Microbiome Data

Nutrient intake strongly affects human health and diseases such as obesity, whereas gut microbiome composition is an important factor in energy extraction from the diet. We illustrate the usefulness of our proposed methodology by applying it to the data set reported by Wu et al. (2011) and previously studied by Chen and Li (2013) and Lin et al. (2014), where a cross-sectional study of 98 healthy volunteers was carried out to investigate the habitual diet effect on the human gut microbiome. The nutrient intake consisted of 214 micronutrients collected from the volunteers by a food frequency questionnaire. The values were normalized by the residual method to adjust for caloric intake and then standardized to have mean zero and SD one. Similar to Chen and Li (2013), we used one representative for a set of highly correlated micronutrients whose correlation coefficients are larger than 0.9, resulting in 119 representative micronutrients in total. Furthermore, stool samples were collected, and DNA samples were analyzed by Roche 454 pyrosequencing of 16S rDNA gene segments from the V1–V2 region. After taxonomic assignment of the denoised pyrosequences, the operational taxonomic units were combined into 87 genera that appeared in at least one sample. We are interested in identifying the important micronutrients and potential latent factors from the gut microbiome genera that are associated with the BMI.

Because of the high correlations between the micronutrients, we applied NSL via the elastic net (Zou and Hastie 2005) to this data set by treating BMI, nutrient intake, and gut microbiome composition (after the centered log-ratio transformation; Aitchison 1983) as the response, predictors, and covariates of confounding factors, respectively. The data set was split 100 times into a training set of 60 samples and a validation set of the remaining samples. For each

Table 3. Means and Standard Errors (in Parentheses) of Different Performance Measures by All Methods over 200 Simulations in Section 4.1.2

Model	Measure	Lasso	SCAD	Hard	Oracle
M_1	PE	72.33 (1.53)	72.33 (1.53)	76.04 (6.62)	—
	L_2-loss	1.58 (0.24)	1.58 (0.24)	2.25 (1.09)	—
	L_1-loss	4.49 (1.86)	4.49 (1.86)	5.00 (2.10)	—
	L_∞-loss	0.64 (0.13)	0.64 (0.13)	1.50 (1.15)	—
	FP	3.59 (6.52)	3.59 (6.52)	0.49 (0.72)	—
	FN	5.95 (0.23)	5.95 (0.23)	6.00 (0.07)	—
	MSC	0 (0)	0 (0)	0 (0)	—
	Error SD	8.33 (0.59)	8.33 (0.59)	8.20 (0.63)	—
M_2	PE	1.74 (1.08)	1.10 (1.05)	1.04 (0.99)	0.22 (0.01)
	L_2-loss	0.70 (0.22)	0.25 (0.22)	0.16 (0.18)	0.11 (0.03)
	L_1-loss	2.18 (0.50)	0.87 (0.50)	0.39 (0.53)	0.23 (0.07)
	L_∞-loss	0.37 (0.12)	0.13 (0.10)	0.10 (0.09)	0.08 (0.03)
	FP	20.63 (13.60)	23.29 (12.52)	0.70 (3.34)	0 (0)
	FN	0.09 (0.38)	0.15 (0.94)	0.09 (0.63)	0 (0)
	MSC	0 (0)	0.09 (0.29)	0.92 (0.27)	1 (0)
	Error SD	0.74 (0.20)	0.48 (0.21)	0.50 (0.12)	0.45 (0.04)

Notes. M_1, model with only observable predictors; M_2, model includes estimated latent variables. The population error standard deviation $\sigma\sqrt{df/(df-2)}$ equals to 0.45. SCAD, smoothly clipped absolute deviation; PE, prediction error; FP, false positives; FN, false negatives; MSC, model selection consistency; SD, standard deviation.

Table 4. Means and Standard Errors (in Parentheses) of Different Performance Measures for Regression Coefficients of Confounding Factors by All Methods over 200 Simulations in Section 4.1.2

Measure	Lasso	SCAD	Hard	Oracle
L_2-loss	0.08 (0.03)	0.07 (0.04)	0.07 (0.04)	0.01 (0.00)
L_1-loss	0.10 (0.04)	0.09 (0.05)	0.08 (0.05)	0.01 (0.00)
L_∞-loss	0.08 (0.03)	0.06 (0.03)	0.06 (0.03)	0.01 (0.00)
FP_γ	0.21 (0.45)	0.34 (0.60)	0.01 (0.10)	0 (0)
FN_γ	0 (0)	0 (0)	0 (0)	0 (0)
MSC_γ	0.81 (0.39)	0.72 (0.45)	0.99 (0.10)	1 (0)

Notes. The notation 0.00 denotes a number less than 0.005. SCAD, smoothly clipped absolute deviation,

splitting of the data set, we explored two different models, M_1 and M_2, as defined in Section 4.1, with the top 20 sample principal components of gut microbiome composition included in Model M_2 to estimate the potential latent factors. All predictors were rescaled to have a common L_2-norm of $n^{1/2}$, and the tuning parameter was chosen by minimizing the prediction error calculated on the validation set. We summarize in Table 5 the selection probabilities and coefficients of the significant micronutrients and latent variables whose selection probabilities were greater than 0.9 in M_1 or greater than 0.85 in M_2. The means (with standard errors in parentheses) of the prediction errors averaged over 100 random splittings were 167.9 (7.2) in Model M_1 and 110.3 (4.0) in Model M_2, whereas the median model size also decreased from 93 to 69 after applying the NSL methodology. This shows that the prediction performance was improved after using the information on gut microbiome genera.

In view of the model selection results in Table 5, many significant micronutrients in Model M_1 became insignificant after adjusting for the latent substructures, which implies that either they affect BMI through the gut microbiome genera or their combinative effects are captured by the latent variables. This was also evidenced by the reduction in model size mentioned earlier. Moreover, the effects of some micronutrients changed signs in Model M_2, and the subsequent associations with BMI are consistent with scientific discoveries (Gul et al. 2017). For instance, aspartame is a sugar substitute widely used in beverages such as diet soda, and it was negatively associated with BMI in Model M_1 but tended to share a positive association after accounting for the gut microbiome genera. A potential reason is that the people who drink diet soda can have a relatively healthy habitual diet and gut microbiome composition that, in turn, lower the BMI, but the diet soda itself does not reduce fat. Similar phenomena happened with both acrylamide and vitamin E as well.

We also applied the model-free knockoffs (Candès et al. 2018) with a target false discovery rate (FDR) level of 0.2 to Model M_2, and the most significant factors identified were the latent variables of 7th and 9th principal components, which may be explained as BMI-associated enterotypes while adjusting for nutrient intake (Arumugam et al. 2011, Wu et al. 2011). The major gut microbiome genera in the compositions of these two latent variables are displayed in Table 6. At the phylum level, the latent factors mainly consist of Bacteroidetes and Firmicutes, whose relative proportion has been shown to affect human obesity (Ley et al. 2006). In view of the associations with BMI, both the 7th and 9th principal components confirm the claim that the Firmicutes-enriched microbiome holds a greater metabolic potential for energy gain from the diet that results in weight gain (Turnbaugh et al. 2006). Furthermore, one of the major microbiome genera in the latent factor of the 9th principal component, Acidaminococcus, was also found to be positively associated with BMI by Lin et al. (2014), who show that human obesity can be affected at the genus level.

5. Discussion

In this paper, we have introduced a new methodology, NSL, for prediction and variable selection in the

Table 5. Selection Probabilities and Rescaled Coefficients (in Parentheses) of the Most Frequently Selected Predictors by Each Model Across 100 Random Splittings in Section 4.2

Predictor	Model M_1	Model M_2	Predictor	Model M_1	Model M_2
Sodium	0.98 (1.35)	0.67 (0.55)	PC(7th)	—	0.99 (1.76)
Eicosenoic acid	0.98 (–2.47)	0.80 (–1.24)	—	—	0.96 (–1.21)
Vitamin B_{12}	0.96 (0.43)	0.62 (0.30)	Apigenin	0.95 (–1.67)	0.93 (–1.88)
Gallocatechin	0.96 (–4.81)	0.84 (–1.70)	PC(9th)	—	0.88 (–0.87)
Riboflavin pills	0.94 (1.71)	0.55 (0.61)	PC(10th)	—	0.86 (0.78)
Acrylamide	0.94 (–0.34)	0.62 (0.32)	Iron	0.93 (1.22)	0.86 (0.75)
Naringenin	0.94 (1.11)	0.58 (0.32)	Aspartame	0.93 (–0.46)	0.79 (0.59)
Pelargonidin	0.94 (–1.15)	0.75 (–1.03)	Vitamin C	0.93 (–0.71)	0.76 (–0.39)
Lauric acid	0.93 (1.88)	0.71 (0.50)	Vitamin E	0.92 (0.45)	0.65 (–0.29)

Note. M_1, model with only micronutrients as predictors; M_2, model includes latent variables from gut microbiome composition; PC, principal component.

Table 6. Major Gut Microbiome Genera in the Compositions of the Two Significant Latent Variables Identified by the Model-Free Knockoffs in Section 4.2

Latent variable	Phylum	Genus	Weight
PC(7th)	Firmicutes	*Dialister*	−0.40
	Firmicutes	*Eubacterium*	0.39
	Bacteroidetes	*Barnesiella*	−0.28
PC(9th)	Firmicutes	*Acidaminococcus*	−0.51
	Firmicutes	*Megasphaera*	−0.36
	Firmicutes	*Ruminococcus*	−0.30

Note. PC, principal component.

presence of nonsparse coefficient vectors through the factors plus sparsity structure, where latent variables are exploited to capture the nonsparse combinations of either the original predictors or additional covariates. The suggested methodology is ideal for the applications involving two sets of features that are strongly correlated, as in our BMI study. Both theoretical guarantees and empirical performance of the potential latent family incorporating population principal components have been demonstrated. And our methodology is also applicable to more general families with properly estimated latent variables and identifiable models.

It would be interesting to further investigate several problems, such as hypothesis testing and FDR control in nonsparse learning by the idea of NSL. Based on the established model identifiability condition that characterizes the correlations between observable and latent predictors, hypothesis testing can proceed using the debiasing idea in Javanmard and Montanari (2014), van de Geer et al. (2014), Zhang and Zhang (2014). FDR could be controlled by applying the knockoffs inference procedures (Barber and Candès 2015, Candès et al. 2018, Fan et al. 2020) on the latent variable–augmented model. The main difficulty lies in analyzing how the estimation errors of unobservable factors affect the corresponding procedures. Another possible direction is to explore more general ways of modeling the latent variables to deal with the nonsparse coefficient vectors. These problems are beyond the scope of this paper and will be interesting topics for future research.

Acknowledgments
The authors sincerely thank the editors and referees for their valuable comments that helped improve the article substantially.

References
Aitchison J (1983) Principal component analysis of compositional data. *Biometrika* 70(1):57–65.

Alter O, Brown P, Botstein D (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. USA* 97(18):10101–10106.

Artemiou A, Li B (2009) On principal components and regression: A statistical explanation of a natural phenomenon. *Statist. Sinica* 19(4):1557–1565.

Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, Fernandes GR, et al. (2011) Enterotypes of the human gut microbiome. *Nature* 473(7346):174–180.

Bair E, Hastie T, Paul D, Tibshirani R (2006) Prediction by supervised principal components. *J. Amer. Statist. Assoc.* 101(437):119–137.

Barber R, Candès EJ (2015) Controlling the false discovery rate via knockoffs. *Ann. Statist.* 43(5):2055–2085.

Belloni A, Chernozhukov V, Wang L (2011) Square-root lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika* 98(4):791–806.

Belloni A, Chernozhukov V, Chetverikov D, Wei Y (2018) Uniformly valid post-regularization confidence regions for many functional parameters in *Z*-estimation framework. *Ann. Statist.* 46(6B):3643–3675.

Belloni A, Chernozhukov V, Fernández-Val I, Hansen C (2017) Program evaluation and causal inference with high-dimensional data. *Econometrica* 85(1):233–298.

Bickel PJ, Ritov Y, Tsybakov AB (2009) Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.* 37(4):1705–1732.

Bing X, Bunea F, Ning Y, Wegkamp M (2020) Sparse latent factor models with pure variables for overlapping clustering. *Ann. Statist.* 48(4):2055–2081.

Bing X, Bunea F, Wegkamp M, Strimas-Mackey S (2019) Essential regression. arXiv:1905.12696.

Boyle E, Li Y, Pritchard J (2017) An expanded view of complex traits: From polygenic to omnigenic. *Cell* 169(7):1177–1186.

Bradic J, Fan J, Zhu Y (2020) Testability of high-dimensional linear models with non-sparse structures. *Ann. Statist.* Forthcoming.

Bühlmann P, van de Geer S (2011) *Statistics for High-Dimensional Data: Methods, Theory and Applications* (Springer, Berlin).

Candès EJ, Fan Y, Janson L, Lv J (2018) Panning for gold: "Model-X" knockoffs for high-dimensional controlled variable selection. *J. Royal Statist. Soc. B* 80(3):551–577.

Candès EJ, Tao T (2007) The Dantzig selector: Statistical estimation when p is much larger than n (with discussion). *Ann. Statist.* 35(6):2313–2404.

Cavalli-Sforza LL, Menozzi P, Piazza A (1993) Demic expansions and human evolution. *Science* 259(5095):639–646.

Chandrasekaran V, Parrilo PA, Willsky AS (2012) Latent variable graphical model selection via convex optimization (with discussion). *Ann. Statist.* 40(4):1935–1967.

Chen J, Li H (2013) Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. *Ann. Appl. Stat.* 7(1):418–442.

Chen M, Ren Z, Zhao H, Zhou HH (2016) Asymptotically normal and efficient estimation of covariate-adjusted Gaussian graphical model. *J. Amer. Statist. Assoc.* 111513:394–406.

Cook RD (2007) Fisher lecture: Dimension reduction in regression. *Statist. Sci.* 22(1):1–40.

Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* 96(456):1348–1360.

Fan J, Lv J (2010) A selective overview of variable selection in high dimensional feature space. *Statist. Sinica* 20(1):101–148.

Fan J, Lv J (2011) Nonconcave penalized likelihood with NP-dimensionality. *IEEE Trans. Inform. Theory* 57(8):5467–5484.

Fan J, Guo S, Hao N (2012) Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *J. Royal Statist. Soc. B* 74(1):37–65.

Fan J, Liao Y, Micheva M (2013) Large covariance estimation by thresholding principal orthogonal complements (with discussion). *J. Royal Statist. Soc. B* 75(4):603–680.

Fan Y, Lv J (2013) Asymptotic equivalence of regularization methods in thresholded parameter space. *J. Amer. Statist. Assoc.* 108(503): 247–264.

Fan Y, Demirkaya E, Li G, Lv J (2020) RANK: Large-scale inference with graphical nonlinear knockoffs. *J. Amer. Statist. Assoc.* 115(529):362–379.

Gul S, Hamilton AR, Munoz AR, Phupitakphol T, Liu W, Hyoju SK, Economopoulos KP, et al. (2017) Inhibition of the gut enzyme intestinal alkaline phosphatase may explain how aspartame promotes glucose intolerance and obesity in mice. *Appl. Physiol. Nutrition Metabolism* 42(1):77–83.

Hastie T, Tibshirani R, Friedman J (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. (Springer, New York).

Hsu HL, Ing CK, Tong H (2019) On model selection from a finite family of possibly misspecified time series models. *Ann. Statist.* 47(2):1061–1087.

Javanmard A, Montanari A (2014) Confidence intervals and hypothesis testing for high-dimensional regression. *J. Machine Learn. Res.* 15(82):2869–2909.

Johnstone I (2001) On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* 29(2):295–327.

Johnstone I, Lu A (2009) On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.* 104(486):682–693.

Jung S, Marron J (2009) PCA consistency in high dimension, low sample size context. *Ann. Statist.* 37(6):4104–4130.

Kneip A, Sarda P (2011) Factor models and variable selection in high-dimensional regression analysis. *Ann. Statist.* 39(5):2410–2447.

Leek J, Storey J (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics* 3(9): 1724–1735.

Ley RE, Turnbaugh PJ, Klein S, Gordon JI (2006) Human gut microbes associated with obesity. *Nature* 444(7122):1022–1023.

Lin W, Feng R, Li H (2015) Regularization methods for high-dimensional instrumental variables regression with an application to genetical genomics. *J. Amer. Statist. Assoc.* 110(509): 270–288.

Lin W, Shi P, Feng R, Li H (2014) Variable selection in regression with compositional covariates. *Biometrika* 101(4):785–797.

Lv J (2013) Impacts of high dimensionality in finite samples. *Ann. Statist.* 41(4):2236–2262.

Lv J, Liu JS (2014) Model selection principles in misspecified models. *J. Royal Statist. Soc. B* 76(1):141–167.

Mardia K, Kent J, Bibby J (1979) *Multivariate Analysis* (Academic Press, New York).

Menozzi P, Piazza A, Cavalli-Sforza LL (1978) Synthetic maps of human gene frequencies in Europeans. *Science* 201(4358):786–792.

Pan D, He H, Song X, Sun L (2015) Regression analysis of additive hazards model with latent variables. *J. Amer. Statist. Assoc.* 110(511):1148–1159.

Paul D (2007) Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica* 17(4):1617–1642.

Paulson C, Luo L, James G (2018) Efficient large-scale internet media selection optimization for online display advertising. *J. Marketing Res.* 55(4):489–506.

Price A, Patterson N, Plenge R, Weinblatt M, Shadick N, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genetics* 38(8):904–909.

Pritchard J (2001) Are rare variants responsible for susceptibility to complex diseases? *Amer. J. Human Genetics* 69(1):124–137.

Radchenko P, James G (2008) Variable inclusion and shrinkage algorithms. *J. Amer. Statist. Assoc.* 103(483):1304–1315.

Shen D, Shen H, Marron J (2016) A general framework for consistency of principal component analysis. *J. Machine Learn. Res.* 17(150):5218–5251.

Sun T, Zhang C-H (2012) Scaled sparse linear regression. *Biometrika* 99(4):879–898.

Tang CY, Leng C (2010) Penalized high-dimensional empirical likelihood. *Biometrika* 97(4):905–920.

Tibshirani R (1996) Regression shrinkage and selection via the Lasso. *J. Royal Statist. Soc. B* 58(1):267–288.

Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* 444(7122):1027–1031.

Uematsu Y, Tanaka S (2019) High-dimensional macroeconomic forecasting and variable selection via penalized regression. *Econom. J.* 22(1):34–56.

van de Geer S, Bühlmann P, Ritov Y, Dezeure R (2014) On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* 42(3):1166–1202.

Wang W, Fan J (2017) Asymptotics of empirical eigen-structure for high dimensional spiked covariance. *Ann. Statist.* 45(3): 1342–1374.

White H (1982) Maximum likelihood estimation of misspecified models. *Econometrica* 50(1):1–25.

Wu GD, Chen J, Hoffmann C, Bittinger K, Chen Y-Y, Keilbaugh SA, Bewtra M, et al. (2011) Linking long-term dietary patterns with gut microbial enterotypes. *Science* 334(6052):105–108.

Xu H, Caramanis C, Mannor S (2016) Statistical optimization in high dimensions. *Oper. Res.* 64(4):958–979.

Zhang S, Zhang C-H (2014) Confidence intervals for low dimensional parameters in high dimensional linear models. *J. Royal Statist. Soc. B* 76(1):217–242.

Zhao P, Yu B (2006) On model selection consistency of Lasso. *J. Machine Learn. Res.* 7(November):2541–2563.

Zheng Z, Fan Y, Lv J (2014) High-dimensional thresholded regression and shrinkage effect. *J. Royal Statist. Soc. B* 76(3):627–649.

Zhu Y, Bradic J (2018) Linear hypothesis testing in dense high-dimensional linear models. *J. Amer. Statist. Assoc.* 113(524): 1583–1600.

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. *J. Royal Statist. Soc. B* 67(2):301–320.

Zemin Zheng is a professor in the Department of Statistics and Finance of the School of Management at the University of Science and Technology of China (USTC) and a professor in the School of Data Science and International Institute of Finance at the USTC. His research interests include statistics, machine learning, data science, and business applications.

Jinchi Lv is the Kenneth King Stonier Chair in Business Administration, and professor in the Data Sciences and Operations Department of the Marshall School of Business at the University of Southern California (USC), a professor in the Department of Mathematics at USC, and an associate fellow of USC Dornsife Institute for New Economic Thinking. He is the recipient of an Adobe Data Science Research Award (2017), the Royal Statistical Society Guy Medal in Bronze (2015), and the National Science Foundation Faculty Early Career Development Award (2010).

Wei Lin is an associate professor in the Department of Probability and Statistics of the School of Mathematical Sciences and Center for Statistical Science at Peking University. His research interests lie in the broad areas of high-dimensional statistics and statistical machine learning, with particular emphasis on compositional data analysis and statistical learning from high-dimensional complex data.