Submitted to *Operations Research*
manuscript 2019-02-104

Nonsparse Learning with Latent Variables

Zemin Zheng
The School of Management, the School of Data Science, and International Institute of Finance, University of Science and Technology of China, Hefei 230026, China, zhengzm@ustc.edu.cn

Jinchi Lv
Marshall School of Business, University of Southern California, Los Angeles, CA 90089, jinchilv@marshall.usc.edu

Wei Lin
School of Mathematical Sciences and Center for Statistical Science, Peking University, Beijing 100871, China, weilin@math.pku.edu.cn

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 due to 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.

Key words: 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 that the optimization formulations will contain many more variables and constraints than in fact 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 have been developed to generate meaningful and interpretable models, including Tibshirani (1996), Fan and Li (2001), Zou and Hastie (2005), Candès and Tao (2007), Belloni et al. (2011), Sun and Zhang (2012), 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), 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 due to 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 that the genes associated with complex traits tend to be spread across most of the genome. Similarly, it was conjectured earlier in Pritchard (2001) that instead of being sparse, the causal variants responsible for a trait can be distributed. Under such cases, making 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 in Fan et al. (2013) by 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 in Chandrasekaran et al. (2012) for graphical model selection. To control for confounding in genetical 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 EM 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. Both the numbers of 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 way of nonsparse learning when the nonsparse effects of the covariates can be captured by their leading population principal component score vectors, which are unobservable due to 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 2007) 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. By exploring population principal components as common latent variables, it 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 thresholded 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 thresholded 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 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, while 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 $\mathbf{\Sigma}_W$. 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 $\mathbf{\Sigma}_W$ 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 $\boldsymbol{\gamma}_0 = (\gamma_{0,1}, \ldots, \gamma_{0,K})^T$ their true regression coefficient vector. Then model (1) can be rewritten as

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

It is worth pointing out that the latent variables $\mathbf{F}$ are unobservable to us due to the unknown vectors $\mathbf{u}_i$, which makes our work distinct from most of 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 $\boldsymbol{\Sigma}_W$ and we allow it to diverge with the sample size.

Model (2) is applicable to two different situations. The first one is that we aim at recovering the relationship between predictors $\mathbf{X}$ and response $\mathbf{y}$, while the features $\mathbf{W}$ are treated as confounding variables for making correct inference 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. It occurs in applications when the latent variables are also meaningful. For instance, the principal components of genes can be biologically interpretable that they represent 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 since 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 predictors and adjust for confounding variables by incorporating the principal components of gut microbiome composition,

since 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 speaking, predictors $\mathbf{X}$ stand for features with individual effects while $\mathbf{W}$ are covariates reflecting 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 $\mathbf{W}$. Since 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 nonsparse learning with latent variables which minimizes

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

the penalized residual sum of squares with penalty function $p_\lambda(\cdot)$. Here $\boldsymbol{\beta}_* = (\beta_{*,1}, \ldots, \beta_{*,p})^T$ is the Hadamard (componentwise) 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 $\lambda \geq 0$, and assumed to be increasing in both λ and t with $p_\lambda(0) = 0$. We use a compact notation for

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

The proposed methodology in (3) enables the possibility to simultaneously estimate $\boldsymbol{\beta}_0$ and $\boldsymbol{\gamma}_0$, identifying the significant observable predictors and latent factors altogether. However, it is still difficult to obtain accurate estimates since 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 $\mathbb{B}_\rho = \{\boldsymbol{\gamma} \in \mathbb{R}^K : \|\boldsymbol{\gamma}\|_\infty \leq \rho\}$, 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 in 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, benefitting 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 before.

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 the situations where the response can be regressed on the leading principal components of predictors with little loss of information. Probabilistic explanation was provided in 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, while the model interpretability is also retained by exploring the individual effects of observable predictors.

Besides 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 theirs in that we also allow the presence of covariates with sparse individual effects. Another line of work including 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 them, properties including model selection consistency and oracle inequalities will be proved for the proposed methodology via the thresholded 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 in Paul (2007), Jung and Marron (2009), Shen et al. (2016), Wang and Fan (2017), among many others. This model assumes that the first few eigenvalues of the population covariance matrix deviate from one while 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 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 in Paul (2007) and Shen et al. (2016) for multivariate Gaussian observations. The asymptotic distributions of spiked principal components were established in 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 in Jung and Marron (2009) for the covariance structure of covariate matrix $\mathbf{W}$, where the population covariance

matrix $\mathbf{\Sigma}_W$ is assumed to contain K spiked eigenvalues that can be divided into m groups. The eigenvalues grow at the same rate within each group while 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 since 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 non-spiked eigenvalues. Then the set of indices for the lth group of eigenvalues is

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

Although the above eigen-structure 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. It makes the original convergence analysis of sample eigenvalues and eigenvectors invalid since 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 in, 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 eigen-decomposition of the population covariance matrix $\mathbf{\Sigma}_W$ is given by $\mathbf{\Sigma}_W = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$, where $\mathbf{\Lambda}$ 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 eigen-decomposition of $\mathbf{S} = \widehat{\mathbf{U}}\widehat{\mathbf{\Lambda}}\widehat{\mathbf{U}}^T$ provides the diagonal matrix $\widehat{\mathbf{\Lambda}}$ 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

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

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. Thresholded regression using hard-thresholding

As discussed in Section 1, there are a large spectrum of regularization methods for sparse learning in high dimensions. It has been demonstrated in Fan and Lv (2013) that the popular L_1-regularization of 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 the Lasso. Therefore, we will show theoretical properties of the proposed methodology via a specific concave regularization method, the thresholded regression using hard-thresholding (Zheng et al. 2014). It utilizes either the hard-thresholding penalty $p_{H,\lambda}(t) = \frac{1}{2}\left[\lambda^2 - (\lambda - t)_+^2\right]$ 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 in 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, while 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 $i \in J_l$, $1 \le l \le m$,*

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

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*

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

(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 $\cos(\omega_{jj}) \ge 1 - \frac{c_2^2 \log n}{8K^2\rho^2 n}$ with probability $1 - \theta_1$ that converges to one as $n \to \infty$.*

Condition 4 *The inequality* $\|n^{-1/2}(\mathbf{X},\mathbf{F})\boldsymbol{\delta}\|_2 \geq c\|\boldsymbol{\delta}\|_2$ *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*

$$P\Big(\cap_{j=1}^{p}\big\{L^{-1} \leq \frac{\|\mathbf{x}_j\|_2}{\sqrt{n}} \leq L\big\}\Big) = 1-\theta_3,$$

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

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

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 while their limits can be different, depending on the constants c_i. It is weaker than those usually imposed in the literature such as 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 non-spiked 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 that 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 that 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 since 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. It 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 certain sparsity level, such as $rspark_c(\mathbf{X}, \mathbf{F})$, the model will become identifiable since $\mathbf{F}$ cannot be represented by sparse linear combinations of $\mathbf{X}$. Utilizing 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 (Lv 2013, Theorem 2) 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 $M = \tilde{c}n/(\log p)$. It 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.

While no distributional assumptions are imposed on the random design matrix $\mathbf{X}$, Condition 5 puts 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. It 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 thresholded regression. The first part puts a sparsity constraint on the true model size s for model identifiability as discussed after Condition 4, while 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 one 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 that 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 $K = m$.

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

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

where $A(t) = \big(\sum_{l=t+1}^{m} k_l q^{\alpha_l} + k_{m+1}\big) K^{-1} q^{\alpha - \alpha_t}$ 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*

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

Part (a) of Theorem 1 provides the uniform convergence rates of sample principal components to the corresponding subspaces for general spiked covariance structure with possibly tiered eigenvalues under mild conditions. It holds even if the principal components are not separable, so that the results also apply to the first kind of applications of model (2) discussed in Section 2.1. Since the convergence rates of θ_{il}^2 to zero and $\cos^2(\theta_{il})$ to one are the same by L'Hospital's rule, both of them are $\sum_{t=1}^{l-1} \left[\prod_{i=t+1}^{l-1}(1+k_i)\right] O\{k_t A(t)\} + O\{A(l)\}$ 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. It shows the "blessing of dimensionality" under the spiked covariance structure since 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 $k_l = 1$. Since 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 since 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, it leads to the following proposition.

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

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

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*

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

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

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

(c) $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_q \leq 2c_1^{-2}c_2Ls^{1/q}\sqrt{(\log n)/n}$, $\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|_q \leq 2c_1^{-2}c_2s^{1/q}\sqrt{(\log n)/n}$ *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 that 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). It 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 since their roles in inferring the individual effects and prediction are equivalent to those of the latent variables.

The proposed methodology of nonsparse learning with latent variables 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 since the number of potential latent variables is generally a small proportion of that of the predictors. It 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 it useful for combining additional sources.

4. Numerical studies

In this section, we investigate the finite sample performance of NSL via three regularization methods of the Lasso (Tibshirani 1996), SCAD (Fan and Li 2001), and the thresholded regression using hard-thresholding (Hard) (Zheng et al. 2014). All three methods are implemented through the ICA algorithm (Fan and Lv 2011) since coordinate optimization enjoys scalability for large-scale problems. The oracle procedure (Oracle) which 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, while in the second example multiple latent factors stem from additional observable covariates and the error vector is relatively heavy-tailed with 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 $(p, k, K, \sigma) = (1000, 3, 10, 0.4)$ such that there are six nonzero components with magnitude 0.6 in the true coefficient vector $\boldsymbol{\beta}_0$ and ten 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 $N(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = \frac{1}{2}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)$, where $\boldsymbol{\Sigma}_1 = (0.5^{|i-j|})_{1 \le i,j \le p}$ and $\boldsymbol{\Sigma}_2 = 0.5\mathbf{I}_p + 0.5\mathbf{1}\mathbf{1}^T$. The choice of $\boldsymbol{\Sigma}_1$ allows for correlation between the predictors at the population level and $\boldsymbol{\Sigma}_2$ has an eigen-structure such that the spiked eigenvalue is comparable with p. Based on the construction of $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, it is easy to check that $\boldsymbol{\Sigma}$ 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 while the oracle procedure used the true confounding factor instead of the estimated one. We applied the 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 the oracle procedure, $\widehat{\mathbf{f}}$ 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. On the other hand, 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 the oracle procedure 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.

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

Table 1 Means and standard errors (in parentheses) of different performance measures by all methods over 200 simulations in Section 4.1.1; M_1: model with only observable predictors, M_2: model includes estimated latent variables

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)

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, while the dimensionality q of covariates $\mathbf{W}$ equals 1000, which is also large.

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 (The notation 0.00 denotes a number less than 0.005.)

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)

For the covariance structure of $\mathbf{X}$, we set $\mathbf{\Sigma}_X = (0.5^{|i-j|})_{1\le i,j\le p}$ to allow for correlation at the population level. On the other hand, 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

$$\mathbf{\Sigma}_W = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_{22} \end{pmatrix},$$

where $\mathbf{\Sigma}_{11} = \frac{3}{4}(\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2)_{1\le i,j\le 200}$ and $\mathbf{\Sigma}_{22} = \frac{1}{2}(\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2)_{1\le i,j\le 800}$ with the defintions of $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ similar to those in Section 4.1.1 except for different dimensions. Under such construction, the two largest eigenvalues of $\mathbf{\Sigma}_W$ are 201.75 and 77.61, respectively, while the others are less than 2.63. Based on the aforementioned covariance structures, for each data set, the rows of $\mathbf{X}$ and $\mathbf{W}$ 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 the 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 which 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

Table 3 Means and standard errors (in parentheses) of different performance measures by all methods over 200 simulations in Section 4.1.2; M_1: model with only observable predictors, M_2: model includes estimated latent variables, population error standard deviation $\sigma\sqrt{df/(df-2)}$ equals to 0.45

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)

via the 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.

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 (The notation 0.00 denotes a number less than 0.005.)

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)

4.2. Application to nutrient intake and human gut microbiome data

Nutrient intake strongly affects human health or diseases such as obesity, while 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 in 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 standard deviation 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 454/Roche 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 which 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 body mass index (BMI).

Due to 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 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 (PCs) 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 above 0.9 in M_1 or above 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, while the median model size also reduced from 93 to 69 after applying the NSL methodology. It shows that the prediction performance was improved after utilizing the information of 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 the model size mentioned before. 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 the diet coke, 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 coke can have a relatively healthy habitual diet and gut microbiome composition which in turn lower the BMI, but the diet coke itself does not reduce fats. Similar phenomena happened to acrylamide and vitamin E as well.

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; M_1: model with only micronutrients as predictors, M_2: model includes latent variables from gut microbiome composition

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)	PC(6th)	———	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)

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

We also applied the model-free knockoffs (Candès et al. 2018) with the target FDR level 0.2 on model M_2, and the most significant factors identified were the latent variables of 7th and 9th PCs, 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 PCs confirm the claim that firmicutes-enriched microbiome holds a greater metabolic potential for energy gain from the diet which results in the gain of weight (Turnbaugh et al. 2006). Furthermore, one of the major microbiome genera in the latent factor of 9th PC, Acidaminococcus, was also found to be positively associated with the BMI in Lin et al. (2014), which shows that human obesity can be affected at the genus level.

5. Discussions

In this paper, we have introduced a new methodology NSL for prediction and variable selection in the 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 false discovery rate control in nonsparse learning by the idea of NSL. Based on the established model identifiability condition which characterizes the correlations between observable and latent predictors, hypothesis testing can be proceeded using the de-biasing idea in Javanmard and Montanari (2014), van de Geer et al. (2014), Zhang and Zhang (2014). False discovery rate 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 the current paper and will be interesting topics for future research.

Acknowledgments

This work was supported by the National Natural Science Foundation of China-11601501, 11671018, 11671374, 71532001, 71731010, and 71921001, Anhui Provincial Natural Science Foundation-1708085QA02, Fundamental Research Funds for the Central Universities-WK2040160028, a grant from the Simons Foundation, Adobe Data Science Research Award, National Key R&D Program of China-2016YFC0207703, Beijing Natural Science Foundation-Z190001, and Beijing Academy of Artificial Intelligence (BAAI). The authors sincerely thank editors and referees for their valuable comments that helped improve the article substantially.

Biographies

Zemin Zheng is Professor in Department of Statistics and Finance of the School of Management at University of Science and Technology of China, Professor in the School of Data Science and International Institute of Finance at USTC. His research interests include statistics, machine learning, data science, and business applications. His research is supported by the National Natural Science Foundation of China-11601501, 11671374, 71731010, and 71921001, Anhui Provincial Natural Science Foundation-1708085QA02, and Fundamental Research Funds for the Central Universities-WK2040160028.

Jinchi Lv is Kenneth King Stonier Chair in Business Administration and Professor in Data Sciences and Operations Department of the Marshall School of Business at the University of Southern California, Professor in Department of Mathematics at USC, and an Associate Fellow of USC Dornsife Institute for New Economic Thinking (INET). He is the recipient of Fellow of Institute of Mathematical Statistics (2019), Adobe Data Science Research Award (2017), the Royal Statistical Society Guy Medal in Bronze (2015), NSF Faculty Early Career Development (CAREER) Award (2010).

Wei Lin is Assistant Professor in 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. His research is supported by the National Natural Science Foundation of China-11671018 and 71532001, National Key R&D Program of China-2016YFC0207703, Beijing Natural Science Foundation-Z190001, and Beijing Academy of Artificial Intelligence (BAAI).

Bibliography

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

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

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

Arumugam, M., Raes, J., Pelletier, E., et al. (2011). Enterotypes of the human gut microbiome. *Nature*, **473**, 174–180.

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

Barber, R. and Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *Ann. Statist.*, **43**, 2055–2085.

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

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

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

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.*, **37**, 1705–1732.

Bing, X., Bunea, F., Ning, Y., and Wegkamp, M. (2020). Sparse latent factor models with pure variables for overlapping clustering. *Ann. Statist.*, to appear.

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

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

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

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

Candès, E. J., Fan, Y., Janson, L., and Lv, J. (2018). Panning for gold: 'model-X' knockoffs for high-dimensional controlled variable selection. *J. Roy. Statist. Soc. Ser.* B, **80**, 551–577.

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

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

Chandrasekaran, V., Parrilo, P. A., and Willsky, A. S. (2012). Latent variable graphical model selection via convex optimization (with discussion). *Ann. Statist.*, **40**, 1935–1967.

Chen, J. and Li, H. (2013). Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. *Annals of Applied Statistics*, **7**, 418–442.

Chen, M., Ren, Z., Zhao, H., and Zhou, H. H. (2016). Asymptotically normal and efficient estimation of covariate-adjusted Gaussian graphical model. *J. Amer. Statist. Assoc.*, **111**, 394–406.

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

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

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

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

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

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

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

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

Gul, S., Hamilton, A., Munoz, A., et al. (2017). Inhibition of the gut enzyme intestinal alkaline phosphatase may explain how aspartame promotes glucose intolerance and obesity in mice. *Applied Physiology, Nutrition, and Metabolism*, **42**, 77–83.

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

Horn, R. A. and Johnson, C. R. (1990). *Matrix Analysis*. 2nd ed. Cambridge University Press.

Hsu, H. L., Ing, C. K., and Tong, H. (2019). On model selection from a finite family of possibly misspecified time series models. *Ann. Statist.*, **47**, 1061–1087.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, **15**, 2869–2909.

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

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

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

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

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

Ley, R. E., Turnbaugh, P. J., Klein, S., and Gordon, J. I. (2006). Human gut microbes associated with obesity. *Nature*, **444**, 1022–1023.

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

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

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

Lv, J. and Liu, J. S. (2014). Model selection principles in misspecified models. *J. Roy. Statist. Soc. Ser.* B, **76**, 141–167.

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

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

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

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

Paulson, C., Luo, L., and James, G. (2018). Efficient large-scale internet media selection optimization for online display advertising. *Journal of Marketing Research*, **55**, 489–506.

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

Pritchard, J. (2001). Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.*, **69**, 124–137.

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

Shen, D., Shen, H., and Marron, J. (2016). A general framework for consistency of principal component analysis. *Journal of Machine Learning Research*, **17**, 5218–5251.

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

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

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser.* B, **58**, 267–288.

Turnbaugh, P. J., Ley, R. E., Mahowald, M. A., Magrini, V., Mardis, E. R., and Gordon, J. I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, **444**, 1027–1031.

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

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

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

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

Xu, H., Caramanis, C., and Mannor, S. (2016). Statistical optimization in high dimensions. *Operations Research*, **64**, 958–979.

Wu, G. D., Chen, J., Hoffmann, C., et al. (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science*, **334**, 105–108.

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

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research*, **7**, 2541–2563.

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

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

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

E-companion to "Nonsparse Learning with Latent Variables"

This e-companion consists of two parts. Section EC.1 lists the key lemmas and presents the proofs for main results. Additional technical proofs for the lemmas are provided in Section EC.2.

EC.1. Proofs of main results

EC.1.1. Lemmas

The following lemmas are used in the proofs of main results.

LEMMA EC.1 **(Consistency of spiked sample eigenvalues)**. *Under Conditions 1 and 2, with asymptotic probability one, the eigenvalues of the sample covariance matrix* $\mathbf{S}$ *satisfy that for any* l, $1 \le l \le m$, *uniformly over* $i \in J_l$,

$$q^{-\alpha_l}\widehat{\lambda}_i \to c_i \ \text{ as } q \to \infty.$$

LEMMA EC.2. *Denote by* $\mathbf{X}_0$ *and* $\widehat{\mathbf{F}}_0$ *the submatrices of* $\mathbf{X}$ *and* $\widehat{\mathbf{F}}$ *consisting of columns in* $\mathrm{supp}(\boldsymbol{\beta}_0)$ *and* $\mathrm{supp}(\boldsymbol{\gamma}_0)$, *respectively, and* $\widetilde{\boldsymbol{\varepsilon}} = (\mathbf{F} - \widehat{\mathbf{F}})\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$. *For the following two events*

$$\widetilde{\mathcal{E}} = \left\{ \|n^{-1}(\mathbf{X}, \widehat{\mathbf{F}})^T \widetilde{\boldsymbol{\varepsilon}}\|_\infty \le c_2 \sqrt{(\log p)/n} \right\} \quad \text{and}$$
$$\widetilde{\mathcal{E}}_0 = \left\{ \|n^{-1}(\mathbf{X}_0, \widehat{\mathbf{F}}_0)^T \widetilde{\boldsymbol{\varepsilon}}\|_\infty \le c_2 \sqrt{(\log n)/n} \right\}$$

with constant $c_2 > 2\sqrt{2}\sigma$, *when the estimation error bound of* $\widehat{\mathbf{F}}$ *in Condition 3 holds and the columns of* $\mathbf{X}$ *adopt a common scale of* L_2*-norm* $n^{1/2}$, *we have*

$$P(\widetilde{\mathcal{E}} \cap \widetilde{\mathcal{E}}_0) \ge 1 - \frac{4\sqrt{2}\sigma}{c_2\sqrt{\pi \log p}} p^{1 - \frac{c_2^2}{8\sigma^2}} - \frac{2\sqrt{2}\sigma s}{c_2\sqrt{\pi \log n}} n^{-\frac{c_2^2}{8\sigma^2}},$$

which converges to one as $n \to \infty$.

EC.1.2. Proof of Theorem 1

Proof of part (a). In this part, we will focus on the convergence rates of the sample eigenvectors. The key ingredient of this proof is to link the angle between the sample eigenvector and the

space spanned by population eigenvectors with the sum of inner products between the sample and population eigenvectors by the $\cos(\cdot)$ function. In this way, it suffices to show that the sum of inner products converges to one for subspace consistency, and at the same time, deriving the convergence rates by induction. To ease readability, we will finish the proof in four steps.

Step 1: Analysis of the subspace consistency. We first show that for any $i \in J_l$, $1 \le l \le m$, the subspace consistency of the sample eigenvector $\widehat{\mathbf{u}}_i$ is equivalent to

$$\sum_{j \in J_l} p_{ji}^2 \to 1, \tag{EC.1}$$

where $p_{ji} = \mathbf{u}_j^T \widehat{\mathbf{u}}_i$ is the inner product between the population eigenvector $\mathbf{u}_j$ (the jth column of $\mathbf{U}$) and $\widehat{\mathbf{u}}_i$ (the ith column of $\widehat{\mathbf{U}}$).

Since $\mathbf{U}$ and $\widehat{\mathbf{U}}$ are obtained through eigen-decomposition, we know that $\|\mathbf{u}_j\|_2 = 1$ and $\|\widehat{\mathbf{u}}_i\|_2 = 1$ for any i and j, $1 \le i, j \le q$. Note that $\sum_{j \in J_l} (\mathbf{u}_j^T \widehat{\mathbf{u}}_i)\mathbf{u}_j$ is the projection of $\widehat{\mathbf{u}}_i$ on the space $span\{\mathbf{u}_j : j \in J_l\}$. It gives

$$Angle(\widehat{\mathbf{u}}_i, span\{\mathbf{u}_j : j \in J_l\}) = \arccos\Big\{ \frac{\widehat{\mathbf{u}}_i^T [\sum_{j \in J_l} (\mathbf{u}_j^T \widehat{\mathbf{u}}_i)\mathbf{u}_j]}{\|\widehat{\mathbf{u}}_i\|_2 \cdot \|\sum_{j \in J_l} (\mathbf{u}_j^T \widehat{\mathbf{u}}_i)\mathbf{u}_j\|_2} \Big\} =$$
$$\arccos\Big\{ \frac{\sum_{j \in J_l} (\mathbf{u}_j^T \widehat{\mathbf{u}}_i)^2}{[\sum_{j \in J_l} (\mathbf{u}_j^T \widehat{\mathbf{u}}_i)^2]^{1/2}} \Big\} = \arccos\Big\{ \sqrt{\sum_{j \in J_l} (\mathbf{u}_j^T \widehat{\mathbf{u}}_i)^2} \Big\} = \arccos\Big\{ \Big(\sum_{j \in J_l} p_{ji}^2 \Big)^{1/2} \Big\}.$$

Thus, $Angle(\widehat{\mathbf{u}}_i, span\{\mathbf{u}_j : j \in J_l\}) \to 0$ is equivalent to $\sum_{j \in J_l} p_{ji}^2 \to 1$ as $q \to \infty$ for any $i \in J_l$, $1 \le l \le m$. Moreover, the convergence rate of $\sum_{j \in J_l} p_{ji}^2$ indeed provides the convergence rate of the sample eigenvector $\widehat{\mathbf{u}}_i$ to the corresponding space of population eigenvectors.

We will then prove the convergence rates by induction. Hereafter our analysis will be conditional on the event $\mathcal{E}$, which is defined in the proof of Lemma EC.1 for the consistency of the spiked sample eigenvalues and enjoys asymptotic probability one.

Step 2: Convergence rates of sample eigenvectors with indices in J_1. This step aims at proving that uniformly over $i \in J_1$, the convergence rate of $\sum_{j \in J_1} p_{ji}^2$ is given by

$$\sum_{j \in J_1} p_{ji}^2 \ge 1 - O\{ \big(\sum_{l=2}^{m} k_l q^{\alpha_l} + k_{m+1} \big) K^{-1} q^{\alpha - \alpha_1} \} = 1 - O\{A(1)\}, \tag{EC.2}$$

where $A(t)=\big(\sum_{l=t+1}^{m}k_lq^{\alpha_l}+k_{m+1}\big)K^{-1}q^{\alpha-\alpha_t}$ is defined in Theorem 1. It is also the first part of induction. Let $\mathbf{P}=\mathbf{U}^T\widehat{\mathbf{U}}=\{p_{ij}\}_{1\le i,j\le q}$. We have $\sum_{j=1}^{q}p_{ji}^2=1$ for any i since $\mathbf{P}$ is a unitary matrix. To prove (EC.2), it suffices to show

$$\sum_{j\in J_2\cup\cdots\cup J_{m+1}}p_{ji}^2\le O\{A(1)\}.$$

Recall that $\mathbf{Z}=\mathbf{\Lambda}^{-1/2}\mathbf{U}^T\mathbf{W}^T$, $\mathbf{S}=n^{-1}\mathbf{W}^T\mathbf{W}=\widehat{\mathbf{U}}\widehat{\mathbf{\Lambda}}\widehat{\mathbf{U}}^T$. Therefore, we get a connection between $\mathbf{Z}$ and $\mathbf{P}$ that

$$n^{-1}\mathbf{Z}\mathbf{Z}^T=n^{-1}\mathbf{\Lambda}^{-1/2}\mathbf{U}^T\mathbf{W}^T\mathbf{W}\mathbf{U}\mathbf{\Lambda}^{-1/2}=\mathbf{\Lambda}^{-1/2}\mathbf{P}\widehat{\mathbf{\Lambda}}\mathbf{P}^T\mathbf{\Lambda}^{-1/2}.$$

For any j, $1\le j\le q$, in view of the (j,j)th entry, the above equality gives

$$\lambda_j^{-1}\sum_{i=1}^{q}\widehat{\lambda}_ip_{ji}^2=n^{-1}\mathbf{z}_j^T\mathbf{z}_j, \tag{EC.3}$$

where $\mathbf{z}_j$ is the jth column vector of $\mathbf{Z}^T$. It implies for any i, $1\le i\le q$, $\lambda_j^{-1}\widehat{\lambda}_ip_{ji}^2\le n^{-1}\mathbf{z}_j^T\mathbf{z}_j$. Based on this fact, we have

$$\sum_{j\in J_2\cup\cdots\cup J_{m+1}}p_{ji}^2\le\sum_{j\in J_2\cup\cdots\cup J_{m+1}}n^{-1}\mathbf{z}_j^T\mathbf{z}_j\lambda_j/\widehat{\lambda}_i=\sum_{t=1}^{n}\sum_{j\in J_2\cup\cdots\cup J_{m+1}}z_{jt}^2\lambda_j/(n\widehat{\lambda}_i), \tag{EC.4}$$

where z_{jt} is the (j,t)th entry of $\mathbf{Z}$. Conditional on the event $\mathcal{E}$, by Lemma EC.1, Conditions 1 and 2, we have

$$\begin{aligned}\sum_{t=1}^{n}\sum_{j\in J_2\cup\cdots\cup J_{m+1}}z_{jt}^2\lambda_j/(n\widehat{\lambda}_i)&\le\sum_{j\in J_2\cup\cdots\cup J_{m+1}}K^{-1}q^{\alpha}\lambda_j/\widehat{\lambda}_i\\&=O\{K^{-1}q^{\alpha}C\big(\sum_{l=2}^{m}k_lq^{\alpha_l}+k_{m+1}\big)/q^{\alpha_1}\}=O\{A(1)\}.\end{aligned} \tag{EC.5}$$

Since the convergences of $\widehat{\lambda}_i$ are uniform over $i\in J_1$ by Lemma EC.1, the above inequality holds uniformly over $i\in J_1$. Inequalities (EC.4) and (EC.5) together entail $\sum_{j\in J_2\cup\cdots\cup J_{m+1}}p_{ji}^2\le O\{A(1)\}$ uniformly over $i\in J_1$, which implies the convergence rate in (EC.2) for the sample eigenvectors with indices in J_1. It shows that when $s=1$, the convergence rate coincides with our claim that uniformly over $i\in J_l$, $1\le l\le s$,

$$\sum_{j\in J_l}p_{ji}^2\ge 1-\sum_{t=1}^{l-1}\big[\prod_{i=t+1}^{l-1}(1+k_i)\big]O\{k_tA(t)\}-O\{A(l)\}. \tag{EC.6}$$

Note that we define $\sum_{t=a}^{b} s_t = 0$ and $\prod_{t=a}^{b} s_t = 1$ if $b < a$ for any positive sequence $\{s_t\}$.

Step 3: Convergence rates of sample eigenvectors with indices in J_2. Before formally completing the proof by induction, we would like to derive the convergence rates of $\sum_{j\in J_2} p_{ji}^2$ directly for $i \in J_2$ to get the basic idea of induction.

Since we already proved the convergence rate in (EC.2) uniformly over $i \in J_1$ in **Step 2**, summing over $i \in J_1$ gives

$$\sum_{i\in J_1}\sum_{j\in J_1} p_{ji}^2 \geq k_1(1 - O\{A(1)\}) = k_1 - O\{k_1 A(1)\}. \tag{EC.7}$$

Along with the fact that $\sum_{i=1}^{q} p_{ji}^2 = 1$, we get

$$\begin{aligned}\sum_{i\in J_2\cup\cdots\cup J_{m+1}}\sum_{j\in J_1} p_{ji}^2 &= \sum_{i=1}^{q}\sum_{j\in J_1} p_{ji}^2 - \sum_{i\in J_1}\sum_{j\in J_1} p_{ji}^2 = \sum_{j\in J_1}\sum_{i=1}^{q} p_{ji}^2 - \sum_{i\in J_1}\sum_{j\in J_1} p_{ji}^2 \\ &= k_1 - \sum_{i\in J_1}\sum_{j\in J_1} p_{ji}^2 \leq k_1 - (k_1 - O\{k_1 A(1)\}) = O\{k_1 A(1)\}.\end{aligned} \tag{EC.8}$$

The above result is important as it also implies that uniformly over $i \in J_2$,

$$\sum_{j\in J_1} p_{ji}^2 \leq O\{k_1 A(1)\}. \tag{EC.9}$$

For the sample eigenvector $\widehat{u}_i$ with index $i \in J_2$, in order to find a lower bound for $\sum_{j\in J_2} p_{ji}^2$, we write it as

$$\sum_{j\in J_2} p_{ji}^2 = 1 \; - \sum_{j\in J_1} p_{ji}^2 \; - \sum_{j\in J_3\cup\cdots\cup J_{m+1}} p_{ji}^2. \tag{EC.10}$$

The upper bound of $\sum_{j\in J_1} p_{ji}^2$ was provided in (EC.9). For the second term $\sum_{j\in J_3\cup\cdots\cup J_{m+1}} p_{ji}^2$, similar to (EC.4) and (EC.5) in **Step 2**, by Lemma EC.1, Conditions 1 and 2, we have uniformly over $i \in J_2$,

$$\sum_{j\in J_3\cup\cdots\cup J_{m+1}} p_{ji}^2 \leq O\{K^{-1}q^{\alpha}C\big(\sum_{l=3}^{m} k_l q^{\alpha_l} + k_{m+1}\big)/q^{\alpha_2}\} = O\{A(2)\}.$$

Plugging the above two bounds into (EC.10) gives

$$\sum_{j\in J_2} p_{ji}^2 \geq 1 - O\{k_1 A(1)\} - O\{A(2)\},$$

which shows that the uniform convergence rate of the sample eigenvectors $\widehat{u}_i$ over $i \in J_2$. Together with the uniform convergence rate over $i \in J_1$ established in **Step 2**, our claim in (EC.6) gives the uniform convergence rates of the sample eigenvectors $\widehat{u}_i$ over $i \in J_1 \cup J_2$.

Step 4: Convergence rates of sample eigenvectors with indices in J_3 to J_m. In this step, we will complete the proof by induction. Specifically, we show that the claim in (EC.6) holds for any fixed s, $3 \le s \le m$, based on the induction assumption that the claim holds for $s-1$.

By the induction assumption, we have uniformly over $i \in J_l$, $1 \le l \le s-1$,

$$\sum_{j \in J_l} p_{ji}^2 \ge 1 - \sum_{t=1}^{l-1} \Big[\prod_{i=t+1}^{l-1} (1+k_i) \Big] O\{k_t A(t)\} - O\{A(l)\}.$$

By a similar argument as in (EC.7) and (EC.8), it follows that

$$\sum_{i \in J_{l+1} \cup \cdots \cup J_{m+1}} \sum_{j \in J_l} p_{ji}^2 \le k_l \Big(\sum_{t=1}^{l-1} \Big[\prod_{i=t+1}^{l-1} (1+k_i) \Big] O\{k_t A(t)\} + O\{A(l)\} \Big). \tag{EC.11}$$

Similarly as in **Step 3**, for any $i \in J_s$, to get the convergence rate of $\sum_{j \in J_s} p_{ji}^2$, we write it as

$$\sum_{j \in J_s} p_{ji}^2 = 1 \quad - \sum_{j \in J_1 \cup \cdots \cup J_{s-1}} p_{ji}^2 \quad - \sum_{j \in J_{s+1} \cup \cdots \cup J_{m+1}} p_{ji}^2.$$

We will first derive the convergence rate of $\sum_{j \in J_1 \cup \cdots \cup J_{s-1}} p_{ji}^2$. When $1 \le l \le s-1$, we have $i \in J_s \subset J_{l+1} \cup \cdots \cup J_{m+1}$. In view of (EC.11), it gives that uniformly over $i \in J_s$,

$$\sum_{j \in J_l} p_{ji}^2 \le k_l \Big(\sum_{t=1}^{l-1} \Big[\prod_{i=t+1}^{l-1} (1+k_i) \Big] O\{k_t A(t)\} + O\{A(l)\} \Big).$$

Summing over $l = 1, \ldots, s-1$, we get

$$\sum_{j \in J_1 \cup \cdots \cup J_{s-1}} p_{ji}^2 \le \sum_{l=1}^{s-1} k_l \Big(\sum_{t=1}^{l-1} \Big[\prod_{i=t+1}^{l-1} (1+k_i) \Big] O\{k_t A(t)\} + O\{A(l)\} \Big).$$

To simplify the above expression, exchanging the summation order with respect to l and t gives

$$\begin{aligned} \sum_{j \in J_1 \cup \cdots \cup J_{s-1}} p_{ji}^2 &\le \sum_{l=1}^{s-1} \sum_{t=1}^{l-1} k_l \Big[\prod_{i=t+1}^{l-1} (1+k_i) \Big] O\{k_t A(t)\} + \sum_{l=1}^{s-1} O\{k_l A(l)\} \\ &= \sum_{t=1}^{s-2} \sum_{l=t+1}^{s-1} k_l \Big[\prod_{i=t+1}^{l-1} (1+k_i) \Big] O\{k_t A(t)\} + \sum_{t=1}^{s-1} O\{k_t A(t)\}. \end{aligned}$$

Then we combine the coefficients of $k_t A(t)$ to get

$$\sum_{j\in J_1\cup\cdots\cup J_{s-1}} p_{ji}^2 \le \sum_{t=1}^{s-2} O\{k_t A(t)\}\big(1+\sum_{l=t+1}^{s-1} k_l\big[\prod_{i=t+1}^{l-1}(1+k_i)\big]\big)+\sum_{t=s-1} O\{k_t A(t)\}.$$

Since it is immediate to conclude by induction that

$$\begin{aligned}1+\sum_{l=t+1}^{s-1} k_l\big[\prod_{i=t+1}^{l-1}(1+k_i)\big] &= 1+k_{t+1}+k_{t+2}(1+k_{t+1})+\cdots\\ &+k_{s-1}(1+k_{s-2})(1+k_{s-3})\cdots(1+k_{t+2})(1+k_{t+1})=\prod_{i=t+1}^{s-1}(1+k_i),\end{aligned}$$

we then have

$$\begin{aligned}\sum_{j\in J_1\cup\cdots\cup J_{s-1}} p_{ji}^2 &\le \sum_{t=1}^{s-2}\big[\prod_{i=t+1}^{s-1}(1+k_i)\big]O\{k_t A(t)\}+\sum_{t=s-1} O\{k_t A(t)\}\\ &=\sum_{t=1}^{s-1}\big[\prod_{i=t+1}^{s-1}(1+k_i)\big]O\{k_t A(t)\}.\end{aligned}$$

On the other hand, similar to (EC.4) and (EC.5) in **Step 2**, we have uniformly over $i\in J_s$,

$$\sum_{j\in J_{s+1}\cup\cdots\cup J_{m+1}} p_{ji}^2 \le O\{A(s)\}.$$

Combining the above two bounds gives the convergence rate of $\sum_{j\in J_s} p_{ji}^2$ uniformly over $i\in J_s$ as

$$\sum_{j\in J_s} p_{ji}^2 \ge 1-\sum_{t=1}^{s-1}\big[\prod_{i=t+1}^{s-1}(1+k_i)\big]O\{k_t A(t)\}-O\{A(s)\}.$$

Together with the induction assumption that our claim in (EC.6) holds uniformly over $i\in J_l$, $1\le l\le s-1$, we know that the claim also holds uniformly over $i\in J_l$, $1\le l\le s$. Therefore, by induction, the results in part (a) of Theorem 1 hold uniformly over $i\in J_l$, $1\le l\le m$.

Proof of part (b). In this part, we will show that when each group of spiked eigenvalues has size one (that is, $k_l=1$ for any l, $1\le l\le m$), the convergence rates of the angles between the sample score vectors $\mathbf{W}\widehat{\mathbf{u}}_i$ and the population score vectors $\mathbf{W}\mathbf{u}_i$, $1\le i\le K$, are at least as fast as those of the angles between the corresponding sample and population eigenvectors established in part (a) of Theorem 1. The key idea is to conduct delicate analysis on the $\cos(\cdot)$ function of the

angles between the sample score vectors and population score vectors, where some results about the sample eigenvalues derived in the proof of Lemma EC.1 will be used.

When each group has size one, we have $K = m$ and the convergence rates of $\widehat{\mathbf{u}}_i$ $(i \in J_l)$ to the space $span\{\mathbf{u}_j : j \in J_l\}$ become the convergence rates of $\widehat{\mathbf{u}}_i$ to $\mathbf{u}_i$, $1 \le i \le K$. Denote by $\theta_{ii} = Angle(\widehat{\mathbf{u}}_i, \mathbf{u}_i)$ and $\omega_{ii} = Angle(\mathbf{W}\widehat{\mathbf{u}}_i, \mathbf{W}\mathbf{u}_i)$. Then the results in part (a) give that uniformly over $1 \le i \le K$,

$$\cos^2(\theta_{ii}) = p_{ii}^2 \ge 1 - \sum_{t=1}^{i-1} 2^{i-t-1} O\{A(t)\} - O\{A(i)\}. \tag{EC.12}$$

Since $\mathbf{S} = n^{-1}\mathbf{W}^T\mathbf{W} = \widehat{\mathbf{U}}\widehat{\mathbf{\Lambda}}\widehat{\mathbf{U}}^T$, $\widehat{\mathbf{u}}_i$ would be the eigenvector of $\mathbf{W}^T\mathbf{W}$ corresponding to the eigenvalue $n\widehat{\lambda}_i$ with L_2-norm 1. It follows that

$$\cos(\omega_{ii}) = \frac{(\mathbf{W}\mathbf{u}_i)^T\mathbf{W}\widehat{\mathbf{u}}_i}{\|\mathbf{W}\mathbf{u}_i\|_2\|\mathbf{W}\widehat{\mathbf{u}}_i\|_2} = \frac{n\widehat{\lambda}_i\mathbf{u}_i^T\widehat{\mathbf{u}}_i}{\sqrt{n\widehat{\lambda}_i}\|\mathbf{W}\mathbf{u}_i\|_2} = \frac{\sqrt{n\widehat{\lambda}_i}\cos(\theta_{ii})}{\|\mathbf{W}\mathbf{u}_i\|_2}.$$

Squaring both sides above gives

$$\cos^2(\omega_{ii}) = \frac{n\widehat{\lambda}_i\cos^2(\theta_{ii})}{\|\mathbf{W}\mathbf{u}_i\|_2^2}. \tag{EC.13}$$

Therefore, it suffices to show $\|\mathbf{W}\mathbf{u}_i\|_2^2 \le n\widehat{\lambda}_i$.

For the term $\|\mathbf{W}\mathbf{u}_i\|_2^2$, it follows from $\mathbf{W}^T\mathbf{W} = n\widehat{\mathbf{U}}\widehat{\mathbf{\Lambda}}\widehat{\mathbf{U}}^T$ that

$$\|\mathbf{W}\mathbf{u}_i\|_2^2 = \mathbf{u}_i^T\mathbf{W}^T\mathbf{W}\mathbf{u}_i = n\mathbf{u}_i^T\widehat{\mathbf{U}}\widehat{\mathbf{\Lambda}}\widehat{\mathbf{U}}^T\mathbf{u}_i = n\sum_{j=1}^{q}\widehat{\lambda}_j(\mathbf{u}_i^T\widehat{\mathbf{u}}_j)^2 = n\sum_{j=1}^{q}\widehat{\lambda}_j p_{ij}^2.$$

By further making use of equality (EC.3), we have

$$\|\mathbf{W}\mathbf{u}_i\|_2^2 = n\sum_{j=1}^{q}\widehat{\lambda}_j p_{ij}^2 = \lambda_i\mathbf{z}_i^T\mathbf{z}_i,$$

where $\mathbf{z}_i$ is the ith column vector of $\mathbf{Z}^T$. On the other hand, inequality (EC.33) in the proof of Lemma EC.1 gives a lower bound for the sample eigenvalues $\widehat{\lambda}_i$, $1 \le i \le K$. Under the current setting that each group has size one, it gives

$$\widehat{\lambda}_i \ge \varphi_1(n^{-1}\lambda_i\mathbf{z}_i\mathbf{z}_i^T) = \varphi_1(n^{-1}\lambda_i\mathbf{z}_i^T\mathbf{z}_i) = n^{-1}\lambda_i\mathbf{z}_i^T\mathbf{z}_i,$$

where $\varphi_1(\cdot)$ denotes the largest eigenvalue of a given matrix. It follows that

$$n\widehat{\lambda}_i \ge \lambda_i\mathbf{z}_i^T\mathbf{z}_i = \|\mathbf{W}\mathbf{u}_i\|_2^2.$$

Therefore, in view of (EC.13), we get

$$\cos^2(\omega_{ii}) \geq \cos^2(\theta_{ii}),$$

which means that the convergence rate of the sample score vector is at least as good as that of the corresponding sample eigenvector. Then it follows from (EC.12) that uniformly over $1 \leq i \leq K$,

$$\cos^2(\omega_{ii}) \geq 1 - \sum_{t=1}^{i-1} 2^{i-t-1} O\{A(t)\} - O\{A(i)\},$$

which completes the proof of part (b) of Theorem 1.

EC.1.3. Proof of Proposition 1

By Condition 4, the inequality $\|n^{-1/2}(\mathbf{X}, \mathbf{F})\boldsymbol{\delta}\|_2 \geq c\|\boldsymbol{\delta}\|_2$ holds for any $\boldsymbol{\delta}$ satisfying $\|\boldsymbol{\delta}\|_0 < M$ with significant probability $1 - \theta_{n,p}$. We now derive a similar result for $(\mathbf{X}, \widehat{\mathbf{F}})$ by analyzing the estimation errors of confounding factors $\mathbf{F}$.

By the estimation error bound in Condition 3, we have for any $1 \leq j \leq K$,

$$\begin{aligned}\|\mathbf{f}_j - \widehat{\mathbf{f}}_j\|_2^2 &\leq \|\mathbf{f}_j\|_2^2 + \|\widehat{\mathbf{f}}_j\|_2^2 - 2\ \mathbf{f}_j'\widehat{\mathbf{f}}_j = n + n - 2\|\mathbf{f}_j\|_2\|\widehat{\mathbf{f}}_j\|_2 \cos(\omega_{jj}) \\ &= 2n - 2n\cos(\omega_{jj}) = 2n\{1 - \cos(\omega_{jj})\} \leq \frac{c_2^2 \log n}{4K^2\rho^2}.\end{aligned}$$

Since the above bound does not vary with the index j, it gives the uniform confounding factor estimation error bound

$$\max_{1\leq j\leq K} \|\mathbf{f}_j - \widehat{\mathbf{f}}_j\|_2 \leq \frac{c_2}{2K\rho}\sqrt{\log n}. \tag{EC.14}$$

Now we proceed to prove the inequality for $(\mathbf{X}, \widehat{\mathbf{F}})$. First of all, it follows from Condition 4 and the triangular inequality that

$$\begin{aligned}\|n^{-1/2}(\mathbf{X}, \widehat{\mathbf{F}})\boldsymbol{\delta}\|_2 &\geq \|n^{-1/2}(\mathbf{X}, \mathbf{F})\boldsymbol{\delta}\|_2 - \|n^{-1/2}(\mathbf{X}, \mathbf{F})\boldsymbol{\delta} - n^{-1/2}(\mathbf{X}, \widehat{\mathbf{F}})\boldsymbol{\delta}\|_2 \\ &\geq c\|\boldsymbol{\delta}\|_2 - n^{-1/2}\|(\mathbf{F} - \widehat{\mathbf{F}})\boldsymbol{\delta}_1\|_2 \geq c\|\boldsymbol{\delta}\|_2 - n^{-1/2} \max_{1\leq j\leq K} \|(\mathbf{f}_j - \widehat{\mathbf{f}}_j)\|_2 \|\boldsymbol{\delta}_1\|_1,\end{aligned}$$

where $\boldsymbol{\delta}_1$ is a subvector of $\boldsymbol{\delta}$ consisting of the last K components. Note that $\|\boldsymbol{\delta}_1\|_1 \leq \sqrt{K}\|\boldsymbol{\delta}_1\|_2 \leq \sqrt{K}\|\boldsymbol{\delta}\|_2$. Further applying inequality (EC.14) yields

$$\|n^{-1/2}(\mathbf{X}, \widehat{\mathbf{F}})\boldsymbol{\delta}\|_2 \geq c\|\boldsymbol{\delta}\|_2 - n^{-1/2} \cdot \frac{c_2}{2K\rho}\sqrt{\log n} \cdot \sqrt{K}\|\boldsymbol{\delta}\|_2 \geq c_1\|\boldsymbol{\delta}\|_2,$$

where c_1 is some positive constant no larger than $c - \frac{c_2}{2\rho}\sqrt{\frac{\log n}{nK}}$. It is clear that c_1 is smaller than but close to c when n is relatively large. In view of the tail probabilities in Conditions 3 and 4, the above inequality holds with probability at least $1-\theta_1-\theta_2$. Thus, we finish the proof of Proposition 1.

EC.1.4. Proof of Theorem 2

With Proposition 1, we will apply a similar idea as in Zheng et al. (2014) to prove the global properties. The proof consists of two parts. The first part shows the model selection consistency property with the range of λ given in Theorem 2. Based on the first part, several oracle inequalities will then be induced. We will first prove the properties when the columns of design matrix $\mathbf{X}$ have a common scale of L_2-norm $n^{1/2}$ as a benchmark, meaning that $\boldsymbol{\beta}_* = \boldsymbol{\beta}$ and $L = 1$, and then illustrate the results in general cases.

Part 1: Model selection consistency. This part contains two steps. In the first step, it will be shown that when $c_1^{-1}c_2\sqrt{(2s+1)(\log p)/n} < \lambda < b_0$, the number of nonzero elements in $(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T$ is no larger than s conditioning on the event $\widetilde{\mathcal{E}}$ defined in Lemma EC.2. We prove this by using the global optimality of $(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T$.

By the hard-thresholding property (Zheng et al. 2014, Lemma 1) and $\lambda < b_0$, any nonzero component of the true regression coefficient vector $(\boldsymbol{\beta}_0^T, \boldsymbol{\gamma}_0^T)^T$ or of the global minimizer $(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T$ is greater than λ, which ensures $\|p_\lambda\{(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T\}\|_1 = \lambda^2\|(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T\|_0/2$ and $\|p_\lambda\{(\boldsymbol{\beta}_0^T, \boldsymbol{\gamma}_0^T)^T\}\|_1$ $= s\lambda^2/2$. Thus, we get

$$\left\|p_\lambda\left\{(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T\right\}\right\|_1 - \left\|p_\lambda\left\{(\boldsymbol{\beta}_0^T, \boldsymbol{\gamma}_0^T)^T\right\}\right\|_1 = \left\{\|(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T\|_0 - s\right\}\lambda^2/2.$$

Denote by $\boldsymbol{\delta} = (\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T - (\boldsymbol{\beta}_0^T, \boldsymbol{\gamma}_0^T)^T$. Direct calculation yields

$$\begin{aligned} Q\left\{(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T\right\} - Q\left\{(\boldsymbol{\beta}_0^T, \boldsymbol{\gamma}_0^T)^T\right\} &= 2^{-1}\left\|n^{-\frac{1}{2}}(\mathbf{X}, \widehat{\mathbf{F}})\boldsymbol{\delta}\right\|_2^2 - n^{-1}\widetilde{\boldsymbol{\varepsilon}}^T(\mathbf{X}, \widehat{\mathbf{F}})\boldsymbol{\delta} \\ &\quad + \left\{\|(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T\|_0 - s\right\}\lambda^2/2, \end{aligned} \tag{EC.15}$$

where $\widetilde{\boldsymbol{\varepsilon}} = (\mathbf{F} - \widehat{\mathbf{F}})\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$, the sum of the random error vector $\boldsymbol{\varepsilon}$ and estimation errors $(\mathbf{F} - \widehat{\mathbf{F}})\boldsymbol{\gamma}$.

On the other hand, conditional on event $\widetilde{\mathcal{E}}$, we have

$$\begin{aligned} |n^{-1}\widetilde{\boldsymbol{\varepsilon}}^T(\mathbf{X}, \widehat{\mathbf{F}})\boldsymbol{\delta}| \;&\le\; \|n^{-1}\widetilde{\boldsymbol{\varepsilon}}^T(\mathbf{X}, \widehat{\mathbf{F}})\|_\infty \|\boldsymbol{\delta}\|_1 \\ &\le\; c_2\sqrt{(\log p)/n}\|\boldsymbol{\delta}\|_1 \;\le\; c_2\sqrt{(\log p)/n}\|\boldsymbol{\delta}\|_0^{\frac{1}{2}}\|\boldsymbol{\delta}\|_2. \end{aligned} \tag{EC.16}$$

In addition, by Condition 6 and the definition of $\mathbb{S}_{M/2}$, we obtain $\|\boldsymbol{\delta}\|_0 \le \|(\boldsymbol{\beta}_0^T, \boldsymbol{\gamma}_0^T)^T\|_0 + \|(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T\|_0 < M$, where M is the robust spark of $(\mathbf{X}, \widehat{\mathbf{F}})$ with bound c_1 by Proposition 1. Thus, we have

$$\|n^{-\frac{1}{2}}(\mathbf{X}, \widehat{\mathbf{F}})\boldsymbol{\delta}\|_2 \ge c_1\|\boldsymbol{\delta}\|_2. \tag{EC.17}$$

Plugging inequalities (EC.16) and (EC.17) into (EC.15) gives that

$$\begin{aligned} Q\left\{(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T\right\} - Q\left\{(\boldsymbol{\beta}_0^T, \boldsymbol{\gamma}_0^T)^T\right\} \ge 2^{-1}c_1^2\|\boldsymbol{\delta}\|_2^2 - \; & c_2\sqrt{(\log p)/n}\|\boldsymbol{\delta}\|_0^{\frac{1}{2}}\|\boldsymbol{\delta}\|_2 \\ & + \left\{\|(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T\|_0 - s\right\}\lambda^2/2. \end{aligned} \tag{EC.18}$$

Thus, the global optimality of $(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T$ ensures that

$$2^{-1}c_1^2\|\boldsymbol{\delta}\|_2^2 - c_2\sqrt{\frac{\log p}{n}}\|\boldsymbol{\delta}\|_0^{\frac{1}{2}}\|\boldsymbol{\delta}\|_2 + \left\{\|(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T\|_0 - s\right\}\lambda^2/2 \le 0.$$

After completing the squares in the above inequality, we get

$$\left[c_1\|\boldsymbol{\delta}\|_2 - \frac{c_2}{c_1}\sqrt{\frac{\log p}{n}}\|\boldsymbol{\delta}\|_0^{\frac{1}{2}}\right]^2 - \left(\frac{c_2}{c_1}\right)^2\frac{\log p}{n}\|\boldsymbol{\delta}\|_0 + \left\{\|(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T\|_0 - s\right\}\lambda^2 \le 0.$$

Since $\left[c_1\|\boldsymbol{\delta}\|_2 - \frac{c_2}{c_1}\sqrt{\frac{\log p}{n}}\|\boldsymbol{\delta}\|_0^{\frac{1}{2}}\right]^2 \ge 0$, it gives

$$\left\{\|(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T\|_0 - s\right\}\lambda^2 \le \left(\frac{c_2}{c_1}\right)^2\frac{\log p}{n}\|\boldsymbol{\delta}\|_0. \tag{EC.19}$$

We continue to bound the value of $\|(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T\|_0$ by the above inequality. Let $k = \|(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T\|_0$. Then $\|\boldsymbol{\delta}\|_0 = \|(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T - (\boldsymbol{\beta}_0^T, \boldsymbol{\gamma}_0^T)^T\|_0 \le k + s$. Thus, it follows from (EC.19) that

$$(k - s)\lambda^2 \le \left(\frac{c_2}{c_1}\right)^2\frac{\log p}{n}(k + s).$$

Organizing it in terms of k and s, we get

$$k\left(\lambda^2-\left(\frac{c_2}{c_1}\right)^2\frac{\log p}{n}\right)\le s\left(\lambda^2+\left(\frac{c_2}{c_1}\right)^2\frac{\log p}{n}\right). \tag{EC.20}$$

Since $\lambda > c_1^{-1}c_2\sqrt{(2s+1)\log p/n}$, we have $\lambda^2-(c_1^{-1}c_2)^2(2s+1)\frac{\log p}{n}>0$ and $\lambda^2c_1^2n-c_2^2\log p > 2c_2^2 s\log p$. Thus we have $\frac{2c_2^2\log p}{\lambda^2c_1^2n-c_2^2\log p}<1/s$. Then it follows from inequality (EC.20) that

$$k\le s\frac{(\lambda^2+(\frac{c_2}{c_1})^2\frac{\log p}{n})}{(\lambda^2-(\frac{c_2}{c_1})^2\frac{\log p}{n})}=s\left(1+\frac{2c_2^2\log p}{\lambda^2c_1^2n-c_2^2\log p}\right)<s+1.$$

Therefore, the number of nonzero elements in $(\widehat{\boldsymbol{\beta}}^T,\widehat{\boldsymbol{\gamma}}^T)^T$ satisfies

$$\left\|(\widehat{\boldsymbol{\beta}}^T,\widehat{\boldsymbol{\gamma}}^T)^T\right\|_0\le s.$$

The second step is based on the first step, where we will use proof by contradiction to show that $\mathrm{supp}((\boldsymbol{\beta}_0^T,\boldsymbol{\gamma}_0^T)^T)\subset\mathrm{supp}((\widehat{\boldsymbol{\beta}}^T,\widehat{\boldsymbol{\gamma}}^T)^T)$ with the additional assumption $\lambda<b_0c_1/\sqrt{2}$ in the theorem. Suppose that $\mathrm{supp}((\boldsymbol{\beta}_0^T,\boldsymbol{\gamma}_0^T)^T)\not\subset\mathrm{supp}(\widehat{\boldsymbol{\beta}}^T,\widehat{\boldsymbol{\gamma}}^T)^T$, and we denote the number of missed true coefficients as

$$k=\left|\mathrm{supp}\{(\boldsymbol{\beta}_0^T,\boldsymbol{\gamma}_0^T)^T\}\backslash\mathrm{supp}\{(\widehat{\boldsymbol{\beta}}^T,\widehat{\boldsymbol{\gamma}}^T)^T\}\right|\ge 1.$$

Then we have $\|(\widehat{\boldsymbol{\beta}}^T,\widehat{\boldsymbol{\gamma}}^T)^T\|_0\ge s-k$ and $\|\boldsymbol{\delta}\|_0\le\|(\widehat{\boldsymbol{\beta}}^T,\widehat{\boldsymbol{\gamma}}^T)^T\|_0+\|(\boldsymbol{\beta}_0^T,\boldsymbol{\gamma}_0^T)^T\|_0\le 2s$ by the first step. Combining these two results with inequality (EC.18) yields

$$Q\left\{(\widehat{\boldsymbol{\beta}}^T,\widehat{\boldsymbol{\gamma}}^T)^T\right\}-Q\left\{(\boldsymbol{\beta}_0^T,\boldsymbol{\gamma}_0^T)^T\right\}\ge\left(2^{-1}c_1^2\|\boldsymbol{\delta}\|_2-c_2\sqrt{\frac{2s\log p}{n}}\right)\|\boldsymbol{\delta}\|_2-k\lambda^2/2. \tag{EC.21}$$

Note that for each $j\in\mathrm{supp}((\boldsymbol{\beta}_0^T,\boldsymbol{\gamma}_0^T)^T)\setminus\mathrm{supp}((\widehat{\boldsymbol{\beta}}^T,\widehat{\boldsymbol{\gamma}}^T)^T)$, we have $|\delta_j|\ge b_0$ with b_0 the lowest signal strength defined in Condition 6. Thus, $\|\boldsymbol{\delta}\|_2\ge\sqrt{k}b_0$, which together with Condition 6 entails

$$4^{-1}c_1^2\|\boldsymbol{\delta}\|_2\ge 4^{-1}c_1^2\sqrt{k}b_0\ge 4^{-1}c_1^2b_0>c_2\sqrt{(2s\log p)/n}.$$

Thus, it follows from (EC.21) that

$$Q\left\{(\widehat{\boldsymbol{\beta}}^T,\widehat{\boldsymbol{\gamma}}^T)^T\right\}-Q\left\{(\boldsymbol{\beta}_0^T,\boldsymbol{\gamma}_0^T)^T\right\}\ge 4^{-1}c_1^2\|\boldsymbol{\delta}\|_2^2-k\lambda^2/2\ge 4^{-1}c_1^2kb_0^2-k\lambda^2/2>0,$$

where the last step is because of the additional assumption $\lambda < b_0 c_1/\sqrt{2}$. The above inequality contradicts with the global optimality of $(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T$. Thus, we have $\mathrm{supp}((\boldsymbol{\beta}_0^T, \boldsymbol{\gamma}_0^T)^T) \subset \mathrm{supp}((\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T)$. Combining this with $\|(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T\|_0 \le s$ from the first step, we know that $\mathrm{supp}\{(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T\} = \mathrm{supp}\{(\boldsymbol{\beta}_0^T, \boldsymbol{\gamma}_0^T)^T\}$.

Part 2: Prediction and estimation losses. In this part, we will bound the prediction and estimation losses. The idea is to get the L_2-estimation loss bound by the global optimality of $(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T$, conditional on the event $\widetilde{\mathcal{E}} \cap \widetilde{\mathcal{E}}_0$ defined in Lemma EC.2. Then by similar techniques as in the first part, we would derive bounds for the prediction and estimation losses.

Recall that $\mathbf{X}_0$, $\widehat{\mathbf{F}}_0$ are the submatrices of $\mathbf{X}$ and $\widehat{\mathbf{F}}$ consisting of columns in $\mathrm{supp}(\boldsymbol{\beta}_0)$ and $\mathrm{supp}(\boldsymbol{\gamma}_0)$, respectively. Conditioning on $\widetilde{\mathcal{E}} \cap \widetilde{\mathcal{E}}_0$, we have $\|\boldsymbol{\delta}\|_0 \le s$ by the model selection consistency established before. Thus, applying the Cauchy-Schwarz inequality and the definition of $\widetilde{\mathcal{E}}_0$ gives

$$\begin{aligned} |n^{-1}\widetilde{\boldsymbol{\varepsilon}}^T(\mathbf{X}_0, \widehat{\mathbf{F}}_0)\boldsymbol{\delta}| \;&\le\; \|n^{-1}\widetilde{\boldsymbol{\varepsilon}}^T(\mathbf{X}_0, \widehat{\mathbf{F}}_0)\|_\infty \|\boldsymbol{\delta}\|_1 \\ &\le\; c_2\sqrt{\frac{\log n}{n}}\|\boldsymbol{\delta}\|_1 \;\le\; c_2\sqrt{\frac{s\log n}{n}}\|\boldsymbol{\delta}\|_2. \end{aligned} \tag{EC.22}$$

In views of (EC.15) and (EC.17), it follows from inequality (EC.22) and the model selection consistency property $\|(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T\|_0 = s$ that

$$\begin{aligned} &Q\left\{(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T\right\} - Q\left\{(\boldsymbol{\beta}_0^T, \boldsymbol{\gamma}_0^T)^T\right\} \\ &= 2^{-1}\|n^{-1}(\mathbf{X}, \widehat{\mathbf{F}})\boldsymbol{\delta}\|_2^2 - n^{-1}\widetilde{\boldsymbol{\varepsilon}}^T(\mathbf{X}, \widehat{\mathbf{F}})\boldsymbol{\delta} + \left\{\|(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T\|_0 - s\right\}\lambda^2/2 \\ &\ge 2^{-1}c_1^2\|\boldsymbol{\delta}\|_2^2 - n^{-1}\widetilde{\boldsymbol{\varepsilon}}^T(\mathbf{X}_0, \widehat{\mathbf{F}}_0)\boldsymbol{\delta} \ge \left(2^{-1}c_1^2\|\boldsymbol{\delta}\|_2 - c_2\sqrt{\frac{s\log n}{n}}\right)\|\boldsymbol{\delta}\|_2. \end{aligned}$$

Since $(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T$ is the global optimizer of Q, we have

$$2^{-1}c_1^2\|\boldsymbol{\delta}\|_2 - c_2\sqrt{\frac{s\log n}{n}} \le 0,$$

which gives the L_2 and L_∞ estimation loss bounds as

$$\begin{aligned} &\|(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T - (\boldsymbol{\beta}_0^T, \boldsymbol{\gamma}_0^T)^T\|_2 = \|\boldsymbol{\delta}\|_2 \le\; 2c_1^{-2}c_2\sqrt{(s\log n)/n}, \\ &\|(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T - (\boldsymbol{\beta}_0^T, \boldsymbol{\gamma}_0^T)^T\|_\infty \le \|(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T - (\boldsymbol{\beta}_0^T, \boldsymbol{\gamma}_0^T)^T\|_2 \le 2c_1^{-2}c_2\sqrt{(s\log n)/n}. \end{aligned}$$

For L_q-estimation losses with $1 \le q < 2$, applying Hölder's inequality gives

$$\begin{aligned}\|(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T - (\boldsymbol{\beta}_0^T, \boldsymbol{\gamma}_0^T)^T\|_q &= \Big(\sum_j |\delta_j|^q\Big)^{1/q} \le \Big(\sum_j |\delta_j|^2\Big)^{\frac{1}{2}} \Big(\sum_{\delta_j \ne 0} 1^{\frac{2}{2-q}}\Big)^{\frac{1}{q}-\frac{1}{2}} \\ &= \|\boldsymbol{\delta}\|_2 \|\boldsymbol{\delta}\|_0^{\frac{1}{q}-\frac{1}{2}} \le 2c_1^{-2} c_2 s^{\frac{1}{q}} \sqrt{(\log n)/n}.\end{aligned}$$

Next we prove the bound for oracle prediction loss. Since $(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T$ is the global minimizer, it follows from (EC.15) and the model selection consistency property that

$$\begin{aligned}&n^{-1/2} \|(\mathbf{X}, \widehat{\mathbf{F}})\{(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T - (\boldsymbol{\beta}_0^T, \boldsymbol{\gamma}_0^T)^T\}\|_2 \\ \le &\Big\{2n^{-1} \widetilde{\boldsymbol{\varepsilon}}^T (\mathbf{X}, \widehat{\mathbf{F}}) \boldsymbol{\delta}\Big\}^{1/2} \le \Big\{2\|n^{-1}(\mathbf{X}_0, \widehat{\mathbf{F}}_0)^T \tilde{\boldsymbol{\varepsilon}}\|_\infty \|\boldsymbol{\delta}\|_1\Big\}^{1/2} \le 2c_2 c_1^{-1} \sqrt{s(\log n)/n},\end{aligned}$$

where the last step is because of the L_1 estimation loss bound proved before. Then for the oracle prediction loss, together with (EC.34) in the proof of Lemma EC.2, it follows that

$$\begin{aligned}&n^{-1/2} \|(\mathbf{X}, \widehat{\mathbf{F}})(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T - (\mathbf{X}, \mathbf{F})(\boldsymbol{\beta}_0^T, \boldsymbol{\gamma}_0^T)^T\|_2 \\ \le\ & 2c_2 c_1^{-1} \sqrt{s(\log n)/n} + n^{-1/2} \|(\mathbf{F} - \widehat{\mathbf{F}})\boldsymbol{\gamma}_0\|_2 \le (2c_2 c_1^{-1} \sqrt{s} + c_2/2) \sqrt{(\log n)/n}.\end{aligned}$$

Last we will derive our results for general cases when the L_2-norms of columns of $\mathbf{X}$ are not of the common scale $n^{1/2}$. Note that the penalized least squares in (3) can be rewritten as

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

where $\widetilde{\mathbf{X}}$ is the matrix with the L_2-norm of each column rescaled to $n^{1/2}$ and

$$\boldsymbol{\beta}_* = n^{-1/2}(\beta_1 \|\mathbf{x}_1\|_2, \ldots, \beta_p \|\mathbf{x}_p\|_2)^T$$

is the corresponding coefficient vector defined in (3). By Conditions 5 and 6, the same argument applies to derive the model selection consistency property and the bounds on oracle prediction and estimation losses for $(\widehat{\boldsymbol{\beta}}_*^T, \widehat{\boldsymbol{\gamma}}^T)^T$ since the relationship between λ and signal strength keeps the same even if $L \ne 1$. Based on Condition 5, it is clear that the model selection consistency of $\widehat{\boldsymbol{\beta}}_*$ implies

that of $\widehat{\boldsymbol{\beta}}$. And the bound on prediction loss does not change since $\widetilde{\mathbf{X}}\widehat{\boldsymbol{\beta}}_* = \mathbf{X}\widehat{\boldsymbol{\beta}}$. As for the bounds of estimation losses on $\widehat{\boldsymbol{\beta}}$, they can be deduced as

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 \leq 2c_1^{-2}c_2 L\sqrt{(s\log n)/n}, \quad \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_q \leq 2c_1^{-2}c_2 L s^{\frac{1}{q}}\sqrt{(\log n)/n},$$
$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_\infty \leq \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 \leq 2L^{-1}c_1^{-2}c_2\sqrt{(s\log n)/n}.$$

The tail probability for these results to hold is at most the sum of the tail probabilities in Conditions 3-5 and Lemma EC.2. Thus, we know that these properties hold simultaneously with probability at least

$$1 - \frac{4\sqrt{2}\sigma}{c_2\sqrt{\pi\log p}}p^{1-\frac{c_2^2}{8\sigma^2}} + \frac{2\sqrt{2}\sigma s}{c_2\sqrt{\pi\log n}}n^{-\frac{c_2^2}{8\sigma^2}} - \theta_1 - \theta_2 - \theta_3,$$

which concludes the proof of Theorem 2.

EC.2. Additional technical details

The following lemma is needed in proving Lemma EC.1.

LEMMA EC.3 **(Weyl's inequality (Horn and Johnson 1990))**. *If* $\mathbf{A}$ *and* $\mathbf{B}$ *are* $m \times m$ *real symmetric matrices, then for all* $k = 1, \ldots, m$,

$$\left.\begin{array}{c} \varphi_k(\mathbf{A}) + \varphi_m(\mathbf{B}) \\ \varphi_{k+1}(\mathbf{A}) + \varphi_{m-1}(\mathbf{B}) \\ \vdots \\ \varphi_m(\mathbf{A}) + \varphi_k(\mathbf{B}) \end{array}\right\} \leq \varphi_k(\mathbf{A}+\mathbf{B}) \leq \left\{\begin{array}{c} \varphi_k(\mathbf{A}) + \varphi_1(\mathbf{B}) \\ \varphi_{k-1}(\mathbf{A}) + \varphi_2(\mathbf{B}) \\ \vdots \\ \varphi_1(\mathbf{A}) + \varphi_k(\mathbf{B}) \end{array}\right.,$$

where $\varphi_i(\cdot)$ *is the function that takes the* i*th largest eigenvalue of a given matrix.*

EC.2.1. Proof of Lemma EC.1

The main idea of proving Lemma EC.1 is to use induction to show that the sample eigenvalues divided by their corresponding orders of q will be convergent in an event with asymptotic probability one. To ease readability, the proof is divided into three steps.

Step 1: Large probability event $\mathcal{E}$. In this step, we will define an event $\mathcal{E}$ and show that its probability approaches one when q increases to infinity. Our later discussion will be conditional on this event. Denote a series of events by $\mathcal{E}_{jt}$, $1 \le j \le q$, $1 \le t \le n$, such that

$$\mathcal{E}_{jt} = \{z_{jt}^2 \le K^{-1} q^{\alpha}\},$$

where z_{jt} is the (j,t)th entry of $\mathbf{Z}$. By Condition 2, the events $\mathcal{E}_{jt}$ satisfy a uniform tail probability bound $P(\mathcal{E}_{jt}^c) = o(q^{-1}n^{-1})$. Let $\mathcal{E} = \cap_{t=1}^{n} \cap_{j=1}^{q} \mathcal{E}_{jt}$ be the intersection of all events in the series. Then the probability of event $\mathcal{E}$ converges to one since

$$P(\mathcal{E}^c) = P(\cup_{t=1}^{n} \cup_{j=1}^{q} \mathcal{E}_{jt}^c) \le \sum_{t=1}^{n} \sum_{j=1}^{q} P(\mathcal{E}_{jt}^c) = nq \cdot o(q^{-1}n^{-1}) \to 0, \ as \ q \to \infty.$$

Step 2: Convergence of eigenvalues with indices in J_1. This is the first part of induction. We will show that conditional on event $\mathcal{E}$, uniformly over $i \in J_1$, $q^{-\alpha_1}\widehat{\lambda}_i \to c_i$, as $q \to \infty$.

Denote by $\mathbf{C}$ the $q \times q$ diagonal matrix with the first K diagonal components equaling to c_j, $1 \le j \le K$, and the rest diagonal components 1. We decompose $\mathbf{Z}, \mathbf{C}$ and $\mathbf{\Lambda}$ into block matrices according to the index sets $J_1, J_2, \ldots, J_{m+1}$ such that

$$\mathbf{Z} = \begin{pmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \\ \vdots \\ \mathbf{Z}_{m+1} \end{pmatrix}, \mathbf{C} = \begin{pmatrix} \mathbf{C}_1 & \mathbf{O} & \cdots & \mathbf{O} \\ \mathbf{O} & \mathbf{C}_2 & \cdots & \mathbf{O} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{O} & \mathbf{O} & \cdots & \mathbf{C}_{m+1} \end{pmatrix}, \mathbf{\Lambda} = \begin{pmatrix} \mathbf{\Lambda}_1 & \mathbf{O} & \cdots & \mathbf{O} \\ \mathbf{O} & \mathbf{\Lambda}_2 & \cdots & \mathbf{O} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{O} & \mathbf{O} & \cdots & \mathbf{\Lambda}_{m+1} \end{pmatrix}. \tag{EC.23}$$

Then for the dual matrix S_D, we have

$$\mathbf{S}_D = n^{-1}\mathbf{Z}^T\mathbf{\Lambda}\mathbf{Z} = n^{-1}\sum_{l=1}^{m+1} \mathbf{Z}_l^T \mathbf{\Lambda}_l \mathbf{Z}_l. \tag{EC.24}$$

Divided by q^{α_1} on both sides of (EC.24) gives

$$q^{-\alpha_1}\mathbf{S}_D = n^{-1}q^{-\alpha_1}\mathbf{Z}_1^T\mathbf{\Lambda}_1\mathbf{Z}_1 + n^{-1}q^{-\alpha_1}\sum_{l=2}^{m} \mathbf{Z}_l^T\mathbf{\Lambda}_l\mathbf{Z}_l + n^{-1}q^{-\alpha_1}\mathbf{Z}_{m+1}^T\mathbf{\Lambda}_{m+1}\mathbf{Z}_{m+1}. \tag{EC.25}$$

We will show the sum of the last two terms above converges to the zero matrix in Frobenius norm, where the Frobenius norm is defined as $\|\mathbf{A}\|_F = \{\mathrm{tr}(\mathbf{A}\mathbf{A}^T)\}^{1/2}$ for a given matrix $\mathbf{A}$.

For any l, $1 \le l \le m$, let $\lambda_t^{(l)}$ and $c_t^{(l)}$ be the tth diagonal elements of $\mathbf{\Lambda}_l$ and $\mathbf{C}_l$, respectively. Conditional on event $\mathcal{E}$, for any j and k, $1 \le j,k \le n$, the absolute value of the (j,k)th element in $\sum_{l=2}^m \mathbf{Z}_l^T \mathbf{\Lambda}_l \mathbf{Z}_l$ is

$$|\sum_{l=2}^{m}\sum_{t=1}^{k_l} \lambda_t^{(l)} z_{tj}^{(l)} z_{tk}^{(l)}| \le K^{-1} q^{\alpha} \sum_{l=2}^{m}\sum_{t=1}^{k_l} \lambda_t^{(l)},$$

where $z_{tj}^{(l)}$ and $z_{tk}^{(l)}$ are the (t,j)th and (t,k)th elements in $\mathbf{Z}_l$, respectively. By Condition 1, uniformly over $1 \le l \le m$ and $1 \le t \le k_l$, $\lambda_t^{(l)} = O(q^{\alpha_l} c_t^{(l)})$. Then it follows that

$$\begin{aligned}
&\|n^{-1}q^{-\alpha_1}\sum_{l=2}^{m} \mathbf{Z}_l^T \mathbf{\Lambda}_l \mathbf{Z}_l\|_F \le n^{-1}q^{-\alpha_1}(nK^{-1}q^{\alpha}\sum_{l=2}^{m}\sum_{t=1}^{k_l}\lambda_t^{(l)}) \\
&= O\{q^{-\alpha_1}K^{-1}q^{\alpha}\sum_{l=2}^{m}\sum_{t=1}^{k_l} q^{\alpha_l} c_t^{(l)}\} = O\{K^{-1}q^{\alpha}\sum_{l=2}^{m} k_l C q^{\alpha_l}/q^{\alpha_1}\}.
\end{aligned}$$

Similarly we would get

$$\|n^{-1}q^{-\alpha_1}\mathbf{Z}_{m+1}^T \mathbf{\Lambda}_{m+1}\mathbf{Z}_{m+1}\|_F \le K^{-1}q^{\alpha}k_{m+1}C/q^{\alpha_1}.$$

Together with $\alpha < \min\{\Delta, \alpha_m - 1\}$ by Condition 2 and $k_{m+1} < q$, we have

$$\begin{aligned}
&\|n^{-1}q^{-\alpha_1}\sum_{l=2}^{m}\mathbf{Z}_l^T\mathbf{\Lambda}_l\mathbf{Z}_l + n^{-1}q^{-\alpha_1}\mathbf{Z}_{m+1}^T\mathbf{\Lambda}_{m+1}\mathbf{Z}_{m+1}\|_F \\
&\le O\{(\sum_{l=2}^{m} k_l q^{\alpha_l} + k_{m+1})K^{-1}q^{\alpha}C/q^{\alpha_1}\} \to 0, \ as\ q\to\infty. \qquad \text{(EC.26)}
\end{aligned}$$

By a similar argument, under Condition 1, we have

$$\begin{aligned}
&\|n^{-1}q^{-\alpha_1}\mathbf{Z}_1^T\mathbf{\Lambda}_1\mathbf{Z}_1 - n^{-1}\mathbf{Z}_1^T\mathbf{C}_1\mathbf{Z}_1\|_F = \|n^{-1}\mathbf{Z}_1^T(q^{-\alpha_1}\mathbf{\Lambda}_1 - \mathbf{C}_1)\mathbf{Z}_1\|_F \\
&\le n^{-1}[nK^{-1}q^{\alpha}\sum_{t=1}^{k_1}(q^{-\alpha_1}\lambda_t^{(1)} - c_t^{(1)})] \le k_1 K^{-1}q^{\alpha}\cdot O(q^{-\Delta}) \to 0, \ as\ q\to\infty. \qquad \text{(EC.27)}
\end{aligned}$$

In view of (EC.25), it is immediate that

$$\begin{aligned}
&\|q^{-\alpha_1}\mathbf{S}_D - n^{-1}\mathbf{Z}_1^T\mathbf{C}_1\mathbf{Z}_1\|_F \le \|n^{-1}q^{-\alpha_1}\mathbf{Z}_1^T\mathbf{\Lambda}_1\mathbf{Z}_1 - n^{-1}\mathbf{Z}_1^T\mathbf{C}_1\mathbf{Z}_1\|_F \\
&+ \|n^{-1}q^{-\alpha_1}\sum_{l=2}^{m}\mathbf{Z}_l^T\mathbf{\Lambda}_l\mathbf{Z}_l + n^{-1}q^{-\alpha_1}\mathbf{Z}_{m+1}^T\mathbf{\Lambda}_{m+1}\mathbf{Z}_{m+1}\|_F \to 0, \ as\ q\to\infty. \qquad \text{(EC.28)}
\end{aligned}$$

Further applying (Horn and Johnson 1990, Corollary 6.3.8) gives as $q \to \infty$,

$$\max_{1\le i\le n} |\varphi_i(q^{-\alpha_1}\mathbf{S}_D) - \varphi_i(n^{-1}\mathbf{Z}_1^T\mathbf{C}_1\mathbf{Z}_1)| \le \|q^{-\alpha_1}\mathbf{S}_D - n^{-1}\mathbf{Z}_1^T\mathbf{C}_1\mathbf{Z}_1\|_F \to 0. \tag{EC.29}$$

Note that $n^{-1}\mathbf{Z}_1^T\mathbf{C}_1\mathbf{Z}_1$ shares the same nonzero eigenvalues with its due matrix, that is, $n^{-1}\mathbf{C}_1^{1/2}\mathbf{Z}_1\mathbf{Z}_1^T\mathbf{C}_1^{1/2}$ of dimensionality k_1. It follows from (EC.29) that

$$\max_{i\in J_1} |\varphi_i(q^{-\alpha_1}\mathbf{S}_D) - \varphi_i(n^{-1}\mathbf{C}_1^{1/2}\mathbf{Z}_1\mathbf{Z}_1^T\mathbf{C}_1^{1/2})| \to 0. \tag{EC.30}$$

Moreover, by part (b) of Condition 2, we have

$$\max_{i\in J_1} |\varphi_i(n^{-1}\mathbf{C}_1^{1/2}\mathbf{Z}_1\mathbf{Z}_1^T\mathbf{C}_1^{1/2}) - \varphi_i(\mathbf{C}_1)| \le \|\mathbf{C}_1^{1/2}(n^{-1}\mathbf{Z}_1\mathbf{Z}_1^T - \mathbf{I}_{k_1})\mathbf{C}_1^{1/2}\|_F \to 0. \tag{EC.31}$$

Therefore, (EC.30) and (EC.31) together yield that uniformly over $i \in J_1$,

$$q^{-\alpha_1}\widehat{\lambda}_i = \varphi_i(q^{-\alpha_1}\mathbf{S}_D) \to \varphi_i(\mathbf{C}_1) = c_i,$$

as $q \to \infty$. This completes the proof of **Step 2**.

Step 3: Convergence of eigenvalues with indices in $J_2, \ldots, J_m$. As the second part of induction, for any fixed t, $2 \le t \le m$, we will show $q^{-\alpha_t}\widehat{\lambda}_i \to c_i$ for any $i \in J_t$, as $q \to \infty$. The basic idea in this step is to use Weyl's inequality (Lemma EC.3) to get both a lower bound and an upper bound of $q^{-\alpha_t}\widehat{\lambda}_i$, and show that they converge to the same limit.

We derive the upper bound first. Divided by q^{α_t} on both sides of (EC.24) gives

$$q^{-\alpha_t}\mathbf{S}_D = n^{-1}q^{-\alpha_t}\sum_{l=1}^{t-1}\mathbf{Z}_l^T\mathbf{\Lambda}_l\mathbf{Z}_l + n^{-1}q^{-\alpha_t}\sum_{l=t}^{m+1}\mathbf{Z}_l^T\mathbf{\Lambda}_l\mathbf{Z}_l.$$

Applying Weyl's inequality, we get

$$\begin{aligned}\varphi_i(q^{-\alpha_t}\mathbf{S}_D) &\le \varphi_{1+\sum_{l=1}^{t-1}k_l}(\sum_{l=1}^{t-1}\mathbf{Z}_l^T\mathbf{\Lambda}_l\mathbf{Z}_l/nq^{\alpha_t}) + \varphi_{i-\sum_{l=1}^{t-1}k_l}(n^{-1}q^{-\alpha_t}\sum_{l=t}^{m+1}\mathbf{Z}_l^T\mathbf{\Lambda}_l\mathbf{Z}_l)\\ &= \varphi_{i-\sum_{l=1}^{t-1}k_l}(n^{-1}q^{-\alpha_t}\sum_{l=t}^{m+1}\mathbf{Z}_l^T\mathbf{\Lambda}_l\mathbf{Z}_l),\end{aligned} \tag{EC.32}$$

where the first term is indeed zero since $n^{-1}q^{-\alpha_t}\sum_{l=1}^{t-1}\mathbf{Z}_l^T\mathbf{\Lambda}_l\mathbf{Z}_l$ has a rank no more than $\sum_{l=1}^{t-1}k_l$. It gives an upper bound of $\varphi_i(q^{-\alpha_t}\mathbf{S}_D)$. By the same argument as (EC.28) in **Step 2**, under Conditions 1 and 2, we have

$$\|n^{-1}q^{-\alpha_t}\sum_{l=t}^{m+1}\mathbf{Z}_l^T\mathbf{\Lambda}_l\mathbf{Z}_l - n^{-1}\mathbf{Z}_t^T\mathbf{C}_t\mathbf{Z}_t\|_F \to 0,\ as\ q\to\infty.$$

Similar to (EC.29), it implies the upper bound of $\varphi_i(q^{-\alpha_t}\mathbf{S}_D)$ in (EC.32) converges to the same limit as $\varphi_{i-\sum_{l=1}^{t-1}k_l}(n^{-1}\mathbf{Z}_t^T\mathbf{C}_t\mathbf{Z}_t)$ uniformly over $i\in J_t$ as $q\to\infty$.

On the other hand, by Weyl's inequality, we also have

$$\begin{aligned}\varphi_i(q^{-\alpha_t}\mathbf{S}_D) &\geq \varphi_i(n^{-1}q^{-\alpha_t}\sum_{l=1}^{t}\mathbf{Z}_l^T\mathbf{\Lambda}_l\mathbf{Z}_l) + \varphi_n(n^{-1}q^{-\alpha_t}\sum_{l=t+1}^{m+1}\mathbf{Z}_l^T\mathbf{\Lambda}_l\mathbf{Z}_l)\\ &\geq \varphi_i(n^{-1}q^{-\alpha_t}\sum_{l=1}^{t}\mathbf{Z}_l^T\mathbf{\Lambda}_l\mathbf{Z}_l),\end{aligned}$$

where the second term vanishes since the eigenvalues of $\sum_{l=t+1}^{m+1}\mathbf{Z}_l^T\mathbf{\Lambda}_l\mathbf{Z}_l$ are non-negative. In fact, $n^{-1}q^{-\alpha_t}\sum_{l=t+1}^{m+1}\mathbf{Z}_l^T\mathbf{\Lambda}_l\mathbf{Z}_l$ would converge to a zero matrix in Frobenius norm under Conditions 1 and 2, similarly as in (EC.26). For the term $\varphi_i(n^{-1}q^{-\alpha_t}\sum_{l=1}^{t}\mathbf{Z}_l^T\mathbf{\Lambda}_l\mathbf{Z}_l)$, we use Weyl's inequality once more to get

$$\begin{aligned}&\varphi_{\sum_{l=1}^{t}k_l}(n^{-1}q^{-\alpha_t}\sum_{l=1}^{t-1}\mathbf{Z}_l^T\mathbf{\Lambda}_l\mathbf{Z}_l)\\ \leq\ &\varphi_i(n^{-1}q^{-\alpha_t}\sum_{l=1}^{t}\mathbf{Z}_l^T\mathbf{\Lambda}_l\mathbf{Z}_l) + \varphi_{1-i+\sum_{l=1}^{t}k_l}(-n^{-1}q^{-\alpha_t}\mathbf{Z}_t^T\mathbf{\Lambda}_t\mathbf{Z}_t).\end{aligned}$$

Note that the term on the left hand side is indeed zero since the inside matrix has a rank no more than $\sum_{l=1}^{t-1}k_l$. It follows that

$$\begin{aligned}&\varphi_i(n^{-1}q^{-\alpha_t}\sum_{l=1}^{t}\mathbf{Z}_t^T\mathbf{\Lambda}_t\mathbf{Z}_t) \geq -\varphi_{1-i+\sum_{l=1}^{t}k_l}(-n^{-1}q^{-\alpha_t}\mathbf{Z}_t^T\mathbf{\Lambda}_t\mathbf{Z}_t)\\ =&\varphi_{k_t-(1-i+\sum_{l=1}^{t}k_l)+1}(n^{-1}q^{-\alpha_t}\mathbf{Z}_t^T\mathbf{\Lambda}_t\mathbf{Z}_t) = \varphi_{i-\sum_{l=1}^{t-1}k_l}(n^{-1}q^{-\alpha_t}\mathbf{Z}_t^T\mathbf{\Lambda}_t\mathbf{Z}_t),\end{aligned}$$

where we make use of the fact that $\varphi_i(\mathbf{A}) = -\varphi_{n-i+1}(-\mathbf{A})$ for any $n\times n$ real symmetric matrix $\mathbf{A}$, and any $1\leq i\leq n$.

Therefore, we get a lower bound $\varphi_{i-\sum_{l=1}^{t-1}k_l}(n^{-1}q^{-\alpha_t}\mathbf{Z}_t^T\mathbf{\Lambda}_t\mathbf{Z}_t)$ for $\varphi_i(q^{-\alpha_t}\mathbf{S}_D)$. In terms of sample eigenvalues, the above argument shows that for any $\widehat{\lambda}_i$, $i\in J_t$, $1\le t\le m$,

$$\widehat{\lambda}_i=\varphi_i(\mathbf{S}_D)\ge\varphi_{i-\sum_{l=1}^{t-1}k_l}(n^{-1}\mathbf{Z}_t^T\mathbf{\Lambda}_t\mathbf{Z}_t), \tag{EC.33}$$

which is useful in proving the convergence properties of the sample score vectors.

Now we show that the two bounds converge to the same limit. Similar to (EC.27), as $q\to\infty$, we have

$$\|n^{-1}q^{-\alpha_t}\mathbf{Z}_t^T\mathbf{\Lambda}_t\mathbf{Z}_t-n^{-1}\mathbf{Z}_t^T\mathbf{C}_t\mathbf{Z}_t\|_F\to 0,$$

which gives

$$\max_{i\in J_t}\left|\varphi_{i-\sum_{l=1}^{t-1}k_l}(n^{-1}q^{-\alpha_t}\mathbf{Z}_t^T\mathbf{\Lambda}_t\mathbf{Z}_t)-\varphi_{i-\sum_{l=1}^{t-1}k_l}(n^{-1}\mathbf{Z}_t^T\mathbf{C}_t\mathbf{Z}_t)\right|\to 0.$$

It shows that the lower bound of $\varphi_i(q^{-\alpha_t}\mathbf{S}_D)$ converges to the same limit as the term $\varphi_{i-\sum_{l=1}^{t-1}k_l}(n^{-1}\mathbf{Z}_t^T\mathbf{C}_t\mathbf{Z}_t)$ uniformly over $i\in J_t$, so does the upper bound in (EC.32). It follows that $\varphi_i(q^{-\alpha_t}\mathbf{S}_D)$ would also converge to the same limit as $\varphi_{i-\sum_{l=1}^{t-1}k_l}(n^{-1}\mathbf{Z}_t^T\mathbf{C}_t\mathbf{Z}_t)$ uniformly over $i\in J_t$. That is, as $q\to\infty$,

$$\max_{i\in J_t}|\varphi_i(q^{-\alpha_t}\mathbf{S}_D)-\varphi_{i-\sum_{l=1}^{t-1}k_l}(n^{-1}\mathbf{Z}_t^T\mathbf{C}_t\mathbf{Z}_t)|\to 0.$$

By a similar argument as in (EC.30) and (EC.31), we then have

$$\varphi_i(q^{-\alpha_t}\mathbf{S}_D)\to\varphi_{i-\sum_{l=1}^{t-1}k_l}(\mathbf{C}_t)=c_i,$$

uniformly over $i\in J_t$, as $q\to\infty$. Along with the first step of induction in **Step 2**, we finish the proof of Lemma EC.1.

EC.2.2. Proof of Lemma EC.2

To prove the probability bound in Lemma EC.2, we will apply Bonferroni's inequality and Gaussian tail probability bound. Since $\widetilde{\boldsymbol{\varepsilon}}=(\mathbf{F}-\widehat{\mathbf{F}})\boldsymbol{\gamma}+\boldsymbol{\varepsilon}$, some important bounds are needed before

continuation. First, the inequality $\|\boldsymbol{\gamma}\|_1 \le K\rho$ follows immediately from the fact $\|\boldsymbol{\gamma}\|_\infty \le \rho$. Moreover, based on the estimation error bound of $\widehat{\mathbf{F}}$ in Condition 3, we know that inequality (EC.14) holds. These two inequalities yield

$$\|(\mathbf{F}-\widehat{\mathbf{F}})\boldsymbol{\gamma}\|_2 \le \|\boldsymbol{\gamma}\|_1 \cdot \max_{1\le j\le K}\|\mathbf{f}_j-\widehat{\mathbf{f}}_j\|_2 \le K\rho\cdot\frac{c_2}{2K\rho}\sqrt{\log n}=\frac{c_2}{2}\sqrt{\log n}, \tag{EC.34}$$

which gives

$$n^{-1}|\mathbf{x}_i^T(\mathbf{F}-\widehat{\mathbf{F}})\boldsymbol{\gamma}| \le n^{-1/2}\|(\mathbf{F}-\widehat{\mathbf{F}})\boldsymbol{\gamma}\|_2 \le \frac{c_2}{2}\sqrt{\frac{\log n}{n}} \le \frac{c_2}{2}\sqrt{\frac{\log p}{n}}. \tag{EC.35}$$

Similarly we have $n^{-1}|\mathbf{f}_j^T(\mathbf{F}-\widehat{\mathbf{F}})\boldsymbol{\gamma}| \le 2^{-1}c_2\sqrt{(\log n)/n}$.

Now we proceed to prove the probability bounds of the two events. Recall that both $\mathbf{f}_j$ and $\widehat{\mathbf{f}}_j$ have been rescaled to have L_2-norm $n^{1/2}$ and $\boldsymbol{\varepsilon}\sim N(\mathbf{0},\sigma^2\mathbf{I}_n)$ (Section 2). Given $\mathbf{x}_i$ and $\widehat{\mathbf{f}}_j$, it follows that $n^{-1}\mathbf{x}_i^T\boldsymbol{\varepsilon}\sim N(0,\sigma^2/n)$ and $n^{-1}\widehat{\mathbf{f}}_j^T\boldsymbol{\varepsilon}\sim N(0,\sigma^2/n)$ for any i and j. By Bonferroni's inequality, the tail probability of $\widetilde{\mathcal{E}}$ satisfies

$$P(\widetilde{\mathcal{E}}^c) \le \sum_{i=1}^p P\left(|n^{-1}\mathbf{x}_i^T\widetilde{\boldsymbol{\varepsilon}}|>c_2\sqrt{(\log p)/n}\right)+\sum_{j=1}^K P\left(|n^{-1}\widehat{\mathbf{f}}_j^T\widetilde{\boldsymbol{\varepsilon}}|>c_2\sqrt{(\log p)/n}\right).$$

By inequality (EC.35) and Gaussian tail probability bound, for the first term on the right hand side above, we have

$$\begin{aligned}&\sum_{i=1}^p P\left(\frac{|\mathbf{x}_i^T\widetilde{\boldsymbol{\varepsilon}}|}{n}>c_2\sqrt{\frac{\log p}{n}}\right) \le \sum_{i=1}^p P\left(\frac{|\mathbf{x}_i^T\boldsymbol{\varepsilon}|}{n}>c_2\sqrt{\frac{\log p}{n}}-n^{-1}|\mathbf{x}_i^T(\mathbf{F}-\widehat{\mathbf{F}})\boldsymbol{\gamma}|\right)\\ \le &\sum_{i=1}^p P\left(\frac{|\mathbf{x}_i^T\boldsymbol{\varepsilon}|}{n}>\frac{c_2}{2}\sqrt{\frac{\log p}{n}}\right) \le \sum_{j=1}^p \frac{4\sigma}{c_2\sqrt{\log p}}\frac{1}{\sqrt{2\pi}}e^{-\frac{c_2^2\log p}{8\sigma^2}} \le \frac{2\sqrt{2}\sigma}{c_2\sqrt{\pi\log p}}p^{1-\frac{c_2^2}{8\sigma^2}}.\end{aligned}$$

For the second term, similarly we have

$$\sum_{j=1}^K P\left(|n^{-1}\widehat{\mathbf{f}}_j^T\widetilde{\boldsymbol{\varepsilon}}|>c_2\sqrt{(\log p)/n}\right) \le \frac{2\sqrt{2}\sigma K}{c_2\sqrt{\pi\log p}}p^{-\frac{c_2^2}{8\sigma^2}}.$$

As K is no larger than p, the two bounds above give

$$P(\widetilde{\mathcal{E}}^c) \le \frac{4\sqrt{2}\sigma}{c_2\sqrt{\pi\log p}}p^{1-\frac{c_2^2}{8\sigma^2}}.$$

By a similar argument, the bound on $P(\widetilde{\mathcal{E}}_0^c)$ can be derived as

$$P(\widetilde{\mathcal{E}}_0^c) \leq \frac{2\sqrt{2}\sigma s}{c_2\sqrt{\pi \log n}} n^{-\frac{c_2^2}{8\sigma^2}}.$$

Thus, for the intersection event $\widetilde{\mathcal{E}} \cap \widetilde{\mathcal{E}}_0$, we have

$$P\{(\widetilde{\mathcal{E}} \cap \widetilde{\mathcal{E}}_0)^c\} \leq P(\tilde{\mathcal{E}}^c) + P(\widetilde{\mathcal{E}}_0^c) \leq \frac{4\sqrt{2}\sigma}{c_2\sqrt{\pi \log p}} p^{1-\frac{c_2^2}{8\sigma^2}} + \frac{2\sqrt{2}\sigma s}{c_2\sqrt{\pi \log n}} n^{-\frac{c_2^2}{8\sigma^2}},$$

which converges to zero as $n \to \infty$ for $c_2 > 2\sqrt{2}\sigma$. This completes the proof of Lemma EC.2.