Factor Profiling for Ultra High Dimensional Variable Selection

Hansheng Wang

Guanghua School of Management, Peking University

Abstract

We propose here a novel method of factor profiling (FP) for ultra high dimensional variable selection. The new method assumes that the correlation structure of the high dimensional data can be well represented by a set of low-dimensional latent factors (Fan et al., 2008). The latent factors can then be estimated consistently by eigenvalue-eigenvector decomposition. They should be profiled out subsequently from both the response and predictors. Such an operation is referred to as FP. Obviously, FP produces uncorrelated predictors. Thereafter, the method of sure independent screening (Fan and Lv, 2008, SIS) can be applied immediately. This leads to profiled independent screening (PIS). PIS is shown to be selection consistent, even if the predictor dimension is substantially larger than the sample size. To further improve PIS, a novel method of profiled sequential screening (PSS) is proposed. PSS shares similar strength as forward regression (Wang, 2009a) but is computationally even simpler. Numerical studies are presented to corroborate our theoretical findings.

KEY WORDS: Bayesian Information Criterion; Factor Profiling; Forward Regression; Maximum Eigenvalue Ratio Criterion; Profiled Independent Screening; Profiled Sequential Screening; Selection Consistency; Screening Consistency

†Hansheng Wang is from Guanghua School of Management, Peking University, Beijing, 100871, P. R. China. This research is supported in part by a NSFC grant (No. 10771006). The author is very grateful to Mr. Lilun Du, Mr. Wei Lan, Mr. Hanzhong Liu, and Miss. Rui Pan for the careful reading and a lot of helpful comments.

1. INTRODUCTION

For many modern datasets, the predictor dimensions are found substantially larger than the sample sizes. As a consequence, classical methods, such as the ordinary least squares (OLS), are no longer immediately applicable. Then, dimension reduction becomes the central theme of high dimensional data analysis, for which the idea of variable selection has been found very useful (Fan and Li, 2006).

Under a fixed dimension setup, best subset selection in conjunction with two popular criteria has been widely used in practice. Those two criteria are, respectively, the AIC (Akaike, 1973) and BIC (Schwarz, 1978). Despite its usefulness, such a method suffers from high computational cost (Tibshirani, 1996), estimation instability (Breiman, 1996), and also complicated stochastic property (Fan and Li, 2001; Hjort and Claeskens, 2003). As computationally efficient alternatives, various shrinkage methods have been developed and gained a lot of popularity in the past decade. Those methods include, for example, the nonnegative garrotte (Breiman, 1995; Yuan and Lin, 2007), the LASSO (Tibshirani, 1996; Knight and Fu, 2000; Zhao and Yu, 2006), the bridge regression (Fu, 1998; Huang et al., 2007), the SCAD (Fan and Li, 2001; Fan and Peng, 2004; Wang et al., 2007), the elastic net (Zou and Hastie, 2005), the adaptive LASSO (Zou, 2006; Zhang and Lu, 2007; Wang and Leng, 2007), one-step sparse estimation (Zou and Li, 2008), the adaptive elastic net (Zou and Zhang, 2009), and others. Without any doubt, those methods are useful. Many of them have been shown to be consistent for model selection, but under the constraint that the predictor dimension should be no more than the sample size. In contrast, if the predictor dimension is much larger than the sample size, none of them has been shown to be selection consistent under a general design condition (Leng et al., 2006; Zhao and Yu, 2006).

In addition to those shrinkage methods, Fan and Lv (2008) developed the theory

of sure independent screening (SIS). They show that the simple method of marginal correlation estimation is effective for variable screening. Given the fact that SIS is computationally very simple, such a nice property is not only practically useful but also theoretically appealing. Subsequently, Fan and Lv (2008) refer to it as a SIS property, which is also known as screening consistency by Wang (2009a). For a further improved performance, Wang (2009a) investigated another classical variable screening method, that is, forward regression (FR). Wang (2009a) proves that FR also enjoys the SIS or screening consistency property. Our experience suggests that both SIS and FR are useful methods but with two common limitations. Firstly, neither of them is selection consistent (Shao, 1997; Shi and Tsai, 2002). In fact, both of them suffer nonignorable overfitting effects. In other words, to get all relevant variables correctly discovered, both methods might have positive probabilities to have some irrelevant variables included. Secondly, neither of them can handle endogeneity problem, which means that the residual might be correlated with the predictor. Such types of problems are very often found in economics related datasets (Wooldridge, 2001). Consequently, we are motivated to develop a new method, which is capable of handling ultra high dimensional data even in presence of endogeneity problem.

To this end, we propose here a novel method of factor profiling (FP). Our method is well motivated by empirical evidences. Specifically, for many ultra high dimensional datasets, their first few eigenvalues are found to be substantially larger than the rest. Such an observation suggests that the high-dimensional predictors' correlation structure might be represented by a low-dimensional latent factor model (Fan et al., 2008). If those latent factors can be estimated consistently, they can be profiled out from both the predictors and the responses subsequently. For convenience, we refer to such an operation as FP. The consequence of FP is that the profiled predictors are no longer

correlated with each other. Furthermore, FP leads to uncorrelated residuals and thus the problem of endogeneity is fixed. Thereafter, SIS together can be used immediately. This leads to profiled independent screening (PIS). We show that PIS is consistent for model selection. In fact, its performance can be further improved considerably, if relevant predictors' regression effects can be eliminated from the response in a sequential manner. This leads to de-noised new responses and models. As a result, the selection accuracy can be further improved. We refer to such a method as profiled sequential screening (PSS). Obviously, PSS shares similar computational flavor as FR (Wang, 2009a). Nevertheless, PSS is even simpler because the factor profiled predictors are already uncorrelated; see Section 3.3 for a more detailed discussion.

The rest of the article is organized as the follows. Next section introduces FP together with its theoretical properties. The methods of PIS and PSS are investigated in Section 3. Numerical studies are reported in Section 4 and concluding remarks are given in Section 5. Technical details are left to the Appendix.

2. FACTOR PROFILING THEORY

2.1. Model and Notations

Let $Y_i \in \mathbb{R}^1$ be the response collected from the ith $(1 \leq i \leq n)$ subject and $X_i = (X_{i1}, \cdots, X_{ip})^\top \in \mathbb{R}^p$ be the associated p-dimensional predictor. We assume throughout the rest of this article that the predictor dimension p is ultra high and thus is substantially larger than the sample size n. To model the regression relationship between Y_i and X_i, we assume further

$$Y_i = X_i^\top \theta + \varepsilon_i, \tag{2.1}$$

where ε_i is a random noise with mean 0 and variance σ_ε^2; $\theta = (\theta_1, \cdots, \theta_p)^\top \in \mathbb{R}^p$ is a p-dimensional coefficient vector and its true value is given by $\theta_0 = (\theta_{01}, \cdots, \theta_{0p})^\top \in \mathbb{R}^p$. To model the predictor's correlation structure, we follow Fan et al. (2008) and assume

$$X_i = BZ_i + \widetilde{X}_i, \tag{2.2}$$

where $Z_i = (Z_{i1}, \cdots, Z_{id})^\top \in \mathbb{R}^d$ is a d-dimensional latent factor, $B = (b_{jk}) \in \mathbb{R}^{p\times d}$ is the loading matrix, and $\widetilde{X}_i = (\widetilde{X}_{i1}, \cdots, \widetilde{X}_{ip})^\top \in \mathbb{R}^p$ represents the information contained in X_i but missed by Z_i. For $\widetilde{X}_i$ we assume that $\mathrm{cov}(\widetilde{X}_i)$ is a diagonal matrix, that is $\mathrm{cov}(\widetilde{X}_{ij_1}, \widetilde{X}_{ij_2}) = 0$ for any $j_1 \neq j_2$. We assume further that $E(Y_i) = E(X_{ij}) = E(\widetilde{X}_{ij}) = 0$ and $\mathrm{var}(Y_i) = \mathrm{var}(X_{ij}) = 1 \geq \tilde{\sigma}_j^2 = \mathrm{var}(\widetilde{X}_{ij})$. In addition to that, we require $\mathrm{cov}(Z_i) = I$, where I stands for an identity matrix with an appropriate dimension. For example, we should have $I \in \mathbb{R}^{d\times d}$ here because $Z_i \in \mathbb{R}^d$. Otherwise, we can always re-define $Z_i = \mathrm{cov}^{-1/2}(Z_i)Z_i$ and $B = B\mathrm{cov}^{1/2}(Z_i)$, so that the condition $\mathrm{cov}(Z_i) = I$ can be well satisfied.

To reflect the endogeneity problem, we allow that ε_i to be correlated with X_i through the common factor Z_i as

$$\varepsilon_i = Z_i^\top \alpha + \tilde{\varepsilon}_i, \tag{2.3}$$

where $\alpha = (\alpha_1, \cdots, \alpha_d)^\top \in \mathbb{R}^d$ is a d-dimensional vector and its true value is given by $\alpha_0 \in \mathbb{R}^d$. Moreover, $\tilde{\varepsilon}_i$ is some random noise independent of both Z_i and $\widetilde{X}_i$. We then should have $\mathrm{var}(\tilde{\varepsilon}_i) = \tilde{\sigma}_\varepsilon^2 \leq \mathrm{var}(Y_i) = 1$. Because ε_i might be correlated with X_i, the traditional OLS estimate is biased, even if the predictor dimension is fixed and the sample size is infinite (Wooldridge, 2001). However, the story changes, if Z_i can be eliminated from both Y_i and X_i. Specifically, define a profiled response as

$\widetilde{Y}_i = Y_i - Z_i^\top \gamma_0$ with $\gamma_0 = B^\top \theta_0 + \alpha_0$. Next, refer to $\widetilde{X}_i$ and $\tilde{\varepsilon}_i$ as a profiled predictor and noise, respectively. We then have

$$\widetilde{Y}_i = \widetilde{X}_i^\top \theta_0 + \tilde{\varepsilon}_i. \tag{2.4}$$

For model (2.4), not only $\widetilde{X}_i$ and $\tilde{\varepsilon}_i$ are mutually uncorrelated, but also different predictors (i.e., $\widetilde{X}_{ij_1}$ and $\widetilde{X}_{ij_2}$ for any $j_1 \neq j_2$) are mutually uncorrelated. Subsequently, the unknown regression coefficients can be estimated by SIS. This seems to be a fairly appealing procedure due to its computational simplicity. In fact, as we are going to prove later, its theoretical properties are also excellent.

The above discussion motivates us to develop a FP methodology. Before we introduce the details, some notations need to be defined. Let $\mathbb{Y} = (Y_1, \cdots, Y_n)^\top \in \mathbb{R}^n$ be the response vector, $\mathbb{X} = (X_1, \cdots, X_n)^\top \in \mathbb{R}^{n \times p}$ be the design matrix, and $\mathcal{E} = (\varepsilon_1, \cdots, \varepsilon_n)^\top \in \mathbb{R}^n$ be the noise vector. Their profiled versions are defined similarly, which are denoted by $\widetilde{\mathbb{Y}}$, $\widetilde{\mathbb{X}}$, and $\widetilde{\mathcal{E}}$, respectively. Next, define $\mathbb{X}_j = (X_{1j}, \cdots, X_{nj})^\top \in \mathbb{R}^n$ to be the jth column of $\mathbb{X}$. Similarly, $\widetilde{\mathbb{X}}_j$ is the jth column of $\widetilde{\mathbb{X}}$. By models (2.1), (2.2), and (2.3), we know that

$$\mathbb{Y} = \mathbb{Z}\gamma_0 + \widetilde{\mathbb{Y}} = \mathbb{Z}\gamma_0 + \widetilde{\mathbb{X}}\theta_0 + \widetilde{\mathcal{E}} \text{ and } \mathbb{X} = \mathbb{Z}B^\top + \widetilde{\mathbb{X}}, \tag{2.5}$$

where $\mathbb{Z} = (Z_1, \cdots, Z_n)^\top \in \mathbb{R}^{n \times d}$ is the design matrix of the latent factor. By (2.5), we know that the effects due to $\mathbb{Z}$ can be eliminated as long as one can estimate $\mathcal{S}(\mathbb{Z})$ accurately, where $\mathcal{S}(A)$ stands for the linear subspace spanned by the column vectors of an arbitrary matrix A. More specifically, if $\mathcal{S}(\mathbb{Z})$ is known, a projection matrix onto its orthogonal complement can be constructed. Denote such a projection matrix by $Q(\mathbb{Z}) = I - H(\mathbb{Z}) \in \mathbb{R}^{n \times n}$, where $H(\mathbb{Z}) = \mathbb{Z}(\mathbb{Z}^\top \mathbb{Z})^{-1}\mathbb{Z}^\top \in \mathbb{R}^{n \times n}$ is another projection

matrix but onto $\mathcal{S}(\mathbb{Z})$. We can then get $Q(\mathbb{Z})\mathbb{Y} = Q(\mathbb{Z})\mathbb{X}\theta_0 + Q(\mathbb{Z})\mathcal{E}$, which serves as an approximation towards the ideal model (2.4). As a result, we should focus on the factor subspace $\mathcal{S}(\mathbb{Z})$ directly.

2.2. Determining Factor Dimension

To estimate $\mathcal{S}(\mathbb{Z})$ accurately, it is necessary to specify its dimension (denoted by d_0) correctly. Because in real practice the value of d_0 is unknown, one has to estimate it based on data. As a simple and effective solution, we proposed here a method of maximum eigenvalue ratio criterion (Luo et al., 2009, MERC). Specifically, let $(\hat{\lambda}_j, \hat{V}_j)$ be the jth $(1 \leq j \leq n)$ leading eigenvalue-eigenvector pair for the matrix $\mathbb{X}\mathbb{X}^\top/(np) \in \mathbb{R}^{n\times n}$. Practically, they can be easily obtained by eigenvalue-eigenvector decomposition. By definition, we should have $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_n$. Because the true factor dimension is d_0, intuitively we should expect that first d_0 eigenvalues to be relatively large while the rest to be comparatively small. Thus, if we define an eigenvalue ratio criterion as $\hat{\lambda}_j/\hat{\lambda}_{j+1}$ with $\hat{\lambda}_0 = 1$ and $1 \leq j \leq (n-1)$, we should expect its maximum value to happen at $j = d_0$. Consequently, the true structure dimension can be estimated by $\hat{d} = \text{argmax}_{0\leq j\leq d_{\max}}(\hat{\lambda}_j/\hat{\lambda}_{j+1})$, where $d_{\max}$ is a pre-specified maximum factor dimension. We call $\hat{d}$ a MERC estimator.

In theory, we can allow $d_{\max} = n-1$; see the subsequent Theorem 1 and its detailed technical proofs in Appendix D. However, with finite data and limited computational precision, very often the last few eigenvalues (e.g., $\hat{\lambda}_n$) might be estimated badly. Thus, practically it is useful to pre-specify a maximum factor dimension $d_{\max}$. Our experience suggests that the MERC's finite sample performance is rather insensitive to the choice of $d_{\max}$, as long as the last few eigenvalues are excluded. For example, we always set $d_{\max}$ to be the smallest integer such that $(\sum_{j=1}^{d_{\max}} \hat{\lambda}_j)/(\sum_{j=1}^{n} \hat{\lambda}_j) \geq 99\%$. The resulting performance is excellent. Although the idea of MERC is intuitive and simple,

whether it is statistically sound needs to be further justified theoretically. Then, the following theorem rigorously proves that MERC can indeed estimate d_0 consistently; see Appendix D for a detailed proof.

Theorem 1. *Assume technical conditions (A1)–(A3) as given in the Appendix A, then we should have* $P(\hat{d} = d_0) \to 1$ *as* $n \to \infty$.

2.3. Estimating Factor Subspace

By Theorem 1, we know that the true factor dimension d_0 can be estimated consistently. With a correctly specified factor dimension (i.e., $d = d_0$), we can subsequently construct a least squares type objective function as

$$\mathcal{O}(\mathbb{Z}, B) = (np)^{-1} \sum_{j=1}^{p} \left\| \mathbb{X}_j - \mathbb{Z}\beta_j \right\|^2$$

with $\beta_j = (b_{j1}, \cdots, b_{jd})^\top \in \mathbb{R}^d$. We know immediately that $B = (\beta_1, \cdots, \beta_p)^\top \in \mathbb{R}^{p \times d}$. Then, $\mathcal{S}(\mathbb{Z})$ can be estimated by minimizing $\mathcal{O}(\mathbb{Z}, B)$ with respect to both $\mathbb{Z} \in \mathbb{R}^{n \times d}$ and $B \in \mathbb{R}^{p \times d}$. Specifically, for a fixed $\mathbb{Z}$, $\mathcal{O}(\mathbb{Z}, B)$ can be minimized by setting $B = \widehat{B} = (\hat{\beta}_1, \cdots, \hat{\beta}_p)$ and $\hat{\beta}_j = (\mathbb{Z}^\top \mathbb{Z})^{-1}(\mathbb{Z}^\top \mathbb{X}_j) \in \mathbb{R}^d$, which leads to the following profiled objective function

$$\mathcal{O}(\mathbb{Z}) = \mathcal{O}(\mathbb{Z}, \widehat{B}) = (np)^{-1} \sum_{j=1}^{p} \mathbb{X}_j^\top Q(\mathbb{Z}) \mathbb{X}_j = (np)^{-1} tr\Big\{ Q(\mathbb{Z}) \big(\mathbb{X}\mathbb{X}^\top \big) \Big\}, \tag{2.6}$$

where $tr(A)$ stands for the trace of an arbitrary square matrix A. One can then verify that (2.6) can be minimized by setting $\widehat{\mathbb{Z}} = (\hat{V}_1, \cdots, \hat{V}_d) \in \mathbb{R}^{n \times d}$; see Lemma L6 in the Appendix B. Subsequently, $\mathcal{S}(\mathbb{Z})$ can be estimated by $\mathcal{S}(\widehat{\mathbb{Z}})$.

To quantify the estimation accuracy of $\mathcal{S}(\widehat{\mathbb{Z}})$, the following two discrepancy mea-

sures are considered. They are, respectively,

$$D_1(\mathbb{Z}, \widehat{\mathbb{Z}}) = n^{-1} tr\Big\{\mathbb{Z}^\top Q(\widehat{\mathbb{Z}})\mathbb{Z}\Big\} \text{ and } D_2(\mathbb{Z}, \widehat{\mathbb{Z}}) = tr\Big\{H(\mathbb{Z}) - H(\widehat{\mathbb{Z}})\Big\}^2.$$

Obviously, $\mathcal{S}(\widehat{\mathbb{Z}}) = \mathcal{S}(\mathbb{Z})$ implies that $D_1(\mathbb{Z}, \widehat{\mathbb{Z}}) = D_2(\mathbb{Z}, \widehat{\mathbb{Z}}) = 0$. In addition to that, $D_2(\mathbb{Z}, \widehat{\mathbb{Z}}) = 0$ also implies that $\mathcal{S}(\mathbb{Z}) = \mathcal{S}(\widehat{\mathbb{Z}})$; see Xia (2007) and Wang and Xia (2008). It is worthwhile to point out that the first discrepancy measure $D_1(\cdot, \cdot)$ is not symmetric about its two arguments. In other words, it is not necessary that $D_1(\mathbb{Z}, \widehat{\mathbb{Z}}) = D_1(\widehat{\mathbb{Z}}, \mathbb{Z})$. However, the second one $D_2(\cdot, \cdot)$ is. We consider those two measures simply because they both are extensively used in subsequent theoretical development; see for example Appendix D. Then, next theorem says that both of them converge towards 0 at the rate of $O_p(n^{-1})$. See Appendix C for a detailed proof.

Theorem 2. *Assume $d = d_0$ and the technical conditions (A1)–(A3) as given in the Appendix A, then we should have both $D_1(\mathbb{Z}, \widehat{\mathbb{Z}}) = O_p(n^{-1})$ and $D_2(\mathbb{Z}, \widehat{\mathbb{Z}}) = O_p(n^{-1})$.*

3. PROFILED VARIABLE SELECTION

We define a generic notation $\mathcal{M} = \{j_1, \cdots, j_{d^*}\}$ to represent a candidate model, which includes X_{ij} for every $j \in \mathcal{M}$ as relevant variables. We use $|\mathcal{M}|$ to denote the corresponding model size. Thus, $|\mathcal{M}| = d^*$ for this case. We then define the full model as $\mathcal{M}_F = \{j : 1 \leq j \leq p\}$ and the true model as $\mathcal{M}_T = \{j : \theta_{0j} \neq 0\}$.

3.1. Profiled Independent Screening

By Theorem 1 we know that the factor dimension d_0 can be estimated consistently. With a correctly specified factor dimension (i.e., $d = d_0$), Theorem 2 further indicates that the factor subspace $\mathcal{S}(\mathbb{Z})$ can be estimated accurately. Thereafter, we can get factor profiled data as $\widehat{\mathbb{Y}} = Q(\widehat{\mathbb{Z}})\mathbb{Y} \in \mathbb{R}^n$ and $\widehat{\mathbb{X}} = Q(\widehat{\mathbb{Z}})\mathbb{X}$, with $\widehat{\mathbb{X}} = (\widehat{\mathbb{X}}_1, \cdots, \widehat{\mathbb{X}}_p) \in$

$\mathbb{R}^{n\times p}$. Similarly, the factor profiled noise is given by $\widehat{\mathcal{E}} = Q(\widehat{\mathbb{Z}})\mathcal{E}$. Subsequently, the simple method of SIS can be applied to $\widehat{\mathbb{Y}}$ and $\widehat{\mathbb{X}}$ directly, and the resulting estimate is path consistent (Leng et al., 2006). We refer to such a method as PIS. More specifically, PIS estimates θ_j by $\hat{\theta}_j = (n^{-1}\widehat{\mathbb{X}}_j^\top\widehat{\mathbb{X}}_j)^{-1}(n^{-1}\widehat{\mathbb{Y}}^\top\widehat{\mathbb{X}}_j)$. Without loss of generality, we assume further that the predictor indices have been appropriately re-labeled so that $|\hat{\theta}_1| > |\hat{\theta}_2| > \cdots > |\hat{\theta}_p|$. Then, a solution path is given by $\mathbb{M} = \{\mathcal{M}_{(k)} : 0 \le k \le p\}$ with $\mathcal{M}_{(0)} = \emptyset$ and $\mathcal{M}_{(k)} = \{1, \cdots, k\}$ for $1 \le k \le p$. The following theorem implies that $\mathbb{M}$ and thus PIS is path consistent (Leng et al., 2006), that is, $P(\mathcal{M}_T \in \mathbb{M}) \to 1$ as $n \to \infty$. See Appendix E for a detailed proof.

Theorem 3. *Assume $d = d_0$ and the technical conditions (A1)–(A3) as given in the Appendix A, then we should have $\max_{1\le j\le p}|\hat{\theta}_j - \theta_{0j}| = O_p(\sqrt{\log p/n})$ as $n \to \infty$.*

3.2. A Bayesian Information Criterion

Previous subsection proves that PIS is path consistent, which implies that $P(\mathcal{M}_T = \mathcal{M}_{(|\mathcal{M}_T|)}) \to 1$ as $n \to \infty$. However, for a real application, the value of $|\mathcal{M}_T|$ is unknown. Thus, even if the solution path is given, one still needs a statistically sound criterion to decide which model in $\mathbb{M}$ is mostly plausible. To this end, we proposed here the following heuristic BIC-type selection criterion,

$$\mathrm{BIC}(\mathcal{M}) = \log \mathrm{RSS}(\mathcal{M}) + |\mathcal{M}| \cdot \log n \cdot (\log p/n), \tag{3.1}$$

where $\mathrm{RSS}(\mathcal{M}) = \|\widehat{\mathbb{Y}} - \sum_{j\in\mathcal{M}} \hat{\theta}_j \widehat{\mathbb{X}}_j\|^2$ is the residual sum of squares. Then the best model can be selected as $\widehat{\mathcal{M}} = \mathrm{argmin}_{\mathcal{M}\in\mathbb{M}}\mathrm{BIC}(\mathcal{M})$. Comparing (3.1) against the following traditional BIC criterion (Schwarz, 1978),

$$\mathrm{BIC}^*(\mathcal{M}) = \log \mathrm{RSS}(\mathcal{M}) + |\mathcal{M}| \cdot \log n \cdot (1/n),$$

we find that the only difference lies in the last penalization factor, where the classical BIC criterion uses $1/n$ while we consider $\log p/n$. The classical BIC criterion uses $1/n$ because the uniform convergence rate of the OLS estimator (for example) is $O_p(1/\sqrt{n})$, under a fixed dimension setup (Shao, 1997). However, the story changes under an ultra high dimensional setup. By Theorem 3, the uniform convergence rate of the PIS estimator is of a considerably higher order as $O_p(\sqrt{\log p/n})$. Consequently, a heavier penalty factor is inevitable (Chen and Chen, 2008). This motivates us to replace the traditional factor $1/n$ by $\log p/n$, which leads to (3.1). Our numerical experiences, as reported in Section 4, suggests that (3.1) works fairly well.

3.3. Profiled Sequential Screening

As we mentioned earlier, the performance of PIS can be further improved, if the relevant predictor's regression effects can be sequentially removed from the response. To this end, we propose here a profiled sequential screening method (PSS). Specifically, the detailed algorithm is given below.

Step (1) (*Initialization*). Set $\mathcal{M}^*_{(0)} = \emptyset$ and $\widehat{\mathbb{Y}}^{(0)} = \widehat{\mathbb{Y}}$, i.e., the factor profiled response.

Step (2) (*Sequential Screening*).

(2.1) (*Estimation*). In the kth step ($k \geq 1$), we are given $\mathcal{M}^*_{(k-1)}$ and also $\widehat{\mathbb{Y}}^{(k-1)}$. Then, for every $j \in \mathcal{M}_F \backslash \mathcal{M}^*_{(k-1)}$, estimate its regression coefficient as $\hat{\theta}^{(k)}_j = \{\widehat{\mathbb{Y}}^{(k-1)\top}\widehat{\mathbb{X}}_j\}/\|\widehat{\mathbb{X}}_j\|^2$ and its correlation coefficient with the response as $\hat{\zeta}^{(k)}_j = \{\widehat{\mathbb{Y}}^{(k-1)\top}\widehat{\mathbb{X}}_j\}/\{\|\widehat{\mathbb{Y}}^{(k-1)}\| \cdot \|\widehat{\mathbb{X}}_j\|\}$.

(2.2) (*Screening*). We then find $a_k = \text{argmax}_{j \in \mathcal{M}_F \backslash \mathcal{S}^{(k-1)}} |\hat{\zeta}^{(k)}_j|$ and update $\mathcal{M}^*_{(k)} = \mathcal{M}^*_{(k-1)} \bigcup \{a_k\}$ accordingly.

(2.3) (*Elimination*). According to a_k, we then get an updated response vector as $\widehat{\mathbb{Y}}^{(k)} = \widehat{\mathbb{Y}}^{(k-1)} - \hat{\theta}^{(k)}_j \widehat{\mathbb{X}}_j$ with $j = a_k$.

Step (3) (*Solution Path*). Iterating Step (2) for a total of n times, which leads a total of $n+1$ nested candidate models. We then collect those models by a solution path $\mathbb{M}^* = \{\mathcal{M}^*_{(k)} : 0 \leq k \leq n\}$ with $\mathcal{M}^*_{(k)} = \{a_1, \cdots, a_k\}$ for $k > 0$.

Step (4) (*Model Selection*). Select the best model as $\widehat{\mathcal{M}}^* = \text{argmin}_{\mathcal{M} \in \mathbb{M}^*} \text{BIC}(\mathcal{M})$.

As one can see, PSS shares similar computational flavor as FR (Wang, 2009a), in the sense that the effects of relevant predictors are sequentially profiled out from the response. However, the difference is that, for FR, the same also needs to be done for every remaining predictor. Otherwise, they might still correlate with relevant ones seriously. Fortunately, for PSS this is an unnecessary step, because the factor profiled predictors are already uncorrelated. That explains why PSS is computationally even simpler than FR. In fact, we find that the finite sample performance of PSS could be considerably better than that of PIS.

4. NUMERICAL STUDIES

To evaluate the finite sample performance of the proposed methods, we present here three simulation experiments and one real example.

4.1. Different Simulation Models

Example 1. This is an example borrowed from Fan and Lv (2008). Specifically, we fix $d_0 = 1$, $p = 5000$, and $n = 150$. Z_i is generated from $N(0, 1)$. X_i is then simulated as (2.2), where $b_{jk} = 1$ and $\widetilde{X}_i$ follows a p-dimensional standard normal distribution. Following Fan and Lv (2008), we assume the first $|\mathcal{M}_T| = 3$ predictors to be relevant and their coefficients are given by $\theta_{0j} = 5$ for $1 \leq j \leq |\mathcal{M}_T|$. Accordingly, $\theta_{0j} = 0$ for every $j > |\mathcal{M}_T|$. Subsequently, Y_i is given by (2.1), where ε_i follows (2.3) with

$\alpha_0 = 0.8\sigma_\varepsilon$ and $\tilde{\sigma}_\varepsilon = 0.6\sigma_\varepsilon$. Lastly, σ_ε^2 is particularly selected so that the signal-to-noise ratio, i.e., SNR=$\text{var}(X_i^\top \theta_0)/\sigma_\varepsilon^2$, is given by 1, 2, or 5.

Example 2. This is another example revised from Fan and Lv (2008) but with a more sophisticated factor structure. For this example, we have $d_0 = 2$, $p = 10000$, and $n = 400$. $Z_i \in \mathbb{R}^2$ is generated from a bivariate standard normal distribution. X_i is then simulated as (2.2), where both b_{jk} and $\widetilde{X}_{ij}$ are independent and identically distributed as $N(0,1)$. Subsequently, Y_i is given by (2.1), where $\theta_{0j} = (-1)^{R_{aj}}(4\log n/\sqrt{n} + |R_{bj}|)$ for $1 \leq j \leq |\mathcal{M}_T| = 8$ and $\theta_{0j} = 0$ for every $j > |\mathcal{M}_T|$. Here, R_{aj} is a binary random variable with $P(R_{aj} = 1) = 0.4$ and R_{bj} is another $N(0,1)$ variable. Lastly, ε_i is generated according to (2.3) with $\alpha_0 = 0.8\sigma_\varepsilon(1/\sqrt{2}, 1/\sqrt{2})^\top \in \mathbb{R}^2$ and $\tilde{\sigma}_\varepsilon = 0.6\sigma_\varepsilon$. SNR is 1, 2, or 5.

Example 3. This is an example modified from Tibshirani (1996). We have here $d_0 = 3$, $p = 10000$, and $n = 300$. $Z_i \in \mathbb{R}^3$ is generated from a 3-dimensional standard normal random vector. X_i is then simulated as (2.2), where b_{jk} follows $N(0,1)$. However, $\widetilde{X}_i$ follows a p-dimensional normal distribution with $E(\widetilde{X}_{ij}) = 0$ and $\text{cov}(\widetilde{X}_{ij_1}, \widetilde{X}_{ij_2}) = 0.5^{|j_1 - j_2|}$. Subsequently, Y_i is given by (2.1) with $\theta_{01} = 3$, $\theta_{04} = 1.5$, $\theta_{07} = 2$, and $\theta_{0j} = 0$ for any $j \notin \mathcal{M}_T = \{1, 4, 7\}$. Lastly, ε_i is generated according to (2.3) with $\alpha_0 = 0.8\sigma_\varepsilon(1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3})^\top \in \mathbb{R}^3$ and $\tilde{\sigma}_\varepsilon = 0.6\sigma_\varepsilon$. Again, SNR=1, 2, or 5. For this example, the factor model (2.2) is not accurately satisfied, because $\text{cov}(\widetilde{X}_i)$ is not diagonal. Thus, by including this example in our study, we are able to evaluate the sensitivity of the proposed FP methods towards certain model mis-specification, for the factor model (2.2).

4.2. Factor Estimation

For each simulation model, a total of 200 random replications are conducted. For each replication, MERC is used to estimate the structure dimension. We find that the

percentage of the experiments with $\hat{d} = d_0$ is always 100%. Obviously, such an excellent performance is achieved at the sample sizes and predictor dimensions as specified in the previous subsection. Reducing either the sample size and/or predictor dimension should lead to less favorable performance, which is as expected. Simply speaking, our unreported simulation experiments suggest that MERC's performance steadily increases as the sample size and/or predictor dimension increases. All those experiences confirm that MERC is indeed consistent for factor dimension estimation, which corroborates Theorem 1 very well.

In addition to that, for each simulated dataset, we also get the estimated factor subspace $\mathcal{S}(\widehat{\mathbb{Z}})$. Following Xia (2007) and Wang and Xia (2008), we then quantify its estimation error by $D(\mathbb{Z}, \widehat{\mathbb{Z}}) = \lambda_{\max}\{H(\mathbb{Z}) - H(\widehat{\mathbb{Z}})\}$, where $\lambda_{\max}(A)$ stands for the maximal absolute singular value of an arbitrary matrix A. The average values of $D(\mathbb{Z}, \widehat{\mathbb{Z}})$ across the 200 simulation replications are given by, 1.42% for Example 1, 1.05% for Example 2, and 1.41% for Example 3, respectively. Those are very small numbers as compared with similar measures as reported in the past literature; see for example Xia (2007) and Wang and Xia (2008). We thus know that $\mathcal{S}(\widehat{\mathbb{Z}})$ can indeed capture the factor subspace $\mathcal{S}(\mathbb{Z})$ satisfactorily, which corroborates Theorem 2 very well. Qualitatively similar findings are also obtained for both $D_1(\mathbb{Z}, \widehat{\mathbb{Z}})$ and $D_2(\mathbb{Z}, \widehat{\mathbb{Z}})$, which are not further discussed to save space.

4.3. Selection Consistency

We next consider the performances of SIS, PIS, and PSS for variable selection. We use a notation $\overline{\mathcal{M}}$ to represent a model selected by one particular method (e.g., PSS), in conjunction with the proposed BIC criterion (3.1). Following Wang (2009a), we

evaluate $\overline{\mathcal{M}}$'s capability in producing sparse solutions by

$$\% \text{ of Correct Zeros} = 100\% \times \left\{ \left| (\mathcal{M}_F \backslash \overline{\mathcal{M}}) \bigcap (\mathcal{M}_F \backslash \mathcal{M}_T) \right| \right\} \left\{ \left| (\mathcal{M}_F \backslash \mathcal{M}_T) \right| \right\}^{-1}.$$

Obviously, this is only one side of $\overline{\mathcal{M}}$. A method with excellent capability in producing sparse solutions might also suffer serious underfitting effect (e.g., $\overline{\mathcal{M}} = \emptyset$). Thus, it is also important to consider

$$\% \text{ of Incorrect Zeros} = 100\% \times \left\{ \left| (\mathcal{M}_F \backslash \overline{\mathcal{M}}) \bigcap \mathcal{M}_T \right| \right\} \left| \mathcal{M}_T \right|^{-1}.$$

The average values for the percentage of correct and incorrect zeros (across the 200 random replications) are reported respectively in the third and fourth columns in Table 1. We define $\overline{\mathcal{M}}$ to be a correctly fitted model if it is exactly the same as the true model, i.e., $\overline{\mathcal{M}} = \mathcal{M}_T$. Then, the average values of the percentage of the correct fits are reported in the fifth column of Table 1. Next to this column, we also report the average sizes of the selected models.

As one can see from Table 1, except for Example 1 with SNR=1, PIS always outperforms SIS considerably, in terms of both the percentages of incorrect zeros and correct fits. Such a result is not surprising, because by FP, the profiled predictors utilized by PIS is uncorrelated, which enables independent screening to demonstrate its best power. In addition to that, by FP, the problem of endogeneity is fixed, which is another reason. Lastly, we note that the already excellent performance of PIS can be further improved by PSS to much extend. Some times the relative improvement margin could be appreciable; see Examples 1 and 2 with signal to noise ratios being 2 and 5. All those numerical evidences suggest that PSS is an even better choice as compared with SIS and PIS.

4.4. Estimation Accuracy

Lastly, we gauge the performance of different methods in terms of their estimation accuracy. It is worthwhile to mention that we do not advocate the use of the marginal estimator (i.e., $\hat{\theta}_j$ as proposed in Section 3.1) as our final estimator. This is because marginal estimator's estimation accuracy is not optimal, as we have carefully explained in Section 3.3. We thus propose the following OLS type estimator as the final one to use. Before we introduce the detail, some notations need to be introduced. For an arbitrary candidate model $\mathcal{M}$, we use the notation $\mathbb{X}_{(\mathcal{M})} = (\mathbb{X}_j : j \in \mathcal{M}) \in \mathbb{R}^{n \times |\mathcal{M}|}$ to denote its submatrix associated with $\mathcal{M}$. Similarly, $\theta_{(\mathcal{M})} \in \mathbb{R}^{|\mathcal{M}|}$ stands for the corresponding subvector.

Specifically, for a selected model $\overline{\mathcal{M}}$ (e.g., the PSS model), we define an OLS-type estimator as $\hat{\theta}^{\overline{\mathcal{M}}} = (\hat{\theta}_1^{\overline{\mathcal{M}}}, \cdots, \hat{\theta}_p^{\overline{\mathcal{M}}})^\top \in \mathbb{R}^p$, where $\hat{\theta}_j^{\overline{\mathcal{M}}} = 0$ for every $j \notin \overline{\mathcal{M}}$ while

$$\hat{\theta}_{(\overline{\mathcal{M}})}^{\overline{\mathcal{M}}} = \Big(\widehat{\mathbb{X}}_{(\overline{\mathcal{M}})}^\top \widehat{\mathbb{X}}_{(\overline{\mathcal{M}})}\Big)^{-1} \Big(\widehat{\mathbb{X}}_{(\overline{\mathcal{M}})}^\top \widehat{\mathbb{Y}}\Big). \tag{4.1}$$

Simply speaking, $\hat{\theta}^{\overline{\mathcal{M}}}$ is a p-dimensional vector. Its elements associated with irrelevant predictors (i.e., $j \notin \overline{\mathcal{M}}$) are fixed to be 0. On the other hand, its elements associated with relevant predictors (i.e., $j \in \overline{\mathcal{M}}$) are fixed to be the OLS estimator $\hat{\theta}_{(\overline{\mathcal{M}})}^{\overline{\mathcal{M}}}$. The OLS estimator $\hat{\theta}_{(\overline{\mathcal{M}})}^{\overline{\mathcal{M}}}$ is computed based on either the profiled data $(\widehat{\mathbb{X}}_{(\overline{\mathcal{M}})}, \widehat{\mathbb{Y}})$ for PIS and PSS, or the non-profiled data $(\mathbb{X}_{(\overline{\mathcal{M}})}, \mathbb{Y})$ for SIS. Subsequently, we can evaluate its estimation accuracy by the absolute estimation error $\sum_{j=1}^p |\theta_{0j} - \hat{\theta}_j^{\overline{\mathcal{M}}}|$, whose average values are summarized in the last column of Table 1.

By Table 1, we find that the absolute estimation error of SIS is always the worst. Such a result is not surprising because SIS dose not have the capability to fix the endogeneity problem. This leads to poor model selection accuracy and thus unsatisfactory

estimation accuracy. In contrast, both PIS and PSS are free of such an issue. They solve the problem by FP. As a result, their estimation accuracy is considerably better than that of SIS. Comparatively speaking, PSS is even better, because its is more accurate in terms of variable selection.

4.5. A Real Example

To conclude our numerical study, we present here a real example. Specifically, this is a dataset donated by a domestic supermarket located in northern China (Wang, 2009a). It contains a total of $n = 464$ daily records, where the response is the number of customers and the predictors are the sales volumes for a total of $p = 6398$ products. Prior to the formal analysis, both the response and predictors are log-transformed and then further standardized to have zero mean and unit variance.

As our first step, we need to estimate the dimension of the latent factor. We find that the first eigenvalue of the matrix $\mathbb{X}\mathbb{X}^\top/(np)$ is as large as $\hat{\lambda}_1 = 35.4\%$ while the second one is as small as $\hat{\lambda}_2 = 3.5\%$. The big difference as demonstrated between $\hat{\lambda}_1$ and $\hat{\lambda}_2$ suggests that the true factor dimension might be $d_0 = 1$. Such a conjecture is formally confirmed by MERC. We then fix $d = 1$ throughout the rest of this example. Thereafter, the factor subspace $\mathcal{S}(\widehat{\mathbb{Z}})$ can be estimated and the profiled data $(\widehat{\mathbb{Y}}, \widehat{\mathbb{X}})$ can be produced.

For a real problem like this, the value of θ_0 is unknown. We thus have to rely on out-of-sample testing to compare different methods' estimation and/or prediction accuracy. We then conducted a total of 200 random experiments. For each experiment, we randomly split the entire dataset $\mathcal{D} = \{1, \cdots, 464\}$ into two parts. That is $\mathcal{D} = \mathcal{D}_0 \bigcup \mathcal{D}_1$ with $|\mathcal{D}_0| = n_0 = 400$ as the training data and $|\mathcal{D}_1| = n_1 = 64$ as the testing data. Accordingly, we write $\mathbb{X}_0 = \{X_i : i \in \mathcal{D}_0\} \in \mathbb{R}^{n_0 \times p}$, $\mathbb{Y}_0 = \{Y_i : i \in \mathcal{D}_0\} \in \mathbb{R}^{n_0}$,

$\mathbb{X}_1 = \{X_i : i \in \mathcal{D}_1\} \in \mathbb{R}^{n_1 \times p}$, and $\mathbb{Y}_1 = \{Y_i : i \in \mathcal{D}_1\} \in \mathbb{R}^{n_1}$. Notations for $(\widehat{\mathbb{X}}_0, \widehat{\mathbb{X}}_1)$, $(\widehat{\mathbb{Y}}_0, \widehat{\mathbb{Y}}_1)$, and $(\widehat{\mathbb{Z}}_0, \widehat{\mathbb{Z}}_1)$ are defined accordingly.

We apply SIS together with the BIC criterion (3.1) to $(\mathbb{X}_0, \mathbb{Y}_0)$, which produces a candidate model denoted by $\mathcal{M}_{\text{SIS}}$. With the help of $\mathcal{M}_{\text{SIS}}$, we then obtain the OLS estimator $\hat{\theta}^{\mathcal{M}_{\text{SIS}}}$; see (4.1). Similar estimators are also obtained for PIS and PSS but based on an expanded design matrix, where $\widetilde{\mathbb{Z}}_0$ is also included as an additional predictor. Thereafter, the responses in the testing data are predicted, and their median squared prediction error (MSPE) are summarized. Lastly, each method's MSPE values are boxplotted in Figure 1. Obviously, PSS performs the best while SIS is the worst. Across all the experiments, the sizes of both SIS and PIS models are always one, while the size of the PSS model is 3 for 94% of the random experiments.

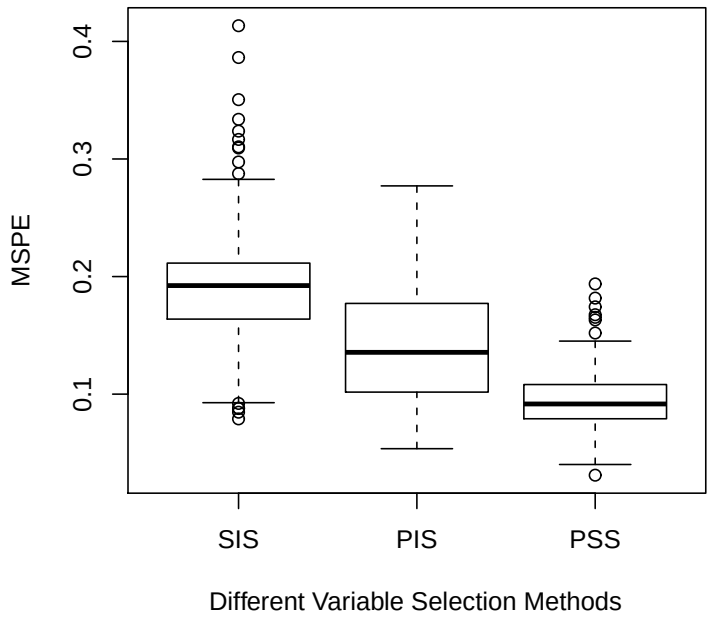

Figure 1: The real supermarket example. Boxplots for the median squared prediction errors (MSPE) based on 200 random replications.

5. CONCLUDING REMARKS

We note that the idea of FP can be viewed as a non-supervised dimension reduction technique, in the sense that the response is completely ignored for factor estimation. As a result, the factors identified by FP might be good for describing a predictor's correlation structure but suboptimal for explaining the response; see for example Li (1991), Cook (1998), Li et al. (2007) and Zhu and Zhu (2009) for some discussions. Although such a phenomenon never happens to our real data example, but there indeed exists such a possibility, at least theoretically. Then, developing a supervised FP method is an interesting future direction.

As one can see, many excellent convergence results have been documented for various estimation methods, under an ultra high dimensional setup. In contrast, little has been obtained for their asymptotic distributions. In a recent work of Chen and Qin (2010), the classical problem of two-sample test has been re-investigated for high dimensional data. Then, whether similar testing procedure can be developed with FP is of great interest (Friguet et al., 2009).

We also note that many research efforts on ultra high dimensional data analysis have been focusing on linear regression. Nevertheless, much limited has been done for other regression models, which include generalized linear models (Fan and Song, 2009) and various semiparametric models (Härdle et al., 2000). The underlying low dimensional factor structure, as described by (2.2), should provide a natural bridge to connect those powerful low-dimensional regression ideas with various ultra high dimensional problems (Wang, 2009b).

To conclude the article, we point out that our experience with FP is still preliminary but rather encouraging. Given the fact that clear factor structures have been witnessed for many ultra high dimensional datasets, we conjecture FP's wide applicability across

a large number of model classes. Further research along this direction is necessary, promising, and also very exciting!

REFERENCES

Akaike, H. (1973), "Information theory and an extension of the maximum likelihood principle," *In 2nd International Symposium on Information Theory, Ed. B. N. Petrov & F. Csaki*, 267–281. Budapest: Akademia Kiado.

Bickel, P. J. and Levina, E. (2008), "Regularized estimation of large covariance matrix," *The Annals of Statistics*, 36, 199–277.

Breiman, L. (1995), "Better subset selection using nonnegative garrote," *Technometrics*, 37, 373–384.

— (1996), "Heuristics of instability and stabilization in model selection," *The Annals of Statistics*, 24, 2350–2383.

Chen, J. and Chen, Z. (2008), "Extended Bayesian information criterion for model selection with large model spaces," *Biometrika*, 95, 759–771.

Chen, S. X. and Qin, Y. L. (2010), "A two sample test for high dimensional data with application to gene-set testing," *The Annals of Statistics*, To appear.

Cook, R. D. (1998), *Regression Graphics*, John Wiley: New York.

Fan, J., Fan, Y., and Lv, J. (2008), "High dimensional covariance matrix estimation using a factor model," *Journal of Econometrics*, 147, 186–197.

Fan, J. and Li, R. (2001), "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, 96, 1348–1360.

— (2006), "Statistical Challenges with High Dimensionality: Feature Selection in Knowledge Discovery," *Proceedings of the International Congress of Mathematicians (M. Sanz-Sole, J. Soria, J.L. Varona, J. Verdera, eds.), Vol. III, European Mathematical Society, Zurich*, 595–622.

Fan, J. and Lv, J. (2008), "Sure independence screening for ultra-high dimensional feature space (with discussion)," *Journal of the Royal Statistical Society, Series B*, 70, 849–911.

Fan, J. and Peng, H. (2004), "On non-concave penalized likelihood with diverging number of parameters," *The Annals of Statistics*, 32, 928–961.

Fan, J. and Song, R. (2009), "Sure independent screening in generalized linear models with NP-dimensionality," *The Annals of Statistics*, To appear.

Friguet, C., Kloareg, M., and Causeur, D. (2009), "A factor model approach to multiple testing under dependence," *Journal of the American Statistical Association*, 104, 1406–1415.

Fu, W. J. (1998), "Penalized regression: the bridge versus the LASSO," *Journal of Computational and Graphical Statistics*, 7, 397–416.

Härdle, W., Liang, H., and Gao, J. (2000), *Partially Linear Models*, Heidelberg: Springer.

Hjort, N. L. and Claeskens, G. (2003), "Frequentist model average estimators (with discussion)," *Journal of the American Statistical Association*, 98, 879–899.

Huang, J., Horowitz, J., and Ma, S. (2007), "Asymptotic properties of bridge estimators in sparse high-dimensional regression models," *The Annals of Statistics*, 36, 587–613.

Knight, K. and Fu, W. (2000), "Asymptotics for lasso-type estimators," *The Annals of Statistics*, 28, 1356–1378.

Leng, C., Lin, Y., and Wahba, G. (2006), "A note on lasso and related procedures in model selection," *Statistica Sinica*, 16, 1273–1284.

Li, K.-C. (1991), "Sliced inverse regression for dimension reduction," *Journal of the American Statistical Association*, 86, 316–327.

Li, L., Cook, R. D., and Tsai, C. L. (2007), "Partial inverse regression," *Biometrika*, 94, 615–625.

Luo, R., Wang, H., and Tsai, C. L. (2009), "Contour projected dimension reduction," *The Annals of Statistics*, To appear.

Schwarz, G. (1978), "Estimating the dimension of a model," *The Annals of Statistics*, 6, 461–464.

Shao, J. (1997), "An asymptotic theory for linear model selection," *Statistica Sinica*, 7, 221–264.

Shi, P. and Tsai, C. L. (2002), "Regression model selection - a residual likelihood approach," *Journal of Royal Statistical Society, Series B*, 64, 237–252.

Tibshirani, R. J. (1996), "Regression shrinkage and selection via the LASSO," *Journal of the Royal Statistical Society, Series B*, 58, 267–288.

Wang, H. (2009a), "Forward regression for ultra-high dimensional variable screening," *Journal of the American Statistical Association*, 104, 1512–1524.

— (2009b), "Rank reducible varying coefficient model," *Journal of Statistical Planning and Inference*, 139, 999–1011.

Wang, H. and Leng, C. (2007), "Unified lasso estimation via least squares approximation," *Journal of the American Statistical Association*, 101, 1418–1429.

Wang, H., Li, B., and Leng, C. (2009), "Shrinkage tuning parameter selection with a diverging number of parameters," *Journal of Royal Statistical Society, Series B*, 71, 671–683.

Wang, H., Li, R., and Tsai, C. L. (2007), "Tuning parameter selectors for the smoothly clipped absolute deviation method," *Biometrika*, 94, 553–558.

Wang, H. and Xia, Y. (2008), "Sliced regression for dimension reduction," *Journal of the American Statistical Associate*, 103, 811–821.

Wooldridge, J. M. (2001), *Econometric Analysis of Cross Section and Panel Data*, The MIT Press, Cambridge, Massachusetts.

Xia, Y. (2007), "A constructive approach to the estimation of dimension reduction directions," *The Annals of Statistics*, 35, 2654–2690.

Yuan, M. and Lin, Y. (2007), "On the nonnegative garrote estimator," *Journal of the Royal Statistical Society, Series B*, 69, 143–161.

Zhang, C. H. and Huang, J. (2008), "The sparsity and bias of the lasso selection in high-dimensional linear regression," *The Annals of Statistics*, 36, 1567–1594.

Zhang, H. H. and Lu, W. (2007), "Adaptive lasso for Cox's proportional hazard model," *Biometrika*, 94, 691–703.

Zhao, P. and Yu, B. (2006), "On model selection consistency of lasso," *Journal of Machine Learning Research*, 7, 2541–2567.

Zhu, L. P. and Zhu, L. X. (2009), "On distribution-weighted partial least squares with diverging number of highly correlated predictors," *Journal of the Royal Statistical Society, Series B*, 71, 525–548.

Zou, H. (2006), "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, 101, 1418–1429.

Zou, H. and Hastie, T. (2005), "Regression shrinkage and selection via the elastic net with application to microarrays," *Journal of the Royal Statistical Society, Series B*, 67, 301–320.

Zou, H. and Li, R. (2008), "One-step sparse estimates in nonconcave penalized likelihood models (with discussion)," *The Annals of Statistics*, 36, 1509–1533.

Zou, H. and Zhang, H. H. (2009), "On the adaptive elastic-net with a diverging number of parameters," *The Annals of Statistics*, 37, 1733–1751.

Table 1: Detailed Simulation Results

Signal Noise Ratio	Variable Selection Method	% of Correct Zeros	% of Incorrect Zeros	% of Correct fit	Average Model Size	Absolute Estimation Error
			Example 1			
1	SIS	100.0	77.2	0.0	1.0	25.4
	PIS	100.0	95.8	0.5	0.1	14.6
	PSS	100.0	95.8	0.5	0.1	14.6
2	SIS	100.0	70.3	0.0	1.0	21.3
	PIS	100.0	46.3	40.0	1.6	7.9
	PSS	100.0	43.3	45.5	1.7	7.4
5	SIS	100.0	67.0	0.0	1.0	18.4
	PIS	100.0	0.2	99.5	3.0	1.0
	PSS	100.0	0.0	100.0	3.0	0.9
			Example 2			
1	SIS	100.0	95.9	0.0	1.0	17.6
	PIS	100.0	75.6	1.0	1.9	11.5
	PSS	100.0	73.6	2.0	2.1	11.2
2	SIS	100.0	93.2	0.0	1.0	16.6
	PIS	100.0	47.4	15.0	4.2	7.2
	PSS	100.0	41.5	30.0	4.7	6.5
5	SIS	100.0	90.8	0.0	1.1	15.9
	PIS	100.0	16.6	34.5	6.7	2.7
	PSS	100.0	6.3	80.5	7.5	1.5
			Example 3			
1	SIS	100.0	92.2	0.0	1.0	8.4
	PIS	100.0	53.5	9.5	1.4	3.4
	PSS	100.0	49.5	11.5	1.5	3.1
2	SIS	100.0	88.0	1.0	1.0	7.7
	PIS	100.0	34.5	24.5	2.0	2.2
	PSS	100.0	23.7	44.0	2.3	1.6
5	SIS	100.0	81.8	1.0	1.1	6.8
	PIS	100.0	20.2	50.5	2.4	1.4
	PSS	100.0	5.3	85.5	2.8	0.6

APPENDIX

Appendix A. Technical Conditions

To gain theoretical insights about the proposed FP methods and also facilitate an easy proof, the following conditions are needed.

(A1). *(Normality Assumption)* Assume that there exists a specification for the factor model (2.2), such that both the latent factor $\mathbb{Z}$ and the profiled predictor $\widetilde{\mathbb{X}}$ are normally distributed. Furthermore, we assume that there exists a positive constant $\tilde{\sigma}^2_{\min} > 0$ such that $\min_{1\le j\le p} \tilde{\sigma}^2_j \ge \tilde{\sigma}^2_{\min}$.

(A2). *(Law of Large Numbers)* Assume both the factor loading β_j and the variance $\tilde{\sigma}^2_j$ admit the following types of large number laws, that is $p^{-1}B^\top B = p^{-1}\sum \beta_j\beta_j^\top = \Sigma_\beta + O_p(p^{-1/2})$ and $p^{-1}\sum \tilde{\sigma}^2_j = \tilde{\sigma}^2_0 + O_p(p^{-1/2})$, where $\Sigma_\beta \in \mathbb{R}^{d\times d}$ is a positive definite matrix and $\tilde{\sigma}^2_0 \ge \tilde{\sigma}^2_{\min} > 0$ is a positive constant.

(A3). *(Predictor Dimension)* Assume both the factor dimension d and the true model size $|\mathcal{M}_T|$ are fixed while the sample size $n \to \infty$. Moreover, we assume that $\xi_{\min} \le n^{-\hbar}\log p \le \xi_{\max}$ for some $0 < \xi_{\min} \le \xi_{\max} < \infty$ and $0 < \hbar < 1$.

Note that the normality assumptions similar to (A1) have been popularly assumed in the past literature to facilitate an easy theoretical proof; see for example Fan and Lv (2008), Zhang and Huang (2008), Bickel and Levina (2008), and Wang (2009a). In particular, it implies the following exponential inequalities

$$P\left(\left|p^{-1}\sum_{j=1}^{p}\left\{\widetilde{X}_{i_1j}\widetilde{X}_{i_2j} - E\big(\widetilde{X}_{i_1j}\widetilde{X}_{i_2j}\big)\right\}\right| > \nu\right) \le C_1\exp(-C_2 p\nu^2) \tag{A4.a}$$

$$P\left(\left|n^{-1}\sum_{i=1}^{n}\widetilde{Y}_i\widetilde{X}_{ij} - E(\widetilde{Y}_i\widetilde{X}_{ij})\right| > \nu\right) \le C_1\exp(-C_2 n\nu^2) \tag{A4.b}$$

$$P\left(\left|n^{-1}\sum_{i=1}^{n}\widetilde{X}_{ij}^2-\tilde{\sigma}_j^2\right|>\nu\right)\leq C_1\exp(-C_2n\nu^2), \tag{A4.c}$$

which play the key roles in the theoretical treatment of essentially any type of ultra high dimensional problems; see the Lemma A.3 in Bickel and Levina (2008). The condition (A2) is also reasonable. It can be trivially satisfied if (β_j, σ_j)s for different $1 \leq j \leq p$ are generated independently from some distribution with finite moments. Lastly, by (A3), we require that the predictor dimension p to be much larger than the sample size. Nevertheless, (A3) also constraints that $\log p/n \to 0$. In addition to that, (A3) also requires a fixed true model size $|\mathcal{M}_T|$. In fact, we can allow $|\mathcal{M}_T| \to \infty$, as long as its diverging speed is sufficiently slow; see for example Fan and Lv (2008) and Wang (2009a). We decide to make here a slightly stronger assumption. Otherwise, our already lengthy proof would be even more complicated.

Appendix B. Important Lemmas

The following lemmas are useful in the subsequent theorem proofs. We thus present them firstly. For convenience, the following notations need to be defined. For an arbitrary matrix $d_1 \times d_2$ matrix $A \in \mathbb{R}^{d_1 \times d_2}$, we use $\lambda_{\min}(A)$ to denote its minimal absolute singular value. Recall $\lambda_{\max}(A)$ is its maximal absolute singular value. Furthermore, we define a matrix norm as $\|A\|^2 = tr(A^\top A) = tr(AA^\top)$.

Lemma L1. *Let $A_1 \in \mathbb{R}^{q\times q}$ and $A_2 \in \mathbb{R}^{q\times q}$ be two arbitrary semi positive definite matrices with some $q > 0$. We then have $tr(A_1A_2) \geq \lambda_{\min}(A_1)tr(A_2)$.*

Proof. Because A_2 is a semi positive definite matrix, its root matrix $A_2^{1/2}$ is well defined. Let (τ_j, W_j) be the jth eigenvalue-eigenvector pair for $A_2^{1/2}A_1A_2^{1/2}$. Then,

$$tr(A_1A_2) = tr(A_2^{1/2}A_1A_2^{1/2}) = \sum_{j=1}^{q}\tau_j = \sum_{j=1}^{q}W_j^\top\left(A_2^{1/2}A_1A_2^{1/2}\right)W_j$$

$$= \sum_{j=1}^{q} \Big(W_j^\top A_2^{1/2} \Big) A_1 \Big(A_2^{1/2} W_j \Big) \geq \lambda_{\min}(A_1) \sum_{j=1}^{q} W_j^\top A_2 W_j$$

$$= \lambda_{\min}(A_1) tr\Big\{ A_2 \sum_{j=1}^{q} W_j W_j^\top \Big\} = \lambda_{\min}(A_1) \cdot tr(A_2),$$

because $\sum W_j W_j^\top = I$. This completes the proof.

Lemma L2. *Let $A_1 \in \mathbb{R}^{q\times q}$ and $A_2 \in \mathbb{R}^{q\times q}$ be two arbitrary matrices, where A_2 is semi positive definite. We then have $tr(A_1A_2) \leq \lambda_{\max}(A_1)tr(A_2) \leq tr(A_1)tr(A_2)$.*

PROOF. Following the same notation and similar steps as in Lemma L1, we can prove that $tr(A_1A_2) \leq \lambda_{\max}(A_1)tr(A_2) \leq tr(A_1)tr(A_2)$, where the last inequality is due to the fact $\lambda_{\max}(A_1) \leq tr(A_1)$. This completes the proof.

Lemma L3. *Assume conditions (A1)–(A3), then with probability tending to one, we should have $2^{-1}\tilde{\sigma}^2_{\min} \leq \lambda_{\min}(p^{-1}\widetilde{\mathbb{X}}\widetilde{\mathbb{X}}^\top) \leq \lambda_{\max}(p^{-1}\widetilde{\mathbb{X}}\widetilde{\mathbb{X}}^\top) \leq 2$.*

PROOF. Note that $E(p^{-1}\widetilde{\mathbb{X}}\widetilde{\mathbb{X}}^\top) = I\bar{\sigma}^2$, where $\bar{\sigma}^2 = p^{-1}\sum_{j=1}^{p} \tilde{\sigma}_j^2 \to_p \tilde{\sigma}_0^2$ and $\tilde{\sigma}^2_{\min} \leq \tilde{\sigma}_0^2 \leq \text{var}(X_{ij}) = 1$. Furthermore, note that

$$\lambda_{\max}(p^{-1}\widetilde{\mathbb{X}}\widetilde{\mathbb{X}}^\top) \leq \tilde{\sigma}_0^2 + \lambda_{\max}(p^{-1}\widetilde{\mathbb{X}}\widetilde{\mathbb{X}}^\top - \bar{\sigma}^2 I) + o_p(1)$$

$$\lambda_{\min}(p^{-1}\widetilde{\mathbb{X}}\widetilde{\mathbb{X}}^\top) \geq \tilde{\sigma}_0^2 - \lambda_{\max}(p^{-1}\widetilde{\mathbb{X}}\widetilde{\mathbb{X}}^\top - \bar{\sigma}^2 I) + o_p(1).$$

Thus, the conclusion follows if we can prove that $\Delta = \lambda_{\max}(p^{-1}\widetilde{\mathbb{X}}\widetilde{\mathbb{X}}^\top - \bar{\sigma}^2 I) = o_p(1)$. By definition, we have

$$\Delta = \sup_{\eta\in\mathbb{R}^n:\|\eta\|=1} \left| \eta^\top \Big(p^{-1}\widetilde{\mathbb{X}}\widetilde{\mathbb{X}}^\top - \bar{\sigma}^2 I \Big) \eta \right| \leq \sup_{\eta\in\mathbb{R}^n:\|\eta\|=1} \sum_{i_1=1}^{n}\sum_{i_2=1}^{n} \eta_{i_1}\eta_{i_2} \Big| \hat{\delta}_{i_1 i_2} - \delta_{i_1 i_2} \Big|,$$

where $\eta = (\eta_1, \cdots, \eta_n)^\top \in \mathbb{R}^n$, $\hat{\delta}_{i_1 i_2} = p^{-1}\sum_{j=1}^{p} \widetilde{X}_{i_1 j}\widetilde{X}_{i_2 j}$, and $\delta_{i_1 i_2} = \bar{\sigma}^2$ if $i_1 = i_2$ or 0

if $i_1 \neq i_2$. We can further bound the above quantity by

$$\leq \left(\sup_{\eta \in \mathbb{R}^n : \|\eta\|=1} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \eta_{i_1} \eta_{i_2} \right) \left(\max_{1 \leq i_1, i_2 \leq n} \left| \hat{\delta}_{i_1 i_2} - \delta_{i_1 i_2} \right| \right)$$

$$= \sup_{\eta \in \mathbb{R}^n : \|\eta\|=1} \left(\sum_{i=1}^{n} \eta_i \right)^2 \left(\max_{1 \leq i_1, i_2 \leq n} \left| \hat{\delta}_{i_1 i_2} - \delta_{i_1 i_2} \right| \right)$$

$$\leq n \cdot \max_{1 \leq i_1, i_2 \leq n} \left| \hat{\delta}_{i_1 i_2} - \delta_{i_1 i_2} \right|. \tag{B.1}$$

Next, by Bonferroni's inequality, we have

$$P\left(n \cdot \max_{1 \leq i_1, i_2 \leq n} \left| \hat{\delta}_{i_1 i_2} - \delta_{i_1 i_2} \right| > \nu \right) \leq \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} P\Big(\left| \hat{\delta}_{i_1 i_2} - \delta_{i_1 i_2} \right| > n^{-1} \nu \Big)$$

$$\leq n^2 \cdot C_1 \exp\big(- C_2 p n^{-2} \nu^2 \big) = C_1 \exp\big(- C_2 p n^{-2} \nu^2 + 2 \log n \big), \tag{B.2}$$

where the first inequality in (B.2) is due to the exponential inequality (A4.a) as implied by (A1). Next, by (A3) one can verify that $-C_2 p n^{-2} \nu^2 + 2 \log n \to -\infty$, which implies that the right hand side of (B.2) converges to 0. Combining such a result with (B.2), we find that $\Delta = o_p(1)$. This completes the proof.

Lemma L4. *Let $H \in \mathbb{R}^{q \times q}$ be an arbitrary projection matrix with rank K and $A \in \mathbb{R}^{q \times q}$ be another arbitrary square matrix. We then have $|tr(HA)| \leq K \lambda_{\max}(A)$.*

Proof. Because H is a projection matrix with rank K, we should find a set of orthonormal basis $\{W_j : 1 \leq j \leq K\}$ such that $H = \sum W_j W_j^\top$. We then have $|tr(HA)| = |\sum_j W_j W_j^\top A| \leq \sum_{j=1}^{K} |W_j^\top A W_j| \leq K \lambda_{\max}(A)$. This completes the proof.

Lemma L5. *Assume conditions (A1)–(A3), we have $\lambda_{\max}(p^{-1} \widetilde{\mathbb{X}} B \mathbb{Z}^\top) \to_p 0$.*

Proof. Note that $p^{-1} \widetilde{\mathbb{X}} B \mathbb{Z}^\top$ is a $n \times n$ matrix. Denote its (i_1, i_2)th element by $\hat{\tau}_{i_1 i_2}$. Obviously, $E(\hat{\tau}_{i_1 i_2}) = 0$ because $\widetilde{\mathbb{X}}$ and $\mathbb{Z}$ are mutually independent. Following similar

steps as for (B.1), we can define

$$\Omega = \lambda_{\max}\Big(p^{-1}\widetilde{\mathbb{X}}B\mathbb{Z}^\top\Big) \le n \max_{1\le i_1,i_2\le n} |\hat{\tau}_{i_1 i_2}|. \tag{B.3}$$

We next find a uniform bound for $|\hat{\tau}_{i_1 i_2}|$. Specifically, note that

$$\max_{1\le i_1,i_2\le n} |\hat{\tau}_{i_1 i_2}| = \max_{1\le i_1,i_2\le n} \Big|p^{-1}\widetilde{X}_{i_1}^\top B Z_{i_2}\Big| = \max_{1\le i_1,i_2\le n} \left|p^{-1}\sum_{j=1}^{p}\widetilde{X}_{i_1 j}\left(\sum_{k=1}^{d} b_{jk}Z_{i_2 k}\right)\right|$$

$$= \max_{1\le i_1,i_2\le n} \left|\sum_{k=1}^{d} Z_{i_2 k}\left(p^{-1}\sum_{j=1}^{p}\widetilde{X}_{i_1 j}b_{jk}\right)\right|$$

$$\le d\Big(\max_{i,k} Z_{ik}^2\Big)^{1/2}\cdot p^{-1/2}\cdot \max_{1\le k\le d}\max_{1\le i\le n}\Big|p^{-1/2}\sum_{j=1}^{p}\widetilde{X}_{ij}b_{jk}\Big|.$$

By (A1), we know that $\{Z_{ik}^2\}$ with $1 \le i \le n$ and $1 \le k \le d$ constitutes a total of nd $\chi^2(1)$ random variables, where $\chi^2(1)$ stands for a chi-square distribution with 1 degree of freedom. Following similar technique as in Wang et al. (2009), we can prove that $\max_{ik} Z_{ik}^2 \le 2\log(nd)$, with probability tending to one. Thus, the right hand side of the above inequality is bounded by

$$\le d\sqrt{2\log(nd)}\cdot p^{-1/2}\cdot \max_{1\le k\le d}\max_{1\le i\le n}\Big|p^{-1/2}\tilde{\sigma}_{ib}^{-1}\sum_{j=1}^{p}\widetilde{X}_{ij}b_{jk}\Big| \tag{B.4}$$

$$= d\sqrt{2\log(nd)}\cdot p^{-1/2}\cdot\Big\{\max_{k,i}\chi^2(1)\Big\}^{1/2} \le d\sqrt{2\log(nd)}\cdot p^{-1/2}\cdot\sqrt{2\log(nd)}, \tag{B.5}$$

where the inequality (B.4) is due to the fact that $\tilde{\sigma}_{ib}^2 = \mathrm{var}(p^{-1/2}\sum_{j=1}^{p}\widetilde{X}_{ij}b_{jk}) = p^{-1}\sum_{j=1}^{p}\tilde{\sigma}_j^2 b_{jk}^2 \le 1$. Apply (B.5) back to (B.3), we find that, with probability tending to one, $\Omega \le (2nd)\cdot\log(nd)\cdot p^{-1/2} \to 0$ by (A3). This proves the desired lemma conclusion and completes the proof.

Lemma L6. *Assume $d = d_0$, then $(np)^{-1}\|H(\widehat{\mathbb{Z}})\mathbb{X}\|^2 \ge (np)^{-1}\|H(\mathbb{Z})\mathbb{X}\|^2$.*

PROOF. Recall that $(\hat{\lambda}_j, \hat{V}_j)$ with $1 \le j \le n$ is the jth eigenvalue-eigenvector pair for the matrix $\mathbb{X}\mathbb{X}^\top/(np)$. Recall also $\widehat{\mathbb{Z}} = (\hat{V}_1, \cdots, \hat{V}_{d_0})$ and $H(\widehat{\mathbb{Z}}) = \sum_{j=1}^{d_0} \hat{V}_j \hat{V}_j^\top$. Then

$$(np)^{-1}\Big\|H(\mathbb{Z})\mathbb{X}\Big\|^2 = tr\bigg(H(\mathbb{Z}) \cdot \Big\{\mathbb{X}\mathbb{X}^\top/(np)\Big\}\bigg) = tr\big(\hat{H} \cdot \hat{\Lambda}\big), \tag{B.6}$$

where $\hat{\Lambda} = \mathrm{diag}(\hat{\lambda}_1, \cdots, \hat{\lambda}_n) \in \mathbb{R}^{n\times n}$, $\hat{H} = \widehat{\mathbb{V}}^\top H(\mathbb{Z})\widehat{\mathbb{V}}$, and $\widehat{\mathbb{V}} = (\hat{V}_1, \cdots, \hat{V}_n)$ is an orthonormal matrix. One can verify then $\hat{H} = (\hat{h}_{i_1 i_2})$ with $1 \le i_1, i_2 \le n$ is also a projection matrix. As a result, we should have $0 \le \hat{h}_{ii} \le 1$ for every $1 \le i \le n$. Then, the last term in (B.6) equals to

$$(np)^{-1}\Big\|H(\mathbb{Z})\mathbb{X}\Big\|^2 = \sum_{i=1}^{n} \hat{\lambda}_i \hat{h}_{ii}. \tag{B.7}$$

Note that the rank of $\hat{H}$ is d_0, which includes that $\sum_{i=1}^n \hat{h}_{ii} = d_0$. Recall that $0 \le \hat{h}_{ii} \le 1$. We know then the right hand side of (B.7) is no more than $\sum_{i=1}^{d_0} \hat{\lambda}_i = (np)^{-1}\|H(\widehat{\mathbb{Z}})\mathbb{X}\|^2$. This completes the proof.

Lemma L7. *Assume conditions (A1)–(A3), we have* $\max_{1\le j\le p} \|\widetilde{\mathbb{X}}_j\|^2/n = O_p(1)$.

PROOF. Define $\hat{\sigma}_j^2 = \|\widetilde{\mathbb{X}}_j\|^2/n$. Then, the conclusion follows if we can prove that $\max_{1\le j\le p} |\hat{\sigma}_j^2 - \tilde{\sigma}_j^2| = O_p(\sqrt{\log p/n}) = o_p(1)$, due to the fact $0 < \tilde{\sigma}^2_{\min} \le \tilde{\sigma}_j^2 \le 1$ and also the condition (A3); To this end, let $\kappa = (2/C_2)^{1/2}$. Then, by Bonferroni's inequality and (A4.c), we find that

$$P\Big(\max_{1\le j\le p} \big|\hat{\sigma}_j^2 - \tilde{\sigma}_j^2\big| > \kappa\big\{\log p/n\big\}^{1/2}\Big) \le \sum_{j=1}^{p} P\Big(\big|\hat{\sigma}_j^2 - \tilde{\sigma}_j^2\big| > \kappa\big\{\log p/n\big\}^{1/2}\Big)$$

$$\le pC_1 \exp\Big(-C_2\kappa^2 \log p\Big) = C_1 \exp\Big\{(1 - C_2\kappa^2)\log p\Big\} = C_1 \exp(-\log p) \to 0,$$

as $p \to \infty$. Consequently, we know that $\max_j |\hat{\sigma}_j^2 - \tilde{\sigma}_j^2| = O_p(\sqrt{\log p/n})$. This completes

the third step and finishes the entire proof.

Appendix C. Proof of Theorem 2

As one can see subsequently, the theoretical proof of Theorem 1 depends on the result of Theorem 2. In contrast, the proof of Theorem 2 is completely independent of Theorem 1. We thus present here the proof of Theorem 2 first, and delay that of Theorem 1 to the next Appendix D.

Recall that we have assumed that $d = d_0$. Then, the desired theorem conclusion can be proved in three steps. In the first step, we prove that $\mathcal{O}(\mathbb{Z})$, i.e., $\mathcal{O}(\cdot)$ evaluated under the true latent factor, should admit the following asymptotic expression, that is $\mathcal{O}(\mathbb{Z}) = \tilde{\sigma}_0^2 + O_p(n^{-1})$. See (2.6) for the definition of $\mathcal{O}(\cdot)$. In the second step, we show that $\mathcal{O}(\widehat{\mathbb{Z}}) \geq \tilde{\sigma}_0^2 + \lambda_{\min}(\Sigma_\beta) \cdot D_1(\mathbb{Z}, \widehat{\mathbb{Z}}) + O_p(n^{-1})$. By definition we should have $\mathcal{O}(\mathbb{Z}) \geq \mathcal{O}(\widehat{\mathbb{Z}})$, which immediately implies that $D_1(\mathbb{Z}, \widehat{\mathbb{Z}}) = O_p(n^{-1})$. In the last step, we prove $D_2(\mathbb{Z}, \widehat{\mathbb{Z}}) = O_p(n^{-1})$.

The 1st Step. By the definition of $\mathcal{O}(\mathbb{Z})$ and the assumed model (2.2), one can verify easily that $\mathcal{O}(\mathbb{Z}) = (np)^{-1} \sum_{j=1}^{p} \widetilde{\mathbb{X}}_j^\top Q(\mathbb{Z}) \widetilde{\mathbb{X}}_j$. By (A1) we know that $\widetilde{\mathbb{X}}_j$ follows a normal distribution. By definition, $Q(\mathbb{Z})$ is a projection matrix with rank $(n-d)$. We know then $\widetilde{\mathbb{X}}_j^\top Q(\mathbb{Z}) \widetilde{\mathbb{X}}_j / \tilde{\sigma}_j^2$ follows a chi-square distribution with $(n-d)$ degrees of freedom. Consequently, we have

$$E\big\{\mathcal{O}(\mathbb{Z})\big\} = (np)^{-1} \sum_{j=1}^{p} E\Big\{\widetilde{\mathbb{X}}_j^\top Q(\mathbb{Z}) \widetilde{\mathbb{X}}_j\Big\} = (np)^{-1} \sum_{j=1}^{p} \tilde{\sigma}_j^2 (n-d)$$

$$= \left(1 - \frac{d}{n}\right) \left(p^{-1} \sum_{j=1}^{p} \tilde{\sigma}_j^2\right) = \left(1 - \frac{d}{n}\right) \tilde{\sigma}_0^2 + O_p(p^{-1/2}) = \tilde{\sigma}_0^2 + O_p(n^{-1}), \qquad \text{(C.1)}$$

where the second equality in (C.1) is due to (A2) while the last one is due to (A3). Furthermore, recall $\widetilde{\mathbb{X}}_j^\top Q(\mathbb{Z}) \widetilde{\mathbb{X}}_j / \tilde{\sigma}_j^2$ follows a chi-square distribution and the degrees of

freedom is $(n-d)$. We can then verify that

$$\mathrm{var}\big\{\mathcal{O}(\mathbb{Z})\big\} = (np)^{-2}\sum_{j=1}^{p} 2\tilde{\sigma}_j^4(n-d)$$

$$= \left(\frac{2}{np}\right)\left(\frac{n-d}{n}\right)\left(p^{-1}\sum_{j=1}^{p}\tilde{\sigma}_j^4\right) \leq \left(\frac{2}{np}\right) = O\big\{(np)^{-1}\big\}, \tag{C.2}$$

where above inequality is due to the facts $(n-d)/n \leq 1$ and $\tilde{\sigma}_j^2 \leq \mathrm{var}(X_{ij}) = 1$. Then, both (C.1) and (C.2) proves that

$$\mathcal{O}(\mathbb{Z}) = \tilde{\sigma}_0^2 + O_p(n^{-1}). \tag{C.3}$$

The 2nd Step. We next evaluate the asymptotic behavior of $\mathcal{O}(\widehat{\mathbb{Z}})$. By the definition of $\mathcal{O}(\cdot)$ and note $\mathbb{X}_j = \mathbb{Z}\beta_j + \widetilde{\mathbb{X}}_j$, we have the following.

$$\mathcal{O}(\widehat{\mathbb{Z}}) = (np)^{-1}\sum_{j=1}^{p}\mathbb{X}_j^\top Q(\widehat{\mathbb{Z}})\mathbb{X}_j$$

$$= (np)^{-1}\sum_{j=1}^{p}\Big\{\beta_j^\top\mathbb{Z}^\top Q(\widehat{\mathbb{Z}})\mathbb{Z}\beta_j + \widetilde{\mathbb{X}}_j^\top Q(\widehat{\mathbb{Z}})\widetilde{\mathbb{X}}_j + 2\widetilde{\mathbb{X}}_j^\top Q(\widehat{\mathbb{Z}})\mathbb{Z}\beta_j\Big\}$$

$$= tr\left\{n^{-1}\Big(\mathbb{Z}^\top Q(\widehat{\mathbb{Z}})\mathbb{Z}\Big)\left(p^{-1}\sum_{j=1}^{p}\beta_j\beta_j^\top\right)\right\} + (np)^{-1}\sum_{j=1}^{p}\widetilde{\mathbb{X}}_j^\top\widetilde{\mathbb{X}}_j$$

$$-(np)^{-1}\sum_{j=1}^{p}\widetilde{\mathbb{X}}_j^\top H(\widehat{\mathbb{Z}})\widetilde{\mathbb{X}}_j + 2\sum_{k=1}^{d}\left((np)^{-1}\sum_{j=1}^{p}\widetilde{\mathbb{X}}_j^\top Q(\widehat{\mathbb{Z}})\mathbb{Z}_k\beta_{jk}\right),$$

where $\mathbb{Z}_k = (Z_{1k}, \cdots, Z_{nk})^\top \in \mathbb{R}^n$. Then, by Lemma L1, the right hand side of the above equality is no less than

$$\geq D_1(\mathbb{Z}, \widehat{\mathbb{Z}}) \cdot \lambda_{\min}\Big(p^{-1}B^\top B\Big) + (np)^{-1}\sum_{j=1}^{p}\widetilde{\mathbb{X}}_j^\top\widetilde{\mathbb{X}}_j$$

$$-(np)^{-1}tr\left(H(\widehat{\mathbb{Z}})\sum_{j=1}^{p}\widetilde{\mathbb{X}}_j\widetilde{\mathbb{X}}_j^\top\right)+2\sum_{k=1}^{d}\left((np)^{-1}\sum_{j=1}^{p}\widetilde{\mathbb{X}}_j^\top Q(\widehat{\mathbb{Z}})\mathbb{Z}_k\beta_{jk}\right).$$

By Lemma L2, the right hand side of the above inequality can be further bounded by

$$\geq D_1(\mathbb{Z},\widehat{\mathbb{Z}})\cdot\lambda_{\min}\Big(p^{-1}B^\top B\Big)+(np)^{-1}\sum_{j=1}^{p}\widetilde{\mathbb{X}}_j^\top\widetilde{\mathbb{X}}_j \tag{C.4}$$

$$-tr\Big\{n^{-1}H(\widehat{\mathbb{Z}})\Big\}\lambda_{\max}\Big(p^{-1}\widetilde{\mathbb{X}}\widetilde{\mathbb{X}}^\top\Big)+2\sum_{k=1}^{d}\left((np)^{-1}\sum_{j=1}^{p}\widetilde{\mathbb{X}}_j^\top Q(\widehat{\mathbb{Z}})\mathbb{Z}_k\beta_{jk}\right), \tag{C.5}$$

because $p^{-1}\widetilde{\mathbb{X}}\widetilde{\mathbb{X}}^\top=p^{-1}\sum_{j=1}^{p}\widetilde{\mathbb{X}}_j\widetilde{\mathbb{X}}_j^\top$. In what follows, those four terms involved in (C.4) and (C.5) will be evaluated separately. Firstly, by (A2), the first term in (C.4) can be expressed as

$$D_1(\mathbb{Z},\widehat{\mathbb{Z}})\lambda_{\min}\Big(p^{-1}B^\top B\Big)=D_1(\mathbb{Z},\widehat{\mathbb{Z}})\Big\{\lambda_{\min}(\Sigma_\beta)+O_p(p^{-1/2})\Big\}. \tag{C.6}$$

Next, by similar proofs as for (C.3) we know that the second term in (C.4) equals to $\tilde{\sigma}_0^2+O_p(n^{-1})$. Note that $H(\widehat{\mathbb{Z}})$ is a projection matrix, we thus have $tr\{n^{-1}H(\widehat{\mathbb{Z}})\}=d/n$. On the other hand, by Lemma L3, we know that $\lambda_{\max}(p^{-1}\widetilde{\mathbb{X}}\widetilde{\mathbb{X}}^\top)=O_p(1)$. Thus, the first term in (C.5) should be a $O_p(n^{-1})$. Lastly, we consider the second term in (C.5). By (A3), d is a fixed number. We thus can consider an arbitrary but fixed k as

$$\left|(np)^{-1}\sum_{j=1}^{p}\widetilde{\mathbb{X}}_j^\top Q(\widehat{\mathbb{Z}})\mathbb{Z}_k\beta_{jk}\right|=\left|n^{-1}\left(p^{-1}\sum_{j=1}^{p}\beta_{jk}\widetilde{\mathbb{X}}_j\right)^\top Q(\widehat{\mathbb{Z}})\mathbb{Z}_k\right|$$

$$\leq n^{-1}\left\|p^{-1}\sum_{j=1}^{p}\beta_{jk}\widetilde{\mathbb{X}}_j\right\|\cdot\big\|\mathbb{Z}_k\big\|. \tag{C.7}$$

Note $\widetilde{X}_{i_1j_1}$ and $\widetilde{X}_{i_2j_2}$ are mutually uncorrelated, as long as $i_1 \neq i_2$ or $j_1 \neq j_2$. Then,

$$E\left\|p^{-1}\sum_{j=1}^{p}\beta_{jk}\widetilde{\mathbb{X}}_j\right\|^2 = \left(\frac{n}{p}\right)\left(p^{-1}\sum_{j=1}^{p}\beta_{jk}^2\tilde{\sigma}_j^2\right)$$

$$\leq \left(\frac{n}{p}\right)\left(p^{-1}\sum_{j=1}^{p}\beta_{jk}^2\right) = O\big(n/p\big),$$

where the above inequality is due to the fact that $\tilde{\sigma}_j^2 \leq \mathrm{var}(X_{ij}) = 1$ and the equality is due to (A2). Supply this result back to (C.7) and also note that $\|\mathbb{Z}_k\|^2/n = O(1)$, we should know that

$$(np)^{-1}\sum_{j=1}^{p}\widetilde{\mathbb{X}}_j^\top Q(\widehat{\mathbb{Z}})\mathbb{Z}_k\beta_{jk} = O_p(p^{-1/2}). \tag{C.8}$$

Apply those results, e.g., (C.6) and (C.8), back to (C.5), we obtain that

$$\mathcal{O}(\widehat{\mathbb{Z}}) \geq D_1(\mathbb{Z},\widehat{\mathbb{Z}})\cdot\lambda_{\min}(\Sigma_\beta) + \tilde{\sigma}_0^2 + O_p(n^{-1}). \tag{C.9}$$

Combine together the results of (C.3) and (C.9), we prove the desired theorem conclusion, that is $D_1(\mathbb{Z},\widehat{\mathbb{Z}}) = O_p(n^{-1})$.

The 3rd Step. Define $\hat{\Sigma}_{\mathbb{Z}} = n^{-1}\mathbb{Z}^\top\mathbb{Z}$ and $\mathbb{Z}^* = \mathbb{Z}\hat{\Sigma}_{\mathbb{Z}}^{-1/2}$. We know immediately that $H(\mathbb{Z}) = n^{-1}\mathbb{Z}^*\mathbb{Z}^{*\top} \in \mathbb{R}^{p\times p}$ with rank d. We can then write

$$O_p(n^{-1}) = D_1(\mathbb{Z},\widehat{\mathbb{Z}}) = n^{-1}tr\Big(\mathbb{Z}^\top Q(\widehat{\mathbb{Z}})\mathbb{Z}\Big) = n^{-1}tr\Big(\hat{\Sigma}_{\mathbb{Z}}^{1/2}\mathbb{Z}^{*\top}Q(\widehat{\mathbb{Z}})\mathbb{Z}^*\hat{\Sigma}_{\mathbb{Z}}^{1/2}\Big).$$

Recall $Z_i \in \mathbb{R}^d$ $(1 \leq i \leq n)$ are independent random vectors with $E(Z_i) = 0$ and $\mathrm{cov}(Z_i) = I$. We thus have $\hat{\Sigma}_{\mathbb{Z}}^{1/2} = I + O_p(n^{-1/2})$. Then

$$O_p(n^{-1}) = n^{-1}tr\Big(\mathbb{Z}^{*\top}Q(\widehat{\mathbb{Z}})\mathbb{Z}^*\Big)\big\{1 + o_p(1)\big\}$$

$$= n^{-1}tr\Big(Q(\widehat{\mathbb{Z}})\mathbb{Z}^*\mathbb{Z}^{*\top}\Big)\{1+o_p(1)\} = \Big[d - tr\{H(\widehat{\mathbb{Z}})H(\mathbb{Z})\Big\}\Big]\{1+o_p(1)\}, \qquad \text{(C.10)}$$

where the last equality is due to the fact that $H(\mathbb{Z}) = n^{-1}\mathbb{Z}^*\mathbb{Z}^{*\top}$ is a projection matrix with rank d. One can check further that

$$d - tr\Big\{H(\widehat{\mathbb{Z}})H(\mathbb{Z})\Big\} = \frac{1}{2}tr\Big\{H(\mathbb{Z}) - H(\widehat{\mathbb{Z}})\Big\}^2 = \frac{1}{2}D_2(\mathbb{Z}, \widehat{\mathbb{Z}}).$$

Applying this result back to (C.10), we proved that $D_2(\mathbb{Z}, \widehat{\mathbb{Z}}) = O_p(n^{-1})$. This completes the entire theorem proof.

Appendix D. Proof of Theorem 1

We should prove the theorem conclusion in a total of four steps. Specifically, in the first step, we show that, with probability tending to one, $\hat{\lambda}_1 \leq \kappa_1$ for some positive constant $0 < \kappa_1 < \infty$. We next show that, with probability tending to one, $\hat{\lambda}_{d_0} \geq \kappa_2$ for some $0 < \kappa_2 < \infty$. Consequently, as long as n is sufficiently large, we should have

$$\max_{j<d_0}(\hat{\lambda}_j/\hat{\lambda}_{j+1}) \leq \hat{\lambda}_1/\hat{\lambda}_{d_0} \leq \kappa_1/\kappa_2 = O_p(1). \qquad \text{(D.1)}$$

In the third step, we prove that $n\hat{\lambda}_{d_0+1} = O_p(1)$. In the last step we show that, with probability tending to one, $n\hat{\lambda}_n \geq \kappa_3$ for some $0 < \kappa_3 < \infty$. Consequently, as long as the sample size is sufficiently large, we should have

$$\max_{j>d_0}(\hat{\lambda}_j/\hat{\lambda}_{j+1}) \leq \hat{\lambda}_{d_0+1}/\hat{\lambda}_n = (n\hat{\lambda}_{d_0+1})/(n\hat{\lambda}_n) \leq (n\hat{\lambda}_{d_0+1})/\kappa_3 = O_p(1). \qquad \text{(D.2)}$$

Lastly, by the results from the second and third step, we know that

$$\hat{\lambda}_{d_0}/\hat{\lambda}_{d_0+1} = n \cdot \hat{\lambda}_{d_0}/(n\hat{\lambda}_{d_0+1}) \geq n\kappa_2/O_p(1) \to_p \infty. \qquad \text{(D.3)}$$

Combing the results of (D.1), (D.2), and (D.3) together, we obtain $P(\hat{d} = d_0) \to 1$. Thereafter, we should present detailed proofs accordingly.

The 1st Step. Recall $\text{var}(X_{ij}) = 1$ and $\lambda_{\max}(A)$ is the maximum eigenvalue of an arbitrary semi positive definite matrix A. We then have

$$\hat{\lambda}_1 = \lambda_{\max}\Big\{(np)^{-1}\mathbb{X}\mathbb{X}^\top\Big\} \leq tr\Big\{(np)^{-1}\mathbb{X}\mathbb{X}^\top\Big\} = (np)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{p} X_{ij}^2.$$

Recall the fact that $X_{ij} = Z_i^\top \beta_j + \widetilde{X}_{ij}$. Then, with probability tending to one, the right hand side of the above inequality can be bounded by

$$\leq \left(\frac{2}{np}\right)\sum_{i=1}^{n}\sum_{j=1}^{p}\Big\{\big(Z_i^\top\beta_j\big)^2 + \widetilde{X}_{ij}^2\Big\} \tag{D.4}$$

$$= 2\cdot tr\left\{\Big(n^{-1}\sum_{i=1}^{n} Z_i Z_i^\top\Big)\Big(p^{-1}\sum_{j=1}^{p}\beta_j\beta_j^\top\Big)\right\} + \frac{2}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}\widetilde{X}_{ij}^2$$

$$\leq 2d\lambda_{\max}(\Sigma_\beta) + 3, \tag{D.5}$$

where (D.4) is due to Cauchy's inequality and (D.5) is due to the facts $n^{-1}\sum Z_i Z_i^\top \to_p I$, $p^{-1}\sum \beta_j\beta_j^\top \to_p \Sigma_\beta$, and $(np)^{-1}\sum \widetilde{X}_{ij}^2 \to_p \tilde{\sigma}_0^2 \leq \text{var}(X_{ij}) = 1$. Define $\kappa_1 = 2d\lambda_{\max}(\Sigma_\beta) + 3$. This completes the first step proof.

The 2nd Step. Recall $(\hat{\lambda}_j, \hat{V}_j)$ is the jth $(1 \leq j \leq n)$ eigenvalue-eigenvector pair of the matrix $\mathbb{X}\mathbb{X}^\top/(np)$. Consequently, we have

$$\left(\sum_{j=1}^{d_0}\hat{\lambda}_j\right)^{1/2} = tr^{1/2}\Big\{H(\widehat{\mathbb{Z}})\cdot(np)^{-1}\mathbb{X}\mathbb{X}^\top\Big\} = (np)^{-1/2}tr^{1/2}\Big\{\mathbb{X}^\top H(\widehat{\mathbb{Z}})\mathbb{X}\Big\}$$

$$= (np)^{-1/2}\big\|H(\widehat{\mathbb{Z}})\mathbb{X}\big\| \geq (np)^{-1/2}\big\|H(\widehat{\mathbb{Z}})\mathbb{Z}B^\top\big\| - (np)^{-1/2}\big\|H(\widehat{\mathbb{Z}})\widetilde{\mathbb{X}}\big\|$$

$$\geq (np)^{-1/2}\big\|\mathbb{Z}B^\top\big\| - (np)^{-1/2}\big\|Q(\widehat{\mathbb{Z}})\mathbb{Z}B^\top\big\| \tag{D.6}$$

$$-(np)^{-1/2}\Big\|\Big\{H(\widehat{\mathbb{Z}})-H(\mathbb{Z})\Big\}\widetilde{\mathbb{X}}\Big\|-(np)^{-1/2}\Big\|H(\mathbb{Z})\widetilde{\mathbb{X}}\Big\|. \tag{D.7}$$

The four terms involved in the (D.6) and (D.7) are then evaluated as the follows.

$$(np)^{-1}\big\|\mathbb{Z}B^\top\big\|^2=tr\Big\{\big(n^{-1}\mathbb{Z}^\top\mathbb{Z}\big)\big(p^{-1}B^\top B\big)\Big\}\to_p tr(\Sigma_\beta), \tag{D.8}$$

according to the condition (A2) and the fact $\mathrm{cov}(Z_i)=I$. Next, note that

$$(np)^{-1}\big\|Q(\widehat{\mathbb{Z}})\mathbb{Z}B^\top\big\|^2=tr\left[n^{-1}\Big\{\mathbb{Z}^\top Q(\widehat{\mathbb{Z}})\mathbb{Z}\Big\}\Big(p^{-1}B^\top B\Big)\right]$$

$$\le n^{-1}tr\Big\{\mathbb{Z}^\top Q(\widehat{\mathbb{Z}})\mathbb{Z}\Big\}\cdot tr(\Sigma_\beta)\{1+o_p(1)\} \tag{D.9}$$

$$=D_1(\mathbb{Z},\widehat{\mathbb{Z}})\cdot tr(\Sigma_\beta)\{1+o_p(1)\}=O_p(n^{-1}), \tag{D.10}$$

where the inequality in (D.9) is due to Lemma L2 and the last equality in (D.10) is due to the first result of Theorem 2. Similarly, we have

$$(np)^{-1}\Big\|\Big\{H(\widehat{\mathbb{Z}})-H(\mathbb{Z})\Big\}\widetilde{\mathbb{X}}\Big\|^2=tr\left[\Big\{H(\widehat{\mathbb{Z}})-H(\mathbb{Z})\Big\}^2\Big\{\widetilde{\mathbb{X}}\widetilde{\mathbb{X}}^\top/(np)\Big\}\right]$$

$$\le tr\Big\{H(\widehat{\mathbb{Z}})-H(\mathbb{Z})\Big\}^2\cdot\lambda_{\max}\Big\{\widetilde{\mathbb{X}}\widetilde{\mathbb{X}}^\top/(np)\Big\}=O_p(n^{-1}), \tag{D.11}$$

where the last equality is due to the second result of Theorem 2, Lemma L2, and also Lemma L3. Lastly, note that

$$(np)^{-1}\big\|H(\mathbb{Z})\widetilde{\mathbb{X}}\big\|^2=(np)^{-1}\sum_{j=1}^{p}\widetilde{\mathbb{X}}_j^\top H(\mathbb{Z})\widetilde{\mathbb{X}}_j\le(np)^{-1}\sum_{j=1}^{p}\tilde{\sigma}_j^{-2}\widetilde{\mathbb{X}}_j^\top H(\mathbb{Z})\widetilde{\mathbb{X}}_j,$$

where once again the last inequality is due to the fact that $\tilde{\sigma}_j^2\le\mathrm{var}(X_{ij})=1$. Note

that $\text{cov}(\widetilde{\mathbb{X}}_j) = \tilde{\sigma}_j^2 I$, we know then $\sigma_j^{-2}\widetilde{\mathbb{X}}_j^\top H(\mathbb{Z})\widetilde{\mathbb{X}}_j$ follows a $\chi^2(d)$ distribution. Then,

$$E\left\{(np)^{-1}\left\|H(\mathbb{Z})\widetilde{\mathbb{X}}\right\|^2\right\} \leq \frac{d}{n}, \tag{D.12}$$

which implies that that last term in (D.7) is a $O_p(n^{-1/2})$. Applying this result together with (D.8), (D.10), and (D.11) back to (D.7), we know that

$$\sum_{j=1}^{d_0} \hat{\lambda}_j \geq tr(\Sigma_\beta) + o_p(1). \tag{D.13}$$

We next consider what happens to $\sum_{j=1}^{d_0-1}\hat{\lambda}_j$. Define $\hat{\mathbb{Z}}^- = (\hat{V}_1, \cdots, \hat{V}_{d_0-1}) \in \mathbb{R}^{n\times(d_0-1)}$. We then have the following relationship

$$\left(\sum_{j=1}^{d_0-1}\hat{\lambda}_j\right)^{1/2} = (np)^{-1/2}\left\|H(\widehat{\mathbb{Z}}^-)\mathbb{X}\right\| \leq (np)^{-1/2}\left\|H(\widehat{\mathbb{Z}}^-)\mathbb{Z}B^\top\right\| + (np)^{-1/2}\left\|H(\widehat{\mathbb{Z}}^-)\widetilde{\mathbb{X}}\right\|$$

$$= (np)^{-1/2}\left\|H(\widehat{\mathbb{Z}}^-)\mathbb{Z}B^\top\right\| + o_p(1), \tag{D.14}$$

where the above equality is due to the fact $(np)^{-1}\|H(\widehat{\mathbb{Z}}^-)\widetilde{\mathbb{X}}\|^2 \leq (np)^{-1}\|H(\widehat{\mathbb{Z}})\widetilde{\mathbb{X}}\|^2 = o_p(1)$, which has been proved in the first step; see (D.12). We use $\text{rank}(A)$ denote the rank of an arbitrary matrix A, while $\dim(\mathcal{S})$ to represent the dimension of an linear subspace $\mathcal{S}$. We know immediately that $\dim\{\mathcal{S}(A)\} = \text{rank}(A)$. Note that $\text{rank}(\mathbb{Z}) = d_0$ while $\text{rank}\{Q(\widehat{\mathbb{Z}}^-)\} = n - d_0 + 1$. We know immediately that $\mathcal{S}^* = \mathcal{S}(\mathbb{Z})\bigcap\mathcal{S}\{Q(\widehat{\mathbb{Z}}^-)\} \neq \emptyset$, because $\text{rank}(\mathbb{Z}) + \text{rank}\{Q(\widehat{\mathbb{Z}}^-)\} > n$. We can then find a set of orthonormal basis $\{U_j : 1 \leq j \leq n - d_0 + 1\}$ such that $Q(\widehat{\mathbb{Z}}^-) = \sum_{j=1}^{n-d_0+1} U_j U_j^\top$ and $U_1 \in \mathcal{S}^* \subset \mathcal{S}(\mathbb{Z})$. Then

$$(np)^{-1}\left\|Q(\widehat{\mathbb{Z}}^-)\mathbb{Z}B^\top\right\|^2 = (np)^{-1}tr\Big\{Q(\widehat{\mathbb{Z}}^-)\big(\mathbb{Z}B^\top B\mathbb{Z}^\top\big)\Big\}$$

$$\geq (np)^{-1}tr\Big\{\big(U_1U_1^\top\big)\big(\mathbb{Z}B^\top B\mathbb{Z}^\top\big)\Big\} = n^{-1}U_1^\top\mathbb{Z}\Big(p^{-1}B^\top B\Big)\mathbb{Z}^\top U_1$$

$$\geq \lambda_{\min}(\Sigma_\beta)\cdot U_1^\top\Big(\mathbb{Z}\mathbb{Z}^\top/n\Big)U_1\cdot\{1+O_p(n^{-1})\}. \tag{D.15}$$

Because $U_1 \in \mathcal{S}(\mathbb{Z})$ and $\|U_1\| = 1$, we can then write $U_1 = \mathbb{Z}\omega_1$ with $1 = \|U_1\|^2 = \omega_1^\top(\mathbb{Z}^\top\mathbb{Z})\omega_1$. We know then

$$U_1^\top\{\mathbb{Z}\mathbb{Z}^\top/n\}U_1 = \omega_1^\top\big(\mathbb{Z}^\top\mathbb{Z}\big)\big(n^{-1}\mathbb{Z}^\top\mathbb{Z}\big)\omega_1 = \omega_1^\top\big(\mathbb{Z}^\top\mathbb{Z}\big)\omega_1\{1+o_p(1)\} = 1+o_p(1),$$

where the second equality is due to the fact that $n^{-1}\mathbb{Z}^\top\mathbb{Z} = I + o_p(1)$ and the last one is because $\omega_1(\mathbb{Z}^\top\mathbb{Z})\omega_1 = \|U_1\|^2 = 1$. Applying this result back to the right hand side of (D.15), we know then $(np)^{-1}\|Q(\widehat{\mathbb{Z}}^-)\mathbb{Z}B^\top\|^2 \geq \lambda_{\min}(\Sigma_\beta) + o_p(1)$. Note that $tr(\Sigma_\beta) + o_p(1) = (np)^{-1}\|\mathbb{Z}B^\top\|^2 = (np)^{-1}\|Q(\widehat{\mathbb{Z}}^-)\mathbb{Z}B^\top\|^2 + (np)^{-1}\|H(\widehat{\mathbb{Z}}^-)\mathbb{Z}B^\top\|^2$. We know then $(np)^{-1}\|H(\widehat{\mathbb{Z}}^-)\mathbb{Z}B^\top\|^2 \leq tr(\Sigma_\beta) - \lambda_{\min}(\Sigma_\beta) + o_p(1)$. Applying this result back to (D.14), we find that $\sum_{j=1}^{d_0-1}\hat{\lambda}_j \leq tr(\Sigma_\beta) - \lambda_{\min}(\Sigma_\beta) + o_p(1)$. Such a result together with (D.13) implies that, with probability tending to one, $\hat{\lambda}_{d_0} \geq \lambda_{\min}(\Sigma_\beta) + o_p(1) \geq \kappa_2 = 2^{-1}\lambda_{\min}(\Sigma_\beta)$. This completes the second step proof.

The 3rd Step. In this step, we consider $\sum_{j=1}^{d_0+1}\hat{\lambda}_j$. Define $\hat{\mathbb{Z}}^+ = (\hat{V}_1, \cdots, \hat{V}_{d_0+1}) \in \mathbb{R}^{n\times(d_0+1)}$. We then have the following relationship

$$\sum_{j=1}^{d_0+1}\hat{\lambda}_j = (np)^{-1}\Big\|H(\widehat{\mathbb{Z}}^+)\mathbb{X}\Big\|^2$$

$$\leq (np)^{-1}\Big\|H(\widehat{\mathbb{Z}}^+)\mathbb{Z}B^\top\Big\|^2 + (np)^{-1}\Big\|H(\widehat{\mathbb{Z}}^+)\widetilde{\mathbb{X}}\Big\|^2 + 2(np)^{-1}\Big|tr\Big\{H(\widehat{\mathbb{Z}}^+)\widetilde{\mathbb{X}}B\mathbb{Z}^\top\Big\}\Big|$$

$$\leq (np)^{-1}\big\|\mathbb{Z}B^\top\big\|^2 + (np)^{-1}\big\|H(\widehat{\mathbb{Z}}^+)\widetilde{\mathbb{X}}\big\|^2 + 2n^{-1}(d_0+1)\lambda_{\max}\Big\{p^{-1}\widetilde{\mathbb{X}}B\mathbb{Z}^\top\Big\} \tag{D.16}$$

$$= (np)^{-1}\big\|\mathbb{Z}B^\top\big\|^2 + (np)^{-1}tr\Big\{H(\widehat{\mathbb{Z}}^+)\widetilde{\mathbb{X}}\widetilde{\mathbb{X}}^\top\Big\} + o_p(n^{-1}) \tag{D.17}$$

$$\leq (np)^{-1}\left\|\mathbb{Z}B^\top\right\|^2 + (np)^{-1}tr\Big\{H(\widehat{\mathbb{Z}}^+)\Big\}\lambda_{\max}\Big\{\widetilde{\mathbb{X}}\widetilde{\mathbb{X}}^\top\Big\} + o_p(n^{-1}), \tag{D.18}$$

where (D.16) is due to Lemma L4, (D.17) is due to Lemma L5, (D.18) is due to Lemma L2. Lastly, note the fact that $tr\{H(\widehat{\mathbb{Z}}^+)\} = d_0 + 1$. We can then write the right hand side of (D.18) as

$$= (np)^{-1}\left\|\mathbb{Z}B^\top\right\|^2 + \left(\frac{d_0+1}{n}\right)\lambda_{\max}\Big\{p^{-1}\widetilde{\mathbb{X}}\widetilde{\mathbb{X}}^\top\Big\} + o_p(1)$$

$$= (np)^{-1}\left\|\mathbb{Z}B^\top\right\|^2 + O_p(n^{-1}), \tag{D.19}$$

where the last equality is due to the fact that $\lambda_{\max}(p^{-1}\widetilde{\mathbb{X}}\widetilde{\mathbb{X}}^\top) = O_p(1)$; see Lemma L3. Similarly, by Lemma L6, we can prove that

$$\sum_{j=1}^{d_0}\hat{\lambda}_j = (np)^{-1}\Big\|H(\widehat{\mathbb{Z}})\mathbb{X}\Big\|^2 \geq (np)^{-1}\Big\|H(\mathbb{Z})\mathbb{X}\Big\|^2$$

$$\geq (np)^{-1}\Big\|H(\mathbb{Z})\mathbb{Z}B^\top\Big\|^2 - (np)^{-1}\Big\|H(\mathbb{Z})\widetilde{\mathbb{X}}\Big\|^2 - 2(np)^{-1}\Big|tr\Big\{H(\mathbb{Z})\widetilde{\mathbb{X}}B\mathbb{Z}^\top\Big\}\Big|$$

$$= (np)^{-1}\|\mathbb{Z}B^\top\|^2 + O_p(n^{-1}).$$

This together with (D.19) implies $n\hat{\lambda}_{d_0+1} = O_p(1)$.

The 4th Step. In this step, we would like to evaluate $\hat{\lambda}_n$, which by definition is the smallest eigenvalue of $\mathbb{X}\mathbb{X}^\top/(np)$. Moreover, note that $\mathbb{X} = \mathbb{Z}B^\top + \widetilde{\mathbb{X}}$. Then,

$$n\hat{\lambda}_n = \lambda_{\min}\big(p^{-1}\mathbb{X}\mathbb{X}^\top\big) \geq \lambda_{\min}\big(p^{-1}\widetilde{\mathbb{X}}\widetilde{\mathbb{X}}^\top\big) - 2\lambda_{\max}\big(p^{-1}\widetilde{\mathbb{X}}B\mathbb{Z}^\top\big). \tag{D.20}$$

By Lemma L3, we know that, with probability tending to one, $\lambda_{\min}(p^{-1}\widetilde{\mathbb{X}}\widetilde{\mathbb{X}}^\top) > 2^{-1}\tilde{\sigma}^2_{\min} > 0$. In the meanwhile, by Lemma L5, we know that $\lambda_{\max}(p^{-1}\widetilde{\mathbb{X}}B\mathbb{Z}^\top) \to_p 0$. Those results together with (D.20) implies that there exists a positive constant $\kappa_3 > 0$

such that $P(n\hat{\lambda}_n > \kappa_3) \to 1$. This completes the last step proof. The proof of the entire theorem is also accomplished.

Appendix E. Proof of Theorem 3

Define $\tilde{\rho}_j = E(\widetilde{X}_{ij}\widetilde{Y}_i) = \tilde{\sigma}_j^2\theta_{0j}$. By (A1), we know that $\min \tilde{\sigma}_j^2 \geq \tilde{\sigma}_{\min}^2 > 0$. Then, theorem conclusion follows as long as we can prove that: (1) $\max_{1\leq j\leq p} |\hat{\rho}_j - \tilde{\rho}_j| = O_p(\sqrt{\log p/n})$ and (2) $\max_{1\leq j\leq p} |\hat{\sigma}_j^2 - \tilde{\sigma}_j^2| = O_p(\sqrt{\log p/n})$, where $\hat{\rho}_j = \widehat{\mathbb{Y}}^\top\widehat{\mathbb{X}}_j/n$ and $\hat{\sigma}_j^2 = \widehat{\mathbb{X}}_j^\top\widehat{\mathbb{X}}_j/n$. Because the proofs for both statements are very similar, we will supply the details for the first statement only. More specifically, this can be done in three steps. In the first step, we prove that the difference between $(\widetilde{\mathbb{Y}}, \widetilde{\mathbb{X}})$ and $(\widehat{\mathbb{Y}}, \widehat{\mathbb{X}})$ is uniformly small. Subsequently, the same is done for $\hat{\rho}_j$ and $\hat{\rho}_j^* = \widetilde{\mathbb{Y}}^\top\widetilde{\mathbb{X}}_j/n$ in the second step while $\hat{\rho}_j^*$ and $\tilde{\rho}_j$ in the last step.

The 1st Step. Recall $\widetilde{\mathbb{Y}} = \widetilde{\mathbb{X}}\theta_0 + \widetilde{\mathcal{E}}$ and $\widehat{\mathbb{Y}} = Q(\widehat{\mathbb{Z}})\mathbb{Z}\gamma_0 + Q(\widehat{\mathbb{Z}})(\widetilde{\mathbb{X}}\theta_0 + \widetilde{\mathcal{E}})$. Then, the difference between $\widetilde{\mathbb{Y}}$ and $\widehat{\mathbb{Y}}$ can be decomposed as the follows

$$\widehat{\mathbb{Y}} - \widetilde{\mathbb{Y}} = Q(\widehat{\mathbb{Z}})\mathbb{Z}\gamma_0 + \Big\{H(\mathbb{Z}) - H(\widehat{\mathbb{Z}})\Big\}\Big(\widetilde{\mathbb{X}}\theta_0 + \widetilde{\mathcal{E}}\Big) - H(\mathbb{Z})\Big(\widetilde{\mathbb{X}}\theta_0 + \widetilde{\mathcal{E}}\Big).$$

Then, by Cauchy's inequality, we find that

$$n^{-1}\big\|\widehat{\mathbb{Y}} - \widetilde{\mathbb{Y}}\big\|^2/3 \leq \gamma_0^\top\Big(n^{-1}\mathbb{Z}^\top Q(\widehat{\mathbb{Z}})\mathbb{Z}\Big)\gamma_0$$

$$+n^{-1}\Big(\widetilde{\mathbb{X}}\theta_0 + \widetilde{\mathcal{E}}\Big)^\top\Big\{H(\widehat{\mathbb{Z}}) - H(\mathbb{Z})\Big\}^2\Big(\widetilde{\mathbb{X}}\theta_0 + \widetilde{\mathcal{E}}\Big) + n^{-1}\Big(\widetilde{\mathbb{X}}\theta_0 + \widetilde{\mathcal{E}}\Big)^\top H(\mathbb{Z})\Big(\widetilde{\mathbb{X}}\theta_0 + \widetilde{\mathcal{E}}\Big)$$

$$\leq \lambda_{\max}\Big(n^{-1}\mathbb{Z}^\top Q(\widehat{\mathbb{Z}})\mathbb{Z}\Big)\|\gamma_0\|^2 + n^{-1}\lambda_{\max}\Big\{H(\widehat{\mathbb{Z}}) - H(\mathbb{Z})\Big\}^2\big\|\widetilde{\mathbb{X}}\theta_0 + \widetilde{\mathcal{E}}\big\|^2$$

$$+n^{-1}\Big(\widetilde{\mathbb{X}}\theta_0 + \widetilde{\mathcal{E}}\Big)^\top H(\mathbb{Z})\Big(\widetilde{\mathbb{X}}\theta_0 + \widetilde{\mathcal{E}}\Big).$$

Define $\tilde{\sigma}_\theta^2 = \mathrm{var}(\widetilde{X}_i^\top\theta_0 + \tilde{\varepsilon}_i) \leq \mathrm{var}(Y_i) = 1$. Then, by the definition of the discrepancy

measures $D_1(\cdot,\cdot)$ and $D_2(\cdot,\cdot)$, we find that right hand side of the above inequality can be further bounded by

$$\leq D_1(\mathbb{Z},\widehat{\mathbb{Z}})\|\gamma_0\|^2 + D_2(\mathbb{Z},\widehat{\mathbb{Z}})\cdot n^{-1}\Big\|\widetilde{\mathbb{X}}\theta_0+\widetilde{\mathcal{E}}\Big\|^2$$

$$+n^{-1}\tilde{\sigma}_\theta^{-2}\Big(\widetilde{\mathbb{X}}\theta_0+\widetilde{\mathcal{E}}\Big)^\top H(\mathbb{Z})\Big(\widetilde{\mathbb{X}}\theta_0+\widetilde{\mathcal{E}}\Big).$$

By Law of Large Number we know that $n^{-1}\|\widetilde{\mathbb{X}}\theta_0+\widetilde{\mathcal{E}}\|^2 = O_p(1)$. By Theorem 2 we know that both $D_1(\mathbb{Z},\widehat{\mathbb{Z}}) = O_p(n^{-1})$ and $D_2(\mathbb{Z},\widehat{\mathbb{Z}}) = O_p(n^{-1})$. Furthermore, note that $\tilde{\sigma}_\theta^{-2}(\widetilde{\mathbb{X}}\theta_0+\widetilde{\mathcal{E}})^\top H(\mathbb{Z})(\widetilde{\mathbb{X}}\theta_0+\widetilde{\mathcal{E}})$ follows a chi-square distribution with d degrees of freedom. Applying those result back to the above inequality, we find that $n^{-1}\|\widehat{\mathbb{Y}}-\widetilde{\mathbb{Y}}\|^2 = O_p(n^{-1})$. Similarly, because $\mathbb{X}_j = \mathbb{Z}\beta_j + \widetilde{\mathbb{X}}_j$, we can then prove that

$$n^{-1}\big\|\widehat{\mathbb{X}}_j-\widetilde{\mathbb{X}}_j\big\|^2/3 \leq D_1(\mathbb{Z},\widehat{\mathbb{Z}})\cdot\|\beta_j\|^2 + D_2(\mathbb{Z},\widehat{\mathbb{Z}})\Big(n^{-1}\|\widetilde{\mathbb{X}}_j\|^2\Big) + n^{-1}\cdot\chi^2(d).$$

Taking maximum over j at the both sides of the above inequality, we find that

$$\max_{1\leq j\leq p}\Big(n^{-1}\big\|\widehat{\mathbb{X}}_j-\widetilde{\mathbb{X}}_j\big\|^2\Big)/3 \leq D_1(\mathbb{Z},\widehat{\mathbb{Z}})$$

$$+D_2(\mathbb{Z},\widehat{\mathbb{Z}})\cdot\max_j\Big(n^{-1}\|\widetilde{\mathbb{X}}_j\|^2\Big) + n^{-1}\cdot\max_j\chi^2(d),$$

where the above inequality is partially due to the fact that $\|\beta_j\|^2 = \mathrm{var}(\beta_j^\top Z_i) \leq \mathrm{var}(X_{ij}) = 1$. By Lemma L7, we know that $\max_j\|\widetilde{\mathbb{X}}_j\|^2/n = O_p(1)$. By Wang et al. (2009) we know that, with probability tending to one, $\max_{1\leq j\leq p}\chi^2(d) \leq 2\log(pd)$. Thus, the right hand side of the above inequality is a $O_p(\log p/n)$. Thus,

$$\max_{1\leq j\leq p}\Big(n^{-1}\big\|\widehat{\mathbb{X}}_j-\widetilde{\mathbb{X}}_j\big\|^2\Big) = O_p\big(n^{-1}\log p\big) \text{ and } n^{-1}\big\|\widehat{\mathbb{Y}}-\widetilde{\mathbb{Y}}\big\|^2 = O_p(n^{-1}). \tag{E.1}$$

The 2nd Step. We next consider the maximum difference between $\hat{\rho}_j$ and $\hat{\rho}_j^*$. More specifically, it can be bounded as

$$\max_{1\le j\le p}\left|\hat{\rho}_j-\hat{\rho}_j^*\right| \le n^{-1}\max_{1\le j\le p}\left|\widehat{\mathbb{Y}}^\top(\widehat{\mathbb{X}}_j-\widetilde{\mathbb{X}}_j)\right| + n^{-1}\max_{1\le j\le p}\left|\left(\widehat{\mathbb{Y}}-\widetilde{\mathbb{Y}}\right)^\top\widetilde{\mathbb{X}}_j\right|$$

$$\le \left(n^{-1}\left\|\widehat{\mathbb{Y}}\right\|^2\right)^{1/2}\max_j\left(n^{-1}\left\|\widehat{\mathbb{X}}_j-\widetilde{\mathbb{X}}_j\right\|^2\right)^{1/2} + \left(n^{-1}\left\|\widehat{\mathbb{Y}}-\widetilde{\mathbb{Y}}\right\|^2\right)^{1/2}\max_j\left(n^{-1}\left\|\widetilde{\mathbb{X}}_j\right\|^2\right)^{1/2}$$

$$= \left(n^{-1}\left\|\widehat{\mathbb{Y}}\right\|^2\right)^{1/2}O_p\Big(\sqrt{\log p/n}\Big) + O_p(n^{-1/2})\max_j\left(n^{-1}\left\|\widetilde{\mathbb{X}}_j\right\|^2\right)^{1/2}, \qquad \text{(E.2)}$$

due to (E.1). Similarly, we have $n^{-1}\|\widehat{\mathbb{Y}}\|^2 \le 2n^{-1}\|\widetilde{\mathbb{Y}}\|^2 + 2n^{-1}\|\widetilde{\mathbb{Y}}-\widehat{\mathbb{Y}}\|^2 = O_p(1)$. Next note that $n^{-1}\|\widetilde{\mathbb{Y}}\|^2 \to_p \mathrm{var}(\widetilde{Y}_i) \le \mathrm{var}(Y_i) = 1$. Furthermore, note that $n^{-1}\max_j\|\widetilde{\mathbb{X}}_j\|^2 = O_p(1)$; see Lemma L7. Apply those results back to (E.2), we find that $\max_{1\le j\le p}|\hat{\rho}_j - \hat{\rho}_j^*| = O_p(\sqrt{\log p/n})$. Then, theorem conclusion follows if we can further prove that $\max_{1\le j\le p}|\hat{\rho}_j^* - \tilde{\rho}_j| = O_p(\sqrt{\log p/n})$.

The 3rd Step. By the exponential inequality (A4.b) as implied by (A1), we know that there exists two positive constants C_1 and C_2, such that $P(|\hat{\rho}_j^* - \tilde{\rho}_j| > \nu) \le C_1\exp(-C_2 n\nu^2)$, where ν is an arbitrary positive number. Let $\kappa = (2/C_2)^{1/2}$. Then, by Bonferroni's inequality, we have

$$P\Big(\max_{1\le j\le p}\left|\hat{\rho}_j^*-\tilde{\rho}_j\right| > \kappa\big\{\log p/n\big\}^{1/2}\Big) \le \sum_{j=1}^{p}P\Big(\left|\hat{\rho}_j^*-\tilde{\rho}_j\right| > \kappa\big\{\log p/n\big\}^{1/2}\Big)$$

$$\le pC_1\exp\Big(-C_2\kappa^2\log p\Big) = C_1\exp\Big\{(1-C_2\kappa^2)\log p\Big\} = C_1\exp(-\log p)\to 0,$$

as $p\to\infty$. Consequently, we know that $\max_j|\hat{\rho}_j^*-\tilde{\rho}_j| = O_p(\sqrt{\log p/n})$. This completes the third step and finishes the entire proof.