Biometrika (2012), **99**, 1, *pp.* 15–28
© 2012 Biometrika Trust
Printed in Great Britain

doi: 10.1093/biomet/asr074
Advance Access publication 3 February 2012

Factor profiled sure independence screening

By H. WANG

Guanghua School of Management, Peking University, Beijing 100871, China
hansheng@gsm.pku.edu.cn

Summary

We propose a method of factor profiled sure independence screening for ultrahigh-dimensional variable selection. The objective of this method is to identify nonzero components consistently from a sparse coefficient vector. 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, which can be estimated consistently by eigenvalue-eigenvector decomposition. The estimated latent factors should then be profiled out from both the response and the predictors. Such an operation, referred to as factor profiling, produces uncorrelated predictors. Therefore, sure independence screening can be applied subsequently and the resulting screening result is consistent for model selection, a major advantage that standard sure independence screening does not share. We refer to the new method as factor profiled sure independence screening. Numerical studies confirm its outstanding performance.

Some key words: Factor profiled sure independence screening; Factor profiling; Maximum eigenvalue ratio criterion; Screening consistency; Sure independence screening.

1. Introduction

For many modern datasets, the predictor dimension is substantially larger than the sample size, so classical methods such as ordinary least squares are inapplicable. As a result, dimension reduction is the central theme of high-dimensional data analysis, for which both the ideas of factor modelling (Johnson & Wichern, 2003; Fan et al., 2008) and variable selection are very useful.

Under a fixed dimension set-up, best subset selection in conjunction with the Akaike information criterion (Akaike, 1973) and the Bayesian information criterion (Schwarz, 1978) has been widely used in practice. Despite the usefulness of these methods, best subset selection suffers from high computational cost (Tibshirani, 1996), estimation instability (Breiman, 1996) and complicated stochastic properties (Fan & Li, 2001; Hjort & Claeskens, 2003). Various shrinkage methods, developed as computationally efficient alternatives, have gained popularity in the past decade. These include the nonnegative garrotte (Breiman, 1995; Yuan & Lin, 2007), the least absolute shrinkage and selection operator (Tibshirani, 1996; Knight & Fu, 2000; Zhao & Yu, 2006), bridge regression (Fu, 1998; Huang et al., 2007), smoothly clipped absolute deviation (Fan & Li, 2001; Fan & Peng, 2004; Wang et al., 2007), the elastic net (Zou & Hastie, 2005), the adaptive least absolute shrinkage and selection operator (Zou, 2006; Wang & Leng, 2007; Zhang & Lu, 2007), one-step sparse estimation (Zou & Li, 2008) and the adaptive elastic net (Zou & Zhang, 2009). Many of these methods have been shown to be consistent for model selection, but under the constraint that the predictor dimension is smaller than the sample size. In contrast, if the predictor dimension is much larger than the sample size, none has been shown to be consistent for model selection under a general design condition (Leng et al., 2006; Zhao & Yu, 2006). Fan & Lv (2008) developed the theory of sure independence screening. Under a linear regression set-up and assuming the regression coefficient vector to be sparse, they show that marginal correlation estimation is effective for variable screening. Given that sure independence screening is computationally very simple, this nice property is not only practically useful but also theoretically appealing. Fan & Lv (2008) refer to it as the sure independence screening property, also referred to as screening consistency by Wang (2009). Despite its usefulness, sure independence screening suffers an important limitation: its consistency for model selection cannot be assured (Shao, 1997; Shi & Tsai, 2002). In fact, depending on the design conditions, it might perform nonignorable overfitting. In other words, to correctly discover all relevant variables, sure independence screening might have nonzero probabilities of including irrelevant variables, regardless of the sample size (Fan & Lv, 2008).

To achieve selection consistency, we propose a method of factor profiling, which naturally leads to factor profiled sure independence screening. This can be viewed as a combination of factor modelling and sure independence screening. As a result, factor profiled sure independence screening is strong in dimension reduction, which in turn leads to better variable selection. Furthermore, our method is also well motivated empirically. That is, for many ultrahigh-dimensional datasets, their first few eigenvalues are very often found to be substantially larger than the rest. For example, the supermarket dataset presented in § 3·5 has $n = 464$ observations and $p = 6398$ predictors. The leading eigenvalue of the sample covariance matrix accounts for about 35·4% of the total variability, while the second leading eigenvalue accounts for only 3·5% and the rest are even smaller. This suggests that the high-dimensional predictors' correlation structure might be represented by a low-dimensional latent factor model (Fan et al., 2008). Intuitively, if the latent factors can be estimated consistently, they can be profiled out from both the predictors and the responses. This operation is referred to as factor profiling. Factor profiling leads to uncorrelated predictors, which is referred to as factor profiled predictors. Applying standard sure independence screening (Fan & Lv, 2008) on factor profiled predictors leads to consistent model selection results (Shao, 1997; Shi & Tsai, 2002; Leng et al., 2006), even if the data are ultrahigh dimensional. For convenience, we refer to this method as factor profiled sure independence screening.

2. Factor profiling theory

2·1. *Models and notation*

Let $Y_i \in \mathbb{R}^1$ $(i = 1, \ldots, n)$ be the response collected from the ith subject and let $X_i = (X_{i1}, \ldots, X_{ip})^{\mathrm{T}} \in \mathbb{R}^p$ be the associated p-dimensional predictor. To model the regression relationship between Y_i and X_i, we assume that

$$Y_i = X_i^{\mathrm{T}}\theta + \varepsilon_i, \tag{1}$$

where ε_i has mean zero and variance σ_ε^2, and $\theta = (\theta_1, \ldots, \theta_p)^{\mathrm{T}} \in \mathbb{R}^p$ is a p-dimensional coefficient vector with true value $\theta_0 = (\theta_{01}, \ldots, \theta_{0p})^{\mathrm{T}} \in \mathbb{R}^p$. We assume that θ_0 is highly sparse, most of its elements being zero. To model the predictor's correlation structure, we assume that (Johnson & Wichern, 2003; Fan et al., 2008)

$$X_i = BZ_i + \tilde{X}_i, \tag{2}$$

where $Z_i = (Z_{i1}, \ldots, Z_{id})^{\mathrm{T}} \in \mathbb{R}^d$ is a d-dimensional latent factor, $B = (b_{jk}) \in \mathbb{R}^{p\times d}$ is the loading matrix, and $\tilde{X}_i = (\tilde{X}_{i1}, \ldots, \tilde{X}_{ip})^{\mathrm{T}} \in \mathbb{R}^p$ represents the information contained in X_i but missed by Z_i. For $\tilde{X}_i$ we assume that $\mathrm{cov}(\tilde{X}_i)$ is a diagonal matrix, so $\mathrm{cov}(\tilde{X}_{ij_1}, \tilde{X}_{ij_2}) = 0$ for any $j_1 \neq j_2$. We assume further that $E(Y_i) = E(X_{ij}) = E(\tilde{X}_{ij}) = 0$ and $\text{var}(Y_i) = \text{var}(X_{ij}) = 1 \geqslant \tilde{\sigma}_j^2 = \text{var}(\tilde{X}_{ij})$. In addition, we require $\text{cov}(Z_i) = I$, 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 redefine $Z_i = \text{cov}^{-1/2}(Z_i) Z_i$ and $B = B\text{cov}^{1/2}(Z_i)$, so that the condition $\text{cov}(Z_i) = I$ can be satisfied. We also allow ε_i to be correlated with X_i through the common factor Z_i,

$$\varepsilon_i = Z_i^{\mathrm{T}} \alpha + \tilde{\varepsilon}_i, \tag{3}$$

where $\alpha = (\alpha_1, \ldots, \alpha_d)^{\mathrm{T}} \in \mathbb{R}^d$ is a d-dimensional vector with true value $\alpha_0 \in \mathbb{R}^d$. Moreover, $\tilde{\varepsilon}_i$ is independent of both Z_i and $\tilde{X}_i$. We then should have $\text{var}(\tilde{\varepsilon}_i) = \tilde{\sigma}_\varepsilon^2 \leqslant \sigma_\varepsilon^2 \leqslant \text{var}(Y_i) = 1$.

Remark 1. Although models (2) and (3) are presented as two separate factor models, they can be combined as follows

$$X_i^+ = B^+ Z_i + \tilde{X}_i^+, \tag{4}$$

where $X_i^+ = (X_i^{\mathrm{T}}, \varepsilon_i)^{\mathrm{T}} \in \mathbb{R}^{(p+1)}$, $B^+ = (B^{\mathrm{T}}, \alpha)^{\mathrm{T}} \in \mathbb{R}^{(p+1) \times d}$ and $\tilde{X}_i^+ = (\tilde{X}_i^{\mathrm{T}}, \tilde{\varepsilon}_i)^{\mathrm{T}} \in \mathbb{R}^{(p+1)}$. Models like (4) have been previously assumed. For example, Johnson & Wichern (2003, Ch. 9) assume a similar model, with a fixed predictor dimension, and Fan et al. (2008) require the common factor to be observed while we assume it to be latent. Pan & Yao (2008) consider similar models, but under a time series set-up. As a result, they do not differentiate which component in X_i^+ is the response and which are predictors.

By (3), the residual ε_i may be correlated with X_i. As a result, the ordinary least squares estimate is biased, even if the predictor dimension is fixed and the sample size is infinite (Wooldridge, 2001). The story changes, however, if Z_i can be eliminated from both Y_i and X_i. Specifically, define a profiled response as $\tilde{Y}_i = Y_i - Z_i^{\mathrm{T}} \gamma_0$ with $\gamma_0 = B^{\mathrm{T}} \theta_0 + \alpha_0$. Next, refer to $\tilde{X}_i$ and $\tilde{\varepsilon}_i$ as a profiled predictor and noise, respectively. We then have

$$\tilde{Y}_i = \tilde{X}_i^{\mathrm{T}} \theta_0 + \tilde{\varepsilon}_i. \tag{5}$$

For model (5), not only are $\tilde{X}_i$ and $\tilde{\varepsilon}_i$ mutually uncorrelated, but different predictors are also mutually uncorrelated. The unknown regression coefficients can then be estimated consistently by sure independence screening. This is appealing because of its computational simplicity. In fact, as we prove later, its theoretical properties are also excellent.

The above discussion motivates us to develop a factor profiling methodology. Before we introduce the details, some notation is needed. Let $\mathbb{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}} \in \mathbb{R}^n$ be the response vector, $\mathbb{X} = (X_1, \ldots, X_n)^{\mathrm{T}} \in \mathbb{R}^{n \times p}$ the design matrix, and $\mathcal{E} = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}} \in \mathbb{R}^n$ the noise vector. Their profiled versions are defined similarly, and are denoted by $\tilde{\mathbb{Y}}$, $\tilde{\mathbb{X}}$ and $\tilde{\mathcal{E}}$. Next, define $\mathbb{X}_j = (X_{1j}, \ldots, X_{nj})^{\mathrm{T}} \in \mathbb{R}^n$ to be the jth column of $\mathbb{X}$. Similarly, $\tilde{\mathbb{X}}_j$ is the jth column of $\tilde{\mathbb{X}}$. By (1)–(3),

$$\mathbb{Y} = \mathbb{Z}\gamma_0 + \tilde{\mathbb{Y}} = \mathbb{Z}\gamma_0 + \tilde{\mathbb{X}}\theta_0 + \tilde{\mathcal{E}}, \quad \mathbb{X} = \mathbb{Z}B^{\mathrm{T}} + \tilde{\mathbb{X}}. \tag{6}$$

By (6), the effects due to $\mathbb{Z}$ can be eliminated if one can estimate $\mathcal{S}(\mathbb{Z})$ accurately, where $\mathcal{S}(A)$ stands for the linear subspace spanned by the column vectors of a matrix A. More specifically, if $\mathcal{S}(\mathbb{Z})$ is known, a projection matrix can be constructed onto its orthogonal complement. Denote this projection matrix by $Q(\mathbb{Z}) = I - H(\mathbb{Z}) \in \mathbb{R}^{n \times n}$, where $H(\mathbb{Z}) = \mathbb{Z}(\mathbb{Z}^{\mathrm{T}}\mathbb{Z})^{-1}\mathbb{Z}^{\mathrm{T}} \in \mathbb{R}^{n \times n}$ is another projection matrix but onto $\mathcal{S}(\mathbb{Z})$. We then get $Q(\mathbb{Z})\mathbb{Y} = Q(\mathbb{Z})\mathbb{X}\theta_0 + Q(\mathbb{Z})\mathcal{E}$, which serves as an approximation of the ideal model (5). Hence, 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 d_0 correctly. Because in practice d_0 is unknown, one has to estimate it based on data. As a simple and effective solution, we propose a maximum eigenvalue ratio criterion (Luo et al., 2009). Specifically, let $(\hat{\lambda}_j, \hat{V}_j)$ be the jth $(j = 1, \ldots, n)$ leading eigenvalue-eigenvector pair for the matrix $(np)^{-1}\mathbb{X}\mathbb{X}^{\mathrm{T}} \in \mathbb{R}^{n\times n}$. By definition, we should have $\hat{\lambda}_1 \geqslant \cdots \geqslant \hat{\lambda}_n$. Because the true factor dimension is d_0, we expect the first d_0 eigenvalues to be relatively large and 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 $1 \leqslant j \leqslant n-1$, we should expect its maximum value to occur at $j = d_0$. Consequently, the true structure dimension can be estimated by $\hat{d} = \mathrm{argmax}_{0\leqslant j\leqslant d_{\max}}(\hat{\lambda}_j/\hat{\lambda}_{j+1})$, where $d_{\max}$ is a prespecified maximum factor dimension. We call $\hat{d}$ a maximum eigenvalue ratio estimator. Although the idea of maximum eigenvalue ratio is intuitive and simple, whether it is statistically sound needs to be further justified theoretically. The following theorem gives the justification; see the Supplementary Material for a proof.

Theorem 1. *Under Conditions A1–A3 in the Appendix,* $\mathrm{pr}(\hat{d} = d_0) \to 1$ *as* $n \to \infty$.

2·3. *Estimating factor subspace*

By Theorem 1, the true factor dimension d_0 can be estimated consistently. With a correctly specified d_0, we can construct a least squares type objective function as

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

with $\beta_j = (b_{j1}, \ldots, b_{jd})^{\mathrm{T}} \in \mathbb{R}^d$. Since $B = (\beta_1, \ldots, \beta_p)^{\mathrm{T}} \in \mathbb{R}^{p\times d}$, $\mathcal{S}(\mathbb{Z})$ 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 $B = \hat{B} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)$ and $\hat{\beta}_j = (\mathbb{Z}^{\mathrm{T}}\mathbb{Z})^{-1}(\mathbb{Z}^{\mathrm{T}}\mathbb{X}_j) \in \mathbb{R}^d$, leading to the profiled objective function

$$\mathcal{O}(\mathbb{Z}) = \mathcal{O}(\mathbb{Z}, \hat{B}) = (np)^{-1}\sum_{j=1}^{p}\mathbb{X}_j^{\mathrm{T}}Q(\mathbb{Z})\mathbb{X}_j = (np)^{-1}\mathrm{tr}\{Q(\mathbb{Z})(\mathbb{X}\mathbb{X}^{\mathrm{T}})\}, \tag{7}$$

where $\mathrm{tr}(A)$ stands for the trace of a square matrix A. One can then verify that (7) can be minimized by setting $\hat{\mathbb{Z}} = (\hat{V}_1, \ldots, \hat{V}_d) \in \mathbb{R}^{n\times d}$; see Lemma 6 in the Supplementary Material.

Remark 2. Note that $\mathcal{O}(\mathbb{Z})$ is a function in $Q(\mathbb{Z}) = I - H(\mathbb{Z})$, where $H(\mathbb{Z})$ is a projection matrix on $\mathcal{S}(\mathbb{Z})$. As a result, $H(\hat{\mathbb{Z}}) = H(\hat{\mathbb{Z}}A)$ for any orthonormal matrix A. Consequently, $\hat{\mathbb{Z}}$ is not the unique minimizer for $\mathcal{O}(\mathbb{Z})$. Instead, $\hat{\mathbb{Z}}$ is just one possible minimizer, which is obtained by eigenvalue-eigenvector decomposition about $(np)^{-1}\mathbb{X}\mathbb{X}^{\mathrm{T}}$.

Subsequently, $\mathcal{S}(\mathbb{Z})$ can be estimated by $\mathcal{S}(\hat{\mathbb{Z}})$. To quantify the estimation accuracy of $\mathcal{S}(\hat{\mathbb{Z}})$, we consider the discrepancy measures

$$D_1(\mathbb{Z}, \hat{\mathbb{Z}}) = n^{-1}\mathrm{tr}\{\mathbb{Z}^{\mathrm{T}}Q(\hat{\mathbb{Z}})\mathbb{Z}\}, \quad D_2(\mathbb{Z}, \hat{\mathbb{Z}}) = \mathrm{tr}\{H(\mathbb{Z}) - H(\hat{\mathbb{Z}})\}^2.$$

Obviously, $\mathcal{S}(\hat{\mathbb{Z}}) = \mathcal{S}(\mathbb{Z})$ implies that $D_1(\mathbb{Z}, \hat{\mathbb{Z}}) = D_2(\mathbb{Z}, \hat{\mathbb{Z}}) = 0$. In addition, $D_2(\mathbb{Z}, \hat{\mathbb{Z}}) = 0$ implies that $\mathcal{S}(\mathbb{Z}) = \mathcal{S}(\hat{\mathbb{Z}})$; see Xia (2007) and Wang & Xia (2008). The first discrepancy measure, $D_1(\cdot, \cdot)$, is not symmetric in its arguments. In contrast, $D_2(\cdot, \cdot)$ is symmetric, so $D_2(\mathbb{Z}, \hat{\mathbb{Z}}) = D_2(\hat{\mathbb{Z}}, \mathbb{Z})$ for any $\mathbb{Z}$ and $\hat{\mathbb{Z}}$. These two measures are extensively used in subsequent theoretical development; see the Supplementary Material. The next theorem says that both of them converge towards zero at the rate $O_p(n^{-1})$.

THEOREM 2. *If $d = d_0$ and Conditions* A1–A3 *in the Appendix hold*, $D_1(\mathbb{Z}, \hat{\mathbb{Z}}) = O_p(n^{-1})$ *and* $D_2(\mathbb{Z}, \hat{\mathbb{Z}}) = O_p(n^{-1})$, *as $n \to \infty$.*

Remark 3. Let $A_1 \in \mathbb{R}^{n\times n}$ and $A_2 \in \mathbb{R}^{n\times n}$ be two arbitrary orthonormal matrices. Obviously, $H(\mathbb{Z}) = H(\mathbb{Z}A_1)$ and $H(\hat{\mathbb{Z}}) = H(\hat{\mathbb{Z}}A_2)$, so $D_2(\mathbb{Z}, \hat{\mathbb{Z}}) = D_2(\mathbb{Z}A_1, \hat{\mathbb{Z}}A_2)$, which suggests that D_2 is invariant for orthonormal transformation. The same conclusion also holds for D_1.

2·4. *Profiled independence screening*

We define a generic notation $\mathcal{M} = \{j_1, \ldots, j_{d^*}\}$ 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. We then define the full model as $\mathcal{M}_{\mathrm{F}} = \{j : j = 1, \ldots, p\}$ and the true model as $\mathcal{M}_{\mathrm{T}} = \{j : \theta_{0j} \neq 0\}$.

By Theorem 1, the factor dimension d_0 can be estimated consistently. With a correctly specified factor dimension $d = d_0$, Theorem 2 further indicates that the factor subspace $\mathcal{S}(\mathbb{Z})$ can be estimated accurately. Thereafter, we can get factor profiled data as $\hat{\mathbb{Y}} = Q(\hat{\mathbb{Z}})\mathbb{Y} \in \mathbb{R}^n$ and $\hat{\mathbb{X}} = Q(\hat{\mathbb{Z}})\mathbb{X}$, with $\hat{\mathbb{X}} = (\hat{\mathbb{X}}_1, \ldots, \hat{\mathbb{X}}_p) \in \mathbb{R}^{n\times p}$. Similarly, the factor profiled noise is $\hat{\mathcal{E}} = Q(\hat{\mathbb{Z}})\mathcal{E}$. Hence, sure independence screening can be applied to $\hat{\mathbb{Y}}$ and $\hat{\mathbb{X}}$ directly. As mentioned before, this method is referred to as factor profiled sure independence screening. More specifically, it estimates θ_j by $\hat{\theta}_j = (n^{-1}\hat{\mathbb{X}}_j^{\mathrm{T}}\hat{\mathbb{X}}_j)^{-1}(n^{-1}\hat{\mathbb{X}}_j^{\mathrm{T}}\hat{\mathbb{Y}})$. Without loss of generality, we assume further that the predictor indices have been appropriately relabelled so that $|\hat{\theta}_1| > \cdots > |\hat{\theta}_p|$. Then, a solution path is $\mathbb{M} = \{\mathcal{M}_{(k)} : k = 0, \ldots, p\}$ with $\mathcal{M}_{(0)} = \emptyset$ and $\mathcal{M}_{(k)} = \{1, \ldots, k\}$ for $k = 1, \ldots, p$. The following theorem implies that $\mathbb{M}$ and thus factor profiled sure independence screening are path consistent (Leng et al., 2006), that is, $\mathrm{pr}(\mathcal{M}_{\mathrm{T}} \in \mathbb{M}) \to 1$ as $n \to \infty$; see the Appendix for a proof.

THEOREM 3. *When $d = d_0$ and under Conditions* A1–A3 *in the Appendix,* $\max_{1\leqslant j\leqslant p} |\hat{\theta}_j - \theta_{0j}| = O_p\{n^{-1/2}(\log p)^{1/2}\}$ *as $n \to \infty$.*

Remark 4. One key advantage of factor profiled sure independence screening is that one step screening is sufficient for selection consistency (Shao, 1997), while this is generally not true for sure independence screening (Fan & Lv, 2008).

2·5. *Bayesian information criterion*

By the results of Theorem 3, factor profiled sure independence screening is path consistent, which implies that $\mathrm{pr}(\mathcal{M}_{\mathrm{T}} = \mathcal{M}_{(|\mathcal{M}_{\mathrm{T}}|)}) \to 1$ as $n \to \infty$. For a real-world application, however, the value of $|\mathcal{M}_{\mathrm{T}}|$ 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 most plausible. Towards this end, we propose the BIC-type criterion

$$\mathrm{BIC}(\mathcal{M}) = \log \mathrm{RSS}(\mathcal{M}) + |\mathcal{M}| \log n (n^{-1} \log p), \tag{8}$$

where $\mathrm{RSS}(\mathcal{M}) = \|\hat{\mathbb{Y}} - \sum_{j\in\mathcal{M}} \hat{\theta}_j \hat{\mathbb{X}}_j\|^2$ is the residual sum of squares. Then the best model can be selected as $\hat{\mathcal{M}} = \mathrm{argmin}_{\mathcal{M}\in\mathbb{M}} \mathrm{BIC}(\mathcal{M})$. Comparing (8) against the Bayesian information criterion (Schwarz, 1978),

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

we find that the only difference lies in the last penalization factor, where the classical Bayesian information criterion uses n^{-1} while we consider $n^{-1}\log p$. The classical criterion uses n^{-1} because the uniform convergence rate of the ordinary least squares estimator is $O_p(n^{-1/2})$, under a fixed dimension set-up (Shao, 1997). However, under an ultrahigh-dimensional set-up, the uniform convergence rate of the factor profiled sure independence screening estimator is $O_p\{n^{-1/2}(\log p)^{1/2}\}$ by Theorem 3. Consequently, a heavier penalty factor is inevitable (Chen & Chen, 2008). This motivates us to replace the traditional factor n^{-1} by $n^{-1}\log p$ in (8). Our numerical experience suggests that (8) works fairly well.

3. Numerical studies

3·1. *Simulation models*

Example 1. This example is borrowed from Fan & Lv (2008). We fix $d_0 = 1$, $p = 5000$ and $n = 150$. The latent factor Z_i is generated from $N(0, 1)$. The predictor X_i is then simulated according to (2), where $b_{jk} = 1$ and $\tilde{X}_i$ follows a p-dimensional standard normal distribution. Following Fan & Lv (2008), we assume the first $|\mathcal{M}_{\mathrm{T}}| = 3$ predictors to be relevant with coefficients $\theta_{0j} = 5$ for $j = 1, \ldots, |\mathcal{M}_{\mathrm{T}}|$. Consequently, $\theta_{0j} = 0$ for every $j > |\mathcal{M}_{\mathrm{T}}|$. Then, Y_i is generated according to (1), where ε_i follows (3) with $\alpha_0 = 0{\cdot}8\sigma_\varepsilon$ and $\tilde{\sigma}_\varepsilon = 0{\cdot}6\sigma_\varepsilon$. Last, σ_ε^2 is selected so that the signal-to-noise ratio, $\mathrm{var}(X_i^{\mathrm{T}}\theta_0)/\sigma_\varepsilon^2$, is 1, 2 or 5.

Example 2. This example is adapted from Fan & Lv (2008), but with a more sophisticated factor structure. For this example, we have $d_0 = 2$, $p = 10\,000$ and $n = 400$. The latent factor $Z_i \in \mathbb{R}^2$ is generated from a bivariate standard normal distribution. The predictor X_i is simulated as (2), where both b_{jk} and $\tilde{X}_{ij}$ are independent and distributed as $N(0, 1)$. Then, Y_i is generated according to (1), where $\theta_{0j} = (-1)^{R_{aj}}(4n^{-1/2}\log n + |R_{bj}|)$ for $j = 1, \ldots, |\mathcal{M}_{\mathrm{T}}| = 8$ and $\theta_{0j} = 0$ for every $j > |\mathcal{M}_{\mathrm{T}}|$. Here, R_{aj} is a binary random variable with $\mathrm{pr}(R_{aj} = 1) = 0{\cdot}4$ and R_{bj} is another $N(0, 1)$ variable. Lastly, ε_i is generated according to (3) with $\alpha_0 = 0{\cdot}8\sigma_\varepsilon(2^{-1/2}, 2^{-1/2})^{\mathrm{T}} \in \mathbb{R}^2$ and $\tilde{\sigma}_\varepsilon = 0{\cdot}6\sigma_\varepsilon$. For this example, the signal-to-noise ratio is 1, 2 or 5.

Example 3. This example is modified from Tibshirani (1996). We have $d_0 = 3$, $p = 10\,000$ and $n = 300$. The latent factor $Z_i \in \mathbb{R}^3$ is generated from a three-dimensional standard normal random vector. The predictor X_i is then simulated as (2), where $b_{jk} \sim N(0, 1)$. However, $\tilde{X}_i$ follows a p-dimensional normal distribution with $E(\tilde{X}_{ij}) = 0$ and $\mathrm{cov}(\tilde{X}_{ij_1}, \tilde{X}_{ij_2}) = 0{\cdot}5^{|j_1 - j_2|}$, and Y_i is simulated according to (1) with $\theta_{01} = 3$, $\theta_{04} = 1{\cdot}5$, $\theta_{07} = 2$ and $\theta_{0j} = 0$ for any $j \nparallel \in \mathcal{M}_{\mathrm{T}} = \{1, 4, 7\}$. Lastly, ε_i is generated according to (3) with $\alpha_0 = 0{\cdot}8\sigma_\varepsilon(3^{-1/2}, 3^{-1/2}, 3^{-1/2})^{\mathrm{T}} \in \mathbb{R}^3$ and $\tilde{\sigma}_\varepsilon = 0{\cdot}6\sigma_\varepsilon$. Again, the signal-to-noise ratio is 1, 2 or 5. For this example, the factor model (2) is not satisfied, because $\mathrm{cov}(\tilde{X}_i)$ is not diagonal. Thus, by including this example in our study, we evaluate the sensitivity of the proposed methods towards certain model misspecification.

3·2. *Factor estimation*

For each simulation model, 200 random replicates were generated. For each replication, the maximum eigenvalue ratio estimator was used to estimate the factor dimension. The percentage of the experiments with $\hat{d} = d_0$ is always 100%.

In addition, for each simulated dataset, we also get the estimated factor subspace $\mathcal{S}(\hat{\mathbb{Z}})$. Following Xia (2007) and Wang & Xia (2008), we then quantify its estimation error by $D(\mathbb{Z}, \hat{\mathbb{Z}}) = \lambda_{\max}\{H(\mathbb{Z}) - H(\hat{\mathbb{Z}})\}$, where $\lambda_{\max}(A)$ stands for the maximal absolute singular value of an arbitrary matrix A. The average values of $D(\mathbb{Z}, \hat{\mathbb{Z}})$, across the 200 simulation replications, are 1·42% for Example 1, 1·05% for Example 2 and 1·41% for Example 3. Those are very small numbers when compared with similar measures reported in the past literature; see, for example, Xia (2007) and Wang & Xia (2008). We thus know that $\mathcal{S}(\hat{\mathbb{Z}})$ can indeed capture the factor subspace $\mathcal{S}(\mathbb{Z})$ satisfactorily, which corroborates Theorem 2 very well. Qualitatively similar findings were obtained for $D_1(\mathbb{Z}, \hat{\mathbb{Z}})$ and $D_2(\mathbb{Z}, \hat{\mathbb{Z}})$.

3·3. *Selection consistency*

We next consider the performances of sure independence screening and factor profiled sure independence screening for variable selection. We use $\bar{\mathcal{M}}$ to represent a model selected by one particular method, in conjunction with the proposed Bayesian information criterion (8). Following Wang (2009), we evaluate the capability of $\bar{\mathcal{M}}$ in producing sparse solutions by

$$\%\text{ of Correct Zeros} = 100\% \times |(\mathcal{M}_{\mathrm{F}} \setminus \bar{\mathcal{M}}) \cap (\mathcal{M}_{\mathrm{F}} \setminus \mathcal{M}_{\mathrm{T}})||(\mathcal{M}_{\mathrm{F}} \setminus \mathcal{M}_{\mathrm{T}})|^{-1}.$$

This is only one aspect of $\bar{\mathcal{M}}$. A method with excellent capability in producing sparse solutions might also suffer from serious underfitting, for example, $\bar{\mathcal{M}} = \emptyset$. Thus, it is also important to evaluate a method's underfitting effect by

$$\%\text{ of Incorrect Zeros} = 100\% \times |(\mathcal{M}_{\mathrm{F}} \setminus \bar{\mathcal{M}}) \cap \mathcal{M}_{\mathrm{T}}||\mathcal{M}_{\mathrm{T}}|^{-1}.$$

The third and fourth columns in Table 1 report the average percentages of correct and incorrect zeros, respectively, across the 200 replications. We define $\bar{\mathcal{M}}$ to be a correctly fitted model if it is exactly the same as the true model, i.e., $\bar{\mathcal{M}} = \mathcal{M}_{\mathrm{T}}$. Then, the average percentages of correct fits are reported in the fifth column of Table 1. Next to this column, we present the average sizes of the selected models.

Remark 5. Sure independence screening is implemented by self-developed matlab codes, which can be freely obtained from their authors on request. For R users, a package SIS can be downloaded from http://cran.r-project.org/.

As one can see from Table 1, except for Example 1 with a signal-to-noise ratio of 1, factor profiled sure independence screening always greatly outperforms sure independence screening, in terms of both the percentages of incorrect zeros and of correct fits. This is not surprising, because the profiled predictors utilized by factor profiled sure independence screening are uncorrelated, which enables independence screening to demonstrate its best power. Moreover, by factor profiling, the problem of endogeneity is fixed. All numerical evidence suggests that factor profiled sure independence screening is competitive compared with sure independence screening.

3·4. *Estimation accuracy*

Lastly, we gauge the performance of different methods in terms of their estimation accuracy. We do not advocate the use of $\hat{\theta}_j$, as given in § 2·4, as our final estimator, because its estimation accuracy is not optimal (Wang, 2009). Thus, we propose the following ordinary least squares type estimator. 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.

Table 1. *Simulation results based on* 200 *replications with* $\alpha_0 \neq 0$

Signal to 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
	FP-SIS	100·0	95·8	0·5	0·1	14·6
2	SIS	100·0	70·3	0·0	1·0	21·3
	FP-SIS	100·0	46·3	40·0	1·6	7·9
5	SIS	100·0	67·0	0·0	1·0	18·4
	FP-SIS	100·0	0·2	99·5	3·0	1·0
Example 2						
1	SIS	100·0	95·9	0·0	1·0	17·6
	FP-SIS	100·0	75·6	1·0	1·9	11·5
2	SIS	100·0	93·2	0·0	1·0	16·6
	FP-SIS	100·0	47·4	15·0	4·2	7·2
5	SIS	100·0	90·8	0·0	1·1	15·9
	FP-SIS	100·0	16·6	34·5	6·7	2·7
Example 3						
1	SIS	100·0	92·2	0·0	1·0	8·4
	FP-SIS	100·0	53·5	9·5	1·4	3·4
2	SIS	100·0	88·0	1·0	1·0	7·7
	FP-SIS	100·0	34·5	24·5	2·0	2·2
5	SIS	100·0	81·8	1·0	1·1	6·8
	FP-SIS	100·0	20·2	50·5	2·4	1·4

SIS, sure independence screening; FP-SIS, factor profiled sure independence screening.

Specifically, for a selected model $\bar{\mathcal{M}}$, we define an ordinary least squares type estimator as $\hat{\theta}^{\bar{\mathcal{M}}} = (\hat{\theta}_1^{\bar{\mathcal{M}}}, \ldots, \hat{\theta}_p^{\bar{\mathcal{M}}})^{\mathrm{T}} \in \mathbb{R}^p$, where $\hat{\theta}_j^{\bar{\mathcal{M}}} = 0$ for every $j \notin \bar{\mathcal{M}}$ while

$$\hat{\theta}^{\bar{\mathcal{M}}}_{(\bar{\mathcal{M}})} = (\hat{\mathbb{X}}^{\mathrm{T}}_{(\bar{\mathcal{M}})} \hat{\mathbb{X}}_{(\bar{\mathcal{M}})})^{-1} (\hat{\mathbb{X}}^{\mathrm{T}}_{(\bar{\mathcal{M}})} \hat{\mathbb{Y}}). \tag{9}$$

The ordinary least squares estimator $\hat{\theta}^{\bar{\mathcal{M}}}_{(\bar{\mathcal{M}})}$ is computed based on either the profiled data $(\hat{\mathbb{X}}_{(\bar{\mathcal{M}})}, \hat{\mathbb{Y}})$ for factor profiled sure independence screening or the nonprofiled data $(\mathbb{X}_{(\bar{\mathcal{M}})}, \mathbb{Y})$ for sure independence screening. We can evaluate its estimation accuracy by the absolute estimation error $\sum_{j=1}^{p} |\theta_{0j} - \hat{\theta}_j^{\bar{\mathcal{M}}}|$, the average values of which are summarized in the last column of Table 1.

Table 1 shows that the absolute estimation error of factor profiled sure independence screening is always smaller than that of sure independence screening. This is not surprising because the latter does not fix the endogeneity problem, leading to poor model selection accuracy and thus unsatisfactory estimation accuracy.

Remark 6. It is natural to ask whether the relatively inferior performance of sure independence screening is due to the correlation between X_i and ε_i. To address this issue, we replicate the previous three simulation examples but with $\alpha_0 = 0$, so that ε_i is uncorrelated with X_i. Table 2 shows the results. As one can see, by fixing the endogeneity issue, the performance of sure independence screening is indeed improved slightly, in terms of the absolute estimation error. In contrast, the performance of profiled sure independence screening stays the same.

Table 2. *Simulation results based on* 200 *replications with* $\alpha_0 = 0$

Signal to 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	67·8	0·0	1·0	15·7
	PIS	100·0	95·8	0·5	0·1	14·6
2	SIS	100·0	66·3	0·0	1·0	15·2
	PIS	100·0	46·0	40·5	1·6	7·8
5	SIS	100·0	66·2	0·0	1·0	15·1
	PIS	100·0	0·2	99·5	3·0	1·0
Example 2						
1	SIS	100·0	92·4	0·0	1·0	16·1
	PIS	100·0	75·8	1·0	1·9	11·5
2	SIS	100·0	91·8	0·0	1·1	15·9
	PIS	100·0	47·3	15·0	4·2	7·2
5	SIS	100·0	91·6	0·0	1·1	15·9
	PIS	100·0	16·6	34·0	6·7	2·7
Example 3						
1	SIS	100·0	77·7	0·5	1·0	6·1
	PIS	100·0	53·5	9·5	1·4	3·4
2	SIS	100·0	77·8	0·5	1·0	6·1
	PIS	100·0	34·7	24·5	2·0	2·2
5	SIS	100·0	77·7	0·5	1·0	6·0
	PIS	100·0	20·5	50·5	2·4	1·4

SIS, sure independence screening; FP-SIS, factor profiled sure independence screening.

3·5. *Supermarket dataset*

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

As our first step, we estimate the dimension of the latent factor. The first eigenvalue of the matrix $(np)^{-1}\mathbb{X}\mathbb{X}^{\mathrm{T}}$ is $\hat{\lambda}_1 = 35{\cdot}4\%$, while the second is $\hat{\lambda}_2 = 3{\cdot}5\%$. The big difference between $\hat{\lambda}_1$ and $\hat{\lambda}_2$ suggests that the true factor dimension might be $d_0 = 1$, and this is confirmed by the maximum eigenvalue ratio estimator. We fix $d = 1$ and estimate the factor subspace $\mathcal{S}(\hat{\mathbb{Z}})$ and produce the profiled data $(\hat{\mathbb{Y}}, \hat{\mathbb{X}})$.

The value of θ_0 is unknown, so we use out-of-sample testing to compare the different methods' prediction accuracy. We conduct 200 random experiments. For each experiment, we randomly split the dataset $\mathcal{D} = \{1, \ldots, 464\}$ into two parts, $\mathcal{D} = \mathcal{D}_0 \cup \mathcal{D}_1$ 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 $\mathbb{Y}_1 = \{Y_i : i \in \mathcal{D}_1\} \in \mathbb{R}^{n_1}$. Notation for $(\hat{\mathbb{X}}_0, \hat{\mathbb{X}}_1)$, $(\hat{\mathbb{Y}}_0, \hat{\mathbb{Y}}_1)$ and $(\hat{\mathbb{Z}}_0, \hat{\mathbb{Z}}_1)$ is defined accordingly.

We apply sure independence screening with the Bayesian information criterion (8) to $(\mathbb{X}_0, \mathbb{Y}_0)$, which produces a candidate model $\mathcal{M}_{\text{SIS}}$. With the help of $\mathcal{M}_{\text{SIS}}$, we then obtain the ordinary least squares estimator $\hat{\theta}^{\mathcal{M}_{\text{SIS}}}$; see (9). Similar estimators are also obtained for factor profiled sure independence screening but based on an expanded design matrix, where $\hat{\mathbb{Z}}_0$ is also included.

The responses in the testing data are then predicted, and their median squared prediction error is computed for each random experiment. The medians of those median square prediction errors are then summarized across the 200 random experiments. This leads to 19·2% for sure independence screening and 13·6% for factor profiled sure independence screening, respectively. Factor profiled sure independence screening clearly outperforms sure independence screening. Across all the experiments, the sizes of the models selected by both sure independence screening and factor profiled sure independence screening models are always one, i.e. they have only one predictor.

4. Concluding remarks

Factor profiling can be viewed as an unsupervised dimension reduction technique, in the sense that the response is completely ignored for factor estimation. As a result, the factors identified by factor profiling might be good for describing a predictor's correlation structure but suboptimal for explaining the response; see Li (1991), Cook (1998), Li et al. (2007) and Zhu & Zhu (2009) for some discussion. Such a phenomenon never happens in our real data example, but it is a possibility, at least theoretically. Developing a supervised factor profiling method is an interesting future direction.

Acknowledgement

The author is very grateful to the editor, associate editor and two referees for their careful reading and constructive comments. This work was supported by the National Natural Science Foundation of China, the Fundamental Research Funds for the Central Universities, the Research Funds of Renmin University of China, PKU-Guanghua Harvest Institute for Investment and the Centre for Statistical Science at Peking University.

Supplementary material

Supplementary material available at *Biometrika* online includes details for Lemmas 1–7 and also proofs for Theorems 1–2.

Appendix

Technical conditions

To gain theoretical insights into the proposed factor profiling methods and to facilitate an easy proof, the following conditions are needed.

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

Condition A2 (Law of large numbers). Assume that both β_j and $\tilde\sigma_j^2$ admit $p^{-1}B^{\mathrm{T}}B = p^{-1}\sum\beta_j\beta_j^{\mathrm{T}} = \Sigma_\beta + O_p(p^{-1/2})$ and $p^{-1}\sum\tilde\sigma_j^2 = \tilde\sigma_0^2 + O_p(p^{-1/2})$, where $\Sigma_\beta\in\mathbb{R}^{d\times d}$ is a positive definite matrix and $\tilde\sigma_0^2 \geqslant \tilde\sigma^2_{\min} > 0$ is a positive constant.

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

Normality assumptions similar to Condition A1 have been popularly assumed in the past literature to facilitate a proof; see, for example, Fan & Lv (2008), Zhang & Huang (2008), Bickel & Levina (2008) and

Wang (2009). In particular, Condition A1 implies the exponential inequalities

$$\text{pr}\left[\left|p^{-1}\sum_{j=1}^{p}\{\tilde{X}_{i_1 j}\tilde{X}_{i_2 j} - E(\tilde{X}_{i_1 j}\tilde{X}_{i_2 j})\}\right| > \nu\right] \leqslant C_1\exp(-C_2 p\nu^2), \tag{A1}$$

$$\text{pr}\left\{\left|n^{-1}\sum_{i=1}^{n}\tilde{Y}_i\tilde{X}_{ij} - E(\tilde{Y}_i\tilde{X}_{ij})\right| > \nu\right\} \leqslant C_1\exp(-C_2 n\nu^2), \tag{A2}$$

$$\text{pr}\left(\left|n^{-1}\sum_{i=1}^{n}\tilde{X}_{ij}^2 - \tilde{\sigma}_j^2\right| > \nu\right) \leqslant C_1\exp(-C_2 n\nu^2), \tag{A3}$$

for any $\nu < \nu_0$, where ν_0 is some positive constant. The above inequalities play key roles in the theoretical treatment of essentially any type of ultrahigh-dimensional problem (Bickel & Levina, 2008, Lemma A.3). Condition A2 is satisfied if the (β_j, σ_j)s for different $j = 1, \ldots, p$ are generated independently from some distribution with finite moments. Lastly, by Condition A3, we require the predictor dimension p to be much larger than the sample size. Nevertheless, Condition A3 also requires that $n^{-1}\log p \to 0$, and a fixed true model size $|\mathcal{M}_\text{T}|$. In fact, we can allow $|\mathcal{M}_\text{T}| \to \infty$, as long as its speed of divergence is sufficiently slow; see, for example, Fan & Lv (2008) and Wang (2009). We decide to make a slightly stronger assumption. Otherwise, our already lengthy proof would be even more complicated.

Remark A1. The condition $p^{-1}B^\text{T}B = \Sigma_\beta + O_p(p^{-1/2})$ required by Condition A2 has an important theoretical implication: the factors included in Z_i must be so-called strong factors, meaning those factors contained in Z_i must be shared by a nontrivial portion of the predictors. In contrast, weak factors, which are shared by only a small set of predictors, cannot satisfy this condition, because the loading coefficients of a weak factor might be highly sparse. As a result, the probabilistic limit Σ_β cannot be positive definite. Intuitively, it is easy to find many weak factors shared by only a small set of predictors. However, it would be much more difficult to find a strong factor shared by many predictors. That makes the factor dimension d_0 unlikely to be very high.

Important lemmas

A number of lemmas are useful in the subsequent proofs. Only Lemma A1 is presented; details of Lemmas 1–6 can be found in the Supplementary Material. For convenience, the following notation needs to be defined. For an arbitrary 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 $\|A\|^2 = \text{tr}(A^\text{T}A) = \text{tr}(AA^\text{T})$.

Lemma A1. *Under Conditions A1–A3,* $\max_{1\leqslant j\leqslant p} n^{-1}\|\tilde{\mathbb{X}}_j\|^2 = O_p(1)$.

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

$$\begin{aligned}\text{pr}\left\{\max_{1\leqslant j\leqslant p}|\hat{\sigma}_j^2 - \tilde{\sigma}_j^2| > \kappa(n^{-1}\log p)^{1/2}\right\} &\leqslant \sum_{j=1}^{p}\text{pr}\left\{|\hat{\sigma}_j^2 - \tilde{\sigma}_j^2| > \kappa(n^{-1}\log p)^{1/2}\right\}\\ &\leqslant pC_1\exp(-C_2\kappa^2\log p)\\ &= C_1\exp\{(1 - C_2\kappa^2)\log p\}\\ &= C_1\exp(-\log p) \to 0, \quad p \to \infty.\end{aligned}$$

Consequently, $\max_j|\hat{\sigma}_j^2 - \tilde{\sigma}_j^2| = O_p\{n^{-1/2}(\log p)^{1/2}\}$. This completes the third step and finishes the proof. □

Proof of theorems

The proofs of Theorems 1 and 2 can be found in the Supplementary Material. To save space, only the proof for Theorem 3 is presented here.

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

Step 1. Recall that $\tilde{\mathbb{Y}} = \tilde{\mathbb{X}}\theta_0 + \tilde{\mathcal{E}}$ and $\hat{\mathbb{Y}} = Q(\hat{\mathbb{Z}})\mathbb{Z}\gamma_0 + Q(\hat{\mathbb{Z}})(\tilde{\mathbb{X}}\theta_0 + \tilde{\mathcal{E}})$. The difference between $\tilde{\mathbb{Y}}$ and $\hat{\mathbb{Y}}$ can be decomposed as

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

Then, by the Cauchy–Schwarz inequality,

$$\begin{aligned} n^{-1}\|\hat{\mathbb{Y}} - \tilde{\mathbb{Y}}\|^2/3 &\leqslant \gamma_0^{\mathrm{T}}\{n^{-1}\mathbb{Z}^{\mathrm{T}}Q(\hat{\mathbb{Z}})\mathbb{Z}\}\gamma_0 + n^{-1}(\tilde{\mathbb{X}}\theta_0 + \tilde{\mathcal{E}})^{\mathrm{T}}\{H(\hat{\mathbb{Z}}) - H(\mathbb{Z})\}^2(\tilde{\mathbb{X}}\theta_0 + \tilde{\mathcal{E}}) \\ &\quad + n^{-1}(\tilde{\mathbb{X}}\theta_0 + \tilde{\mathcal{E}})^{\mathrm{T}}H(\mathbb{Z})(\tilde{\mathbb{X}}\theta_0 + \tilde{\mathcal{E}}) \\ &\leqslant \lambda_{\max}\{n^{-1}\mathbb{Z}^{\mathrm{T}}Q(\hat{\mathbb{Z}})\mathbb{Z}\}\|\gamma_0\|^2 + n^{-1}\lambda_{\max}\{H(\hat{\mathbb{Z}}) - H(\mathbb{Z})\}^2\|\tilde{\mathbb{X}}\theta_0 + \tilde{\mathcal{E}}\|^2 \\ &\quad + n^{-1}(\tilde{\mathbb{X}}\theta_0 + \tilde{\mathcal{E}})^{\mathrm{T}}H(\mathbb{Z})(\tilde{\mathbb{X}}\theta_0 + \tilde{\mathcal{E}}). \end{aligned}$$

Define $\tilde{\sigma}_\theta^2 = \mathrm{var}(\tilde{X}_i^{\mathrm{T}}\theta_0 + \tilde{\varepsilon}_i) \leqslant \mathrm{var}(Y_i) = 1$. Then, by the definition of the discrepancy measures $D_1(,)$ and $D_2(,)$, we find that the right-hand side of the above inequality can be further bounded above by

$$D_1(\mathbb{Z}, \hat{\mathbb{Z}})\|\gamma_0\|^2 + D_2(\mathbb{Z}, \hat{\mathbb{Z}})n^{-1}\|\tilde{\mathbb{X}}\theta_0 + \tilde{\mathcal{E}}\|^2 + n^{-1}\tilde{\sigma}_\theta^{-2}(\tilde{\mathbb{X}}\theta_0 + \tilde{\mathcal{E}})^{\mathrm{T}}H(\mathbb{Z})(\tilde{\mathbb{X}}\theta_0 + \tilde{\mathcal{E}}). \tag{A4}$$

By the law of large numbers, we know that $n^{-1}\|\tilde{\mathbb{X}}\theta_0 + \tilde{\mathcal{E}}\|^2 = O_p(1)$. By Theorem 2 we know that $D_1(\mathbb{Z}, \hat{\mathbb{Z}}) = O_p(n^{-1})$ and $D_2(\mathbb{Z}, \hat{\mathbb{Z}}) = O_p(n^{-1})$. Furthermore, $\tilde{\sigma}_\theta^{-2}(\tilde{\mathbb{X}}\theta_0 + \tilde{\mathcal{E}})^{\mathrm{T}}H(\mathbb{Z})(\tilde{\mathbb{X}}\theta_0 + \tilde{\mathcal{E}})$ follows a chi-square distribution with d degrees of freedom. Applying these results to (A4), we have $n^{-1}\|\hat{\mathbb{Y}} - \tilde{\mathbb{Y}}\|^2 = O_p(n^{-1})$. Similarly, because $\mathbb{X}_j = \mathbb{Z}\beta_j + \tilde{\mathbb{X}}_j$, we have

$$n^{-1}\|\hat{\mathbb{X}}_j - \tilde{\mathbb{X}}_j\|^2/3 \leqslant D_1(\mathbb{Z}, \hat{\mathbb{Z}})\|\beta_j\|^2 + D_2(\mathbb{Z}, \hat{\mathbb{Z}})(n^{-1}\|\tilde{\mathbb{X}}_j\|^2) + n^{-1}\chi^2(d).$$

Taking maximum over j in both sides of this inequality, we find that

$$\max_{1\leqslant j\leqslant p}(n^{-1}\|\hat{\mathbb{X}}_j - \tilde{\mathbb{X}}_j\|^2)/3 \leqslant D_1(\mathbb{Z}, \hat{\mathbb{Z}}) + D_2(\mathbb{Z}, \hat{\mathbb{Z}})\max_j(n^{-1}\|\tilde{\mathbb{X}}_j\|^2) + n^{-1}\max_j \chi^2(d), \tag{A5}$$

partly because $\|\beta_j\|^2 = \mathrm{var}(\beta_j^{\mathrm{T}}Z_i) \leqslant \mathrm{var}(X_{ij}) = 1$. By Lemma A1, we know that $\max_j n^{-1}\|\tilde{\mathbb{X}}_j\|^2 = O_p(1)$. By Wang et al. (2009) we know that, with probability tending to one, $\max_{1\leqslant j\leqslant p}\chi^2(d) \leqslant 2\log(pd)$. Thus, the right-hand side of (A5) is $O_p(n^{-1}\log p)$. Hence

$$\max_{1\leqslant j\leqslant p}(n^{-1}\|\hat{\mathbb{X}}_j - \tilde{\mathbb{X}}_j\|^2) = O_p(n^{-1}\log p),\ n^{-1}\|\hat{\mathbb{Y}} - \tilde{\mathbb{Y}}\|^2 = O_p(n^{-1}). \tag{A6}$$

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

$$\begin{aligned} \max_{1\leqslant j\leqslant p}|\hat{\rho}_j - \hat{\rho}_j^*| &\leqslant n^{-1}\max_{1\leqslant j\leqslant p}|\hat{\mathbb{Y}}^{\mathrm{T}}(\hat{\mathbb{X}}_j - \tilde{\mathbb{X}}_j)| + n^{-1}\max_{1\leqslant j\leqslant p}|(\hat{\mathbb{Y}} - \tilde{\mathbb{Y}})^{\mathrm{T}}\tilde{\mathbb{X}}_j| \\ &\leqslant (n^{-1}\|\hat{\mathbb{Y}}\|^2)^{1/2}\max_j(n^{-1}\|\hat{\mathbb{X}}_j - \tilde{\mathbb{X}}_j\|^2)^{1/2} + (n^{-1}\|\hat{\mathbb{Y}} - \tilde{\mathbb{Y}}\|^2)^{1/2}\max_j(n^{-1}\|\tilde{\mathbb{X}}_j\|^2)^{1/2} \\ &= (n^{-1}\|\hat{\mathbb{Y}}\|^2)^{1/2}O_p\{n^{-1/2}(\log p)^{1/2}\} + O_p(n^{-1/2})\max_j(n^{-1}\|\tilde{\mathbb{X}}_j\|^2)^{1/2}, \end{aligned} \tag{A7}$$

due to (A6). Similarly, we have $n^{-1}\|\hat{\mathbb{Y}}\|^2 \leqslant 2n^{-1}\|\tilde{\mathbb{Y}}\|^2 + 2n^{-1}\|\tilde{\mathbb{Y}} - \hat{\mathbb{Y}}\|^2 = O_p(1)$. Next note that $n^{-1}\|\tilde{\mathbb{Y}}\|^2 = \mathrm{var}(\tilde{Y}_i) + o_p(1) \leqslant \mathrm{var}(Y_i) = 1$. Furthermore, $n^{-1}\max_j \|\tilde{\mathbb{X}}_j\|^2 = O_p(1)$; see Lemma A1. Applying those results to (A7), we find that $\max_{1\leqslant j\leqslant p} |\hat{\rho}_j - \hat{\rho}_j^*| = O_p\{n^{-1/2}(\log p)^{1/2}\}$. Then the conclusion follows if we can further prove that $\max_{1\leqslant j\leqslant p} |\hat{\rho}_j^* - \tilde{\rho}_j| = O_p\{n^{-1/2}(\log p)^{1/2}\}$.

Step 3. By the exponential inequality (A2) as implied by Condition A1, we know that there exist two positive constants C_1 and C_2, such that $\mathrm{pr}(|\hat{\rho}_j^* - \tilde{\rho}_j| > \nu) \leqslant C_1 \exp(-C_2 n \nu^2)$, where ν is an arbitrary positive number. Let $\kappa = (2/C_2)^{1/2}$. Then, by Bonferroni's inequality,

$$\begin{aligned}
\mathrm{pr}\left\{\max_{1\leqslant j\leqslant p} |\hat{\rho}_j^* - \tilde{\rho}_j| > \kappa(n^{-1}\log p)^{1/2}\right\} &\leqslant \sum_{j=1}^{p} \mathrm{pr}\{|\hat{\rho}_j^* - \tilde{\rho}_j| > \kappa(n^{-1}\log p)^{1/2}\} \\
&\leqslant pC_1 \exp(-C_2\kappa^2 \log p) \\
&= C_1 \exp\{(1 - C_2\kappa^2)\log p\} \\
&= C_1 \exp(-\log p) \to 0,
\end{aligned}$$

as $p \to \infty$. Consequently, $\max_j |\hat{\rho}_j^* - \tilde{\rho}_j| = O_p\{n^{-1/2}(\log p)^{1/2}\}$. This completes the third step and finishes the proof.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd Int. Symp. Info. Theory*, Ed. B. N. Petrov and F. Csaki, pp. 267–81. Budapest: Akademia Kiado.

Bickel, P. J. & Levina, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **36**, 199–277.

Breiman, L. (1995). Better subset selection using nonnegative garrote. *Technometrics* **37**, 373–84.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Ann. Statist.* **24**, 2350–83.

Chen, J. & Chen, Z. (2008). Extended Bayesian information criterion for model selection with large model spaces. *Biometrika* **95**, 759–71.

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

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Assoc.* **96**, 1348–60.

Fan, J. & Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space (with discussion). *J. R. Statist. Soc.* B **70**, 849–911.

Fan, J. & Peng, H. (2004). On non-concave penalized likelihood with diverging number of parameters. *Ann. Statist.* **32**, 928–61.

Fan, J., Fan, Y. & Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *J. Economet.* **147**, 186–97.

Fu, W. J. (1998). Penalized regression: the bridge versus the lasso. *J. Comp. Graph. Statist.* **7**, 397–416.

Hjort, N. L. & Claeskens, G. (2003). Frequentist model average estimators (with discussion). *J. Am. Statist. Assoc.* **98**, 879–99.

Huang, J., Horowitz, J. & Ma, S. (2007). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Statist.* **36**, 587–613.

Johnson, R. A. & Wichern, D. W. (2003). *Applied Multivariate Statistical Analysis*, 5th ed. New York: Pearson Education.

Knight, K. & Fu, W. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.* **28**, 1356–78.

Leng, C., Lin, Y. & Wahba, G. (2006). A note on lasso and related procedures in model selection. *Statist. Sinica* **16**, 1273–84.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *J. Am. Statist. Assoc.* **86**, 316–27.

Li, L., Cook, R. D. & Tsai, C. L. (2007). Partial inverse regression. *Biometrika* **94**, 615–25.

Luo, R., Wang, H. & Tsai, C. L. (2009). Contour projected dimension reduction. *Ann. Statist.* **37**, 3743–78.

Pan, J. & Yao, Q. (2008). Modelling multiple time series via common factors. *Biometrika* **95**, 365–79.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461–4.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statist. Sinica* **7**, 221–64.

Shi, P. & Tsai, C. L. (2002). Regression model selection–a residual likelihood approach. *J. R. Statist. Soc.* B **64**, 237–52.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B **58**, 267–88.

Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *J. Am. Statist. Assoc.* **104**, 1512–24.

Wang, H. & Leng, C. (2007). Unified lasso estimation via least squares approximation. *J. Am. Statist. Assoc.* **101**, 1418–29.

Wang, H. & Xia, Y. (2008). Sliced regression for dimension reduction. *J. Am. Statist. Assoc.* **103**, 811–21.

Wang, H., Li, R. & Tsai, C. L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94**, 553–8.

Wang, H., Li, B. & Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *J. R. Statist. Soc.* B **71**, 671–83.

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

Xia, Y. (2007). A constructive approach to the estimation of dimension reduction directions. *Ann. Statist.* **35**, 2654–90.

Yuan, M. & Lin, Y. (2007). On the nonnegative garrote estimator. *J. R. Statist. Soc.* B **69**, 143–61.

Zhang, C. H. & Huang, J. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *Ann. Statist.* **36**, 1567–94.

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

Zhao, P. & Yu, B. (2006). On model selection consistency of lasso. *J. Mach. Learn. Res.* **7**, 2541–67.

Zhu, L. P. & Zhu, L. X. (2009). On distribution-weighted partial least squares with diverging number of highly correlated predictors. *J. R. Statist. Soc.* B **71**, 525–48.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Am. Statist. Assoc.* **101**, 1418–29.

Zou, H. & Hastie, T. (2005). Regression shrinkage and selection via the elastic net with application to microarrays. *J. R. Statist. Soc.* B **67**, 301–20.

Zou, H. & Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models (with discussion). *Ann. Statist.* **36**, 1509–33.

Zou, H. & Zhang, H. H. (2009). On the adaptive elastic–net with a diverging number of parameters. *Ann. Statist.* **37**, 1733–51.

[*Received January* 2011. *Revised October* 2011]