# Factor Profiling for Ultra-High Dimensional Variable Selection

**Hansheng Wang**

Guanghua School of Management
Peking University

*hansheng.gsm.pku.edu.cn*

# Basic Background

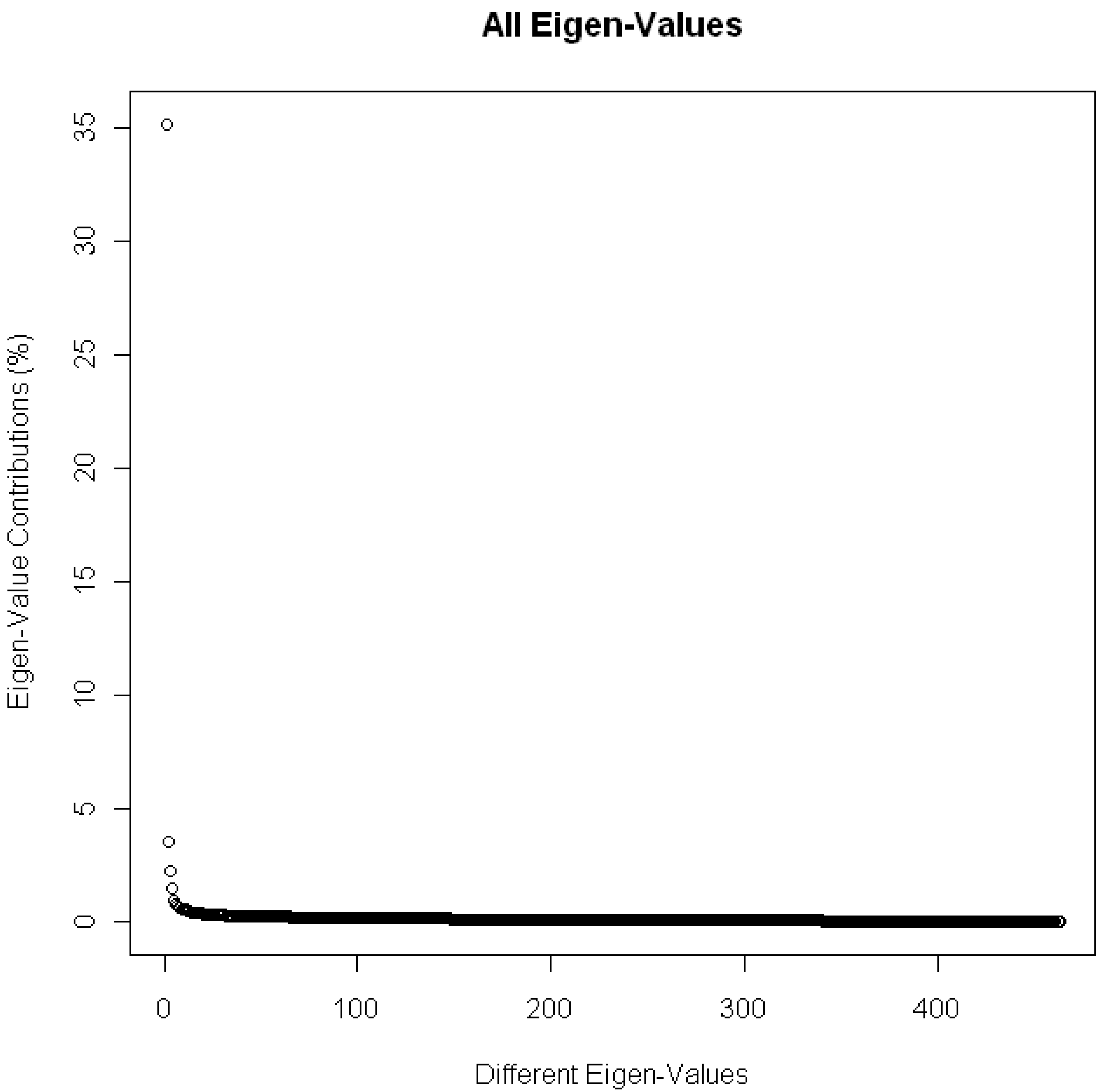
- Practical Motivation
  - Microarray
  - Supermarket
  - Search Engine
- Existing Methods
  - AIC and BIC
  - LASSO and SCAD
  - SIS and FR

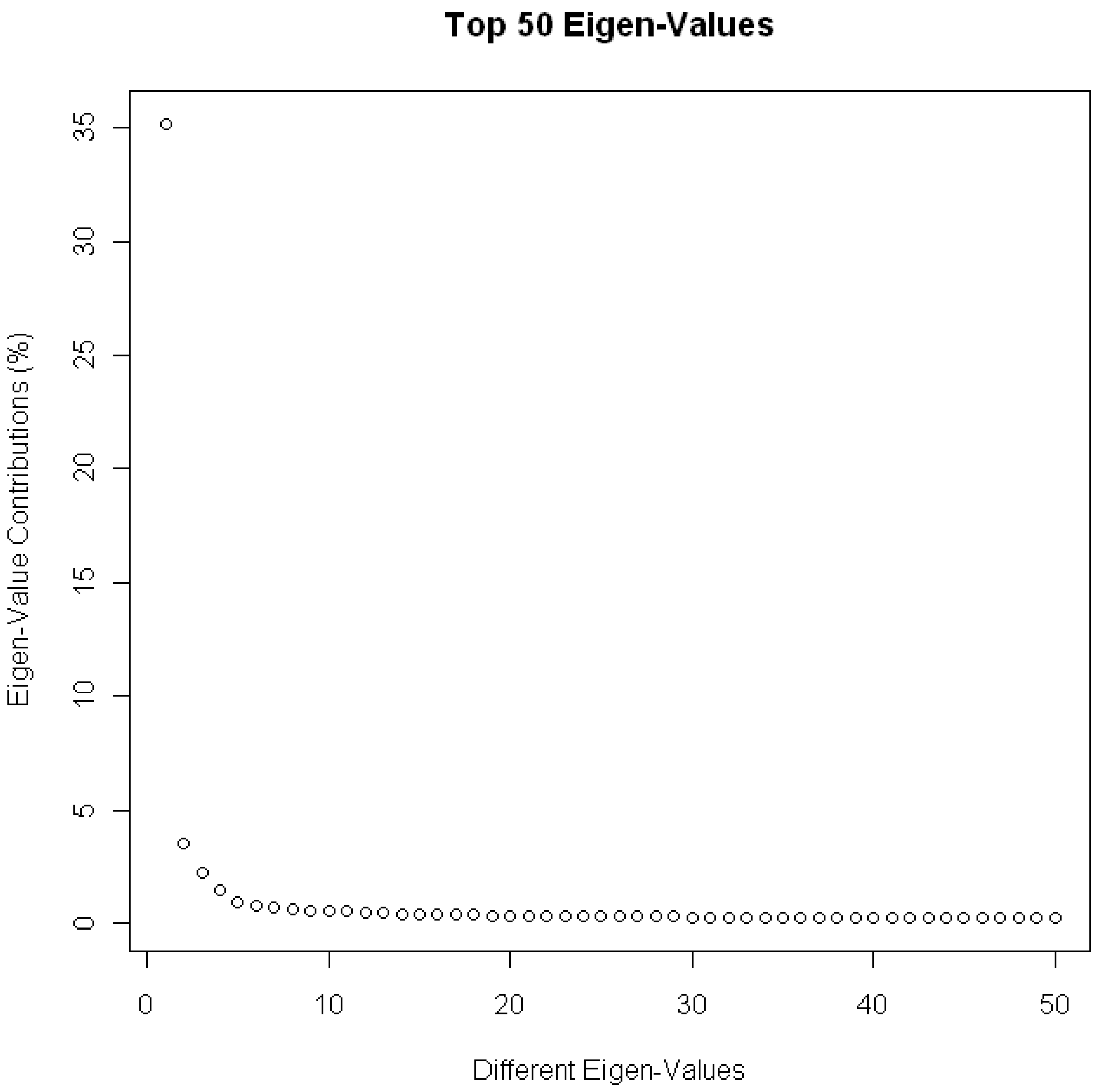
# Screening Methods

- SIS (Fan and Lv, 2008, JRSSB)
- FR (Wang, 2009, JASA)
- We typically wish cov(X) to be well behaved and better not to be highly singular.
- What is the real world?

# A Supermarket Example

- Data Resource:
  - A major domestic super market in Northern China.
- Response:
  - Daily customer volume for a total of 464 days.
- Predictor:
  - Daily sales volume for a total of 6398 products.
- Objective:
  - Predict next day's customer volume.

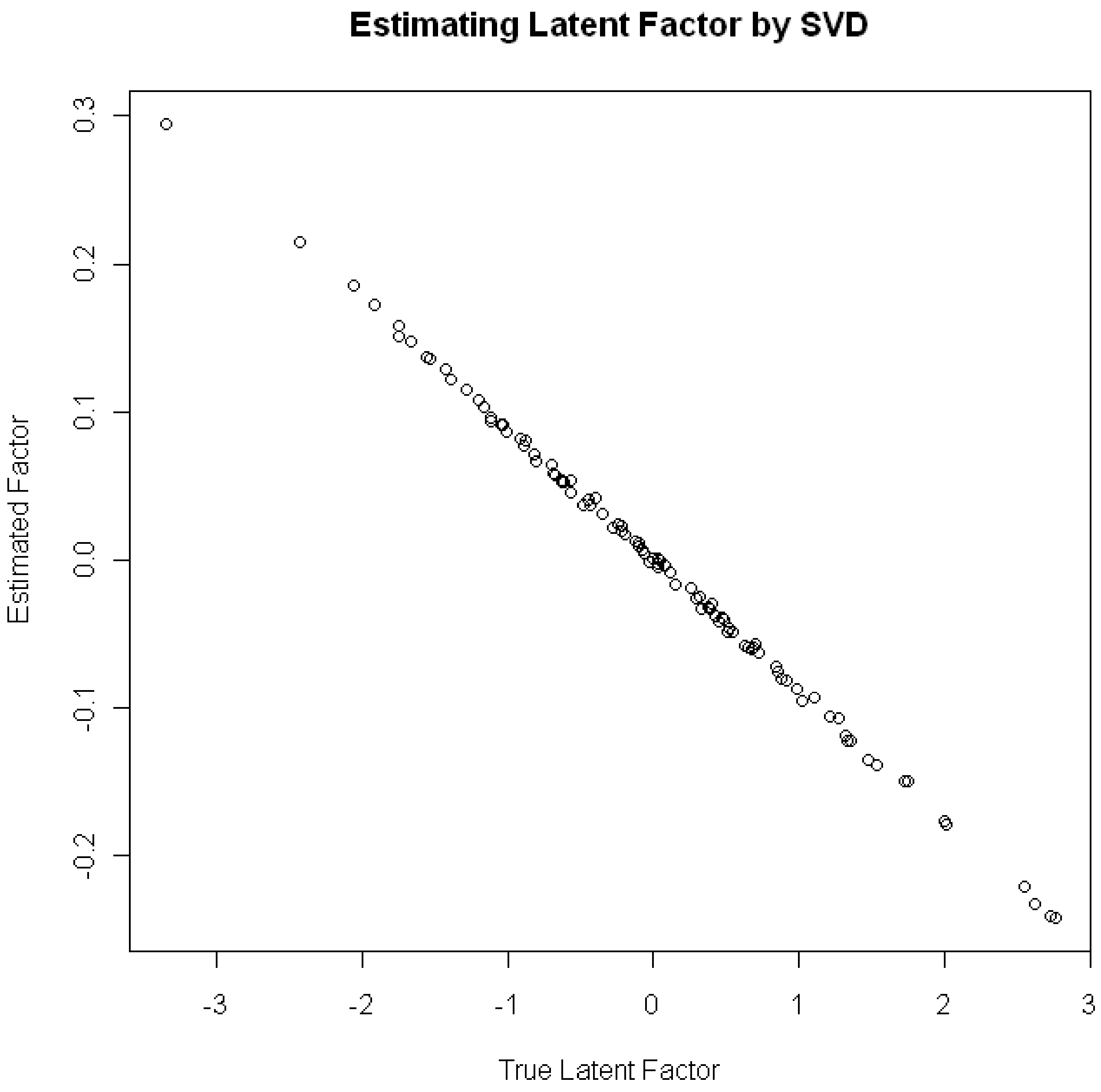

All Eigen-Values
Eigen-Value Contributions (%)
0
5
10
15
20
25
30
35
0
100
200
300
400
Different Eigen-Values



Top 50 Eigen-Values
Eigen-Value Contributions (%)
0
5
10
15
20
25
30
35
0
10
20
30
40
50
Different Eigen-Values

# A Simple Experiment

- Randomly generate a high dimensional data according to a very simple factor model
  - Sample Size = 100;
  - Predictor Dimension = 1000;
  - Factor Model: X=Latent Factor + Error
  - Estimation: Standard SVD
  - Question: Can we capture latent factor consistently or not?


Estimating Latent Factor by SVD
Estimated Factor
True Latent Factor
0.3
0.2
0.1
0.0
-0.1
-0.2
-3
-2
-1
0
1
2
3

# A Theoretical Framework

- To model the regression relationship between $Y_i$ and $X_i$, we assume that

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

where $\varepsilon_i$ is a random noise with mean 0 and variance $\sigma_\varepsilon^2$; $\theta = (\theta_1, \cdots, \theta_p)^\top \in \mathbb{R}^p$ 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 factor 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 $Z_i$.

# Endogeneity Issue

To reflect the endogeneity problem, we allow that $\varepsilon_i$ 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 $\text{var}(\tilde{\varepsilon}_i) = \tilde{\sigma}_\varepsilon^2 \leq \text{var}(Y_i) = 1$.

# Factor Profiling

- Profiled Response: $\widetilde{Y}_i = Y_i - Z_i^\top \gamma_0$ with $\gamma_0 = B^\top \theta_0 + \alpha_0$.
- Profiled Predictor and Noise: $\widetilde{X}_i$ and $\tilde{\varepsilon}_i$.
- Profiled Regression Model: $\widetilde{Y}_i = \widetilde{X}_i^\top \theta_0 + \tilde{\varepsilon}_i$.

# Estimating Factor Dimension

- Let $(\hat{\lambda}_j, \hat{V}_j)$ be the $j$th $(1 \leq j \leq n)$ leading eigenvalue-eigenvector pair for the matrix $\mathbb{X}\mathbb{X}^\top/(np) \in \mathbb{R}^{n\times n}$. Thus, 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

$$\hat{d} = \mathrm{argmax}_{0\leq j\leq d_{\max}}(\hat{\lambda}_j/\hat{\lambda}_{j+1}),$$

where $d_{\max}$ is a pre-specified maximum factor dimension.

# Theoretical Properties

**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* $n \to \infty$.

# Estimating Factor Subspace

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 $B \in \mathbb{R}^{p\times d}$.

# Estimation Accuracy

To quantify the estimation accuracy of $\mathcal{S}(\widehat{\mathbb{Z}})$, the following two discrepancy measures are considered. They are, respectively,

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

**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 $D_2(\mathbb{Z},\widehat{\mathbb{Z}}) = O_p(n^{-1})$.*

# Profiled Independent Screening

- With estimated $d_0$ and $\mathcal{S}(\mathbb{Z})$, 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 $\widehat{\mathbb{X}} = (\widehat{\mathbb{X}}_1, \cdots, \widehat{\mathbb{X}}_p) \in \mathbb{R}^{n\times p}$.
- 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 $\theta_j$ by $\hat{\theta}_j = (n^{-1}\widehat{\mathbb{X}}_j^\top \widehat{\mathbb{X}}_j)^{-1}(n^{-1}\widehat{\mathbb{Y}}^\top \widehat{\mathbb{X}}_j)$.

**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 $n \to \infty$.*

# A BIC 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,

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

where $\text{RSS}(\mathcal{M}) = \|\widehat{\mathbb{Y}} - \sum_{j\in\mathcal{M}} \hat{\theta}_j \widehat{\mathbb{X}}_j\|^2$ is the residual sum of squares. Then the best model can be selected as $\widehat{\mathcal{M}} = \text{argmin}_{\mathcal{M}\in\mathbb{M}} \text{BIC}(\mathcal{M})$.

# Profiled Sequential Screening

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 $k$th 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 $\hat{\theta}_j^{(k)} = \{\widehat{\mathbb{Y}}^{(k-1)\top}\widehat{\mathbb{X}}_j\}/\|\widehat{\mathbb{X}}_j\|^2$ and its correlation coefficient with the response as $\hat{\zeta}_j^{(k)} = \{\widehat{\mathbb{Y}}^{(k-1)\top}\widehat{\mathbb{X}}_j\}/\{\|\widehat{\mathbb{Y}}^{(k-1)}\| \cdot \|\widehat{\mathbb{X}}_j\|\}$.

(2.2) (*Screening*). We then find $a_k = \text{argmax}_{j \in \mathcal{M}_F \backslash \mathcal{S}^{(k-1)}} |\hat{\zeta}_j^{(k)}|$ 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 $\widehat{\mathbb{Y}}^{(k)} = \widehat{\mathbb{Y}}^{(k-1)} - \hat{\theta}_j^{(k)}\widehat{\mathbb{X}}_j$ 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 $\widehat{\mathcal{M}}^* = \text{argmin}_{\mathcal{M} \in \mathbb{M}^*} \text{BIC}(\mathcal{M})$.

# A Simulation Study

*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 $\varepsilon_i$ follows (2.3) with $\alpha_0 = 0.8\sigma_\varepsilon$ and $\tilde{\sigma}_\varepsilon = 0.6\sigma_\varepsilon$. Lastly, $\sigma_\varepsilon^2$ 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.

| 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 |

# Real Example: Factor Dimension

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.

# Out-of-Sample Testing

For a real problem like this, the value of $\theta_0$ 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.

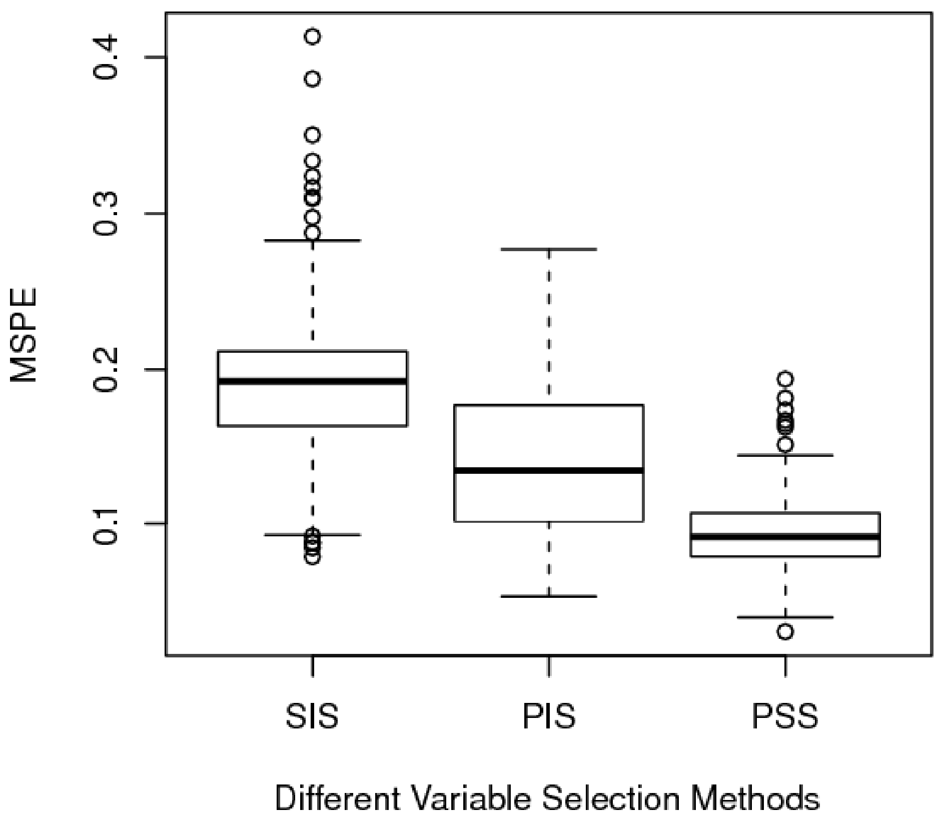



Figure 1: The real supermarket example. Boxplots for the median squared prediction errors (MSPE) based on 200 random replications.

Comments are very welcome!
Many thanks!