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 supermarket 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.
Top 50 Eigen-Values

Eigen-Value Contributions (%)

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: $X = \text{Latent Factor} + \text{Error}$
  - Estimation: Standard SVD
  - Question: Can we capture latent factor consistently or not?
Estimating Latent Factor by SVD

![Graph showing the relationship between True Latent Factor and Estimated Factor. The points form a downward diagonal line, indicating a strong negative correlation.](image-url)
A Theoretical Framework

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

$$Y_i = X_i^\top \theta + \varepsilon_i,$$  \hfill (2.1)

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

$$X_i = BZ_i + \tilde{X}_i,$$  \hfill (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 $\tilde{X}_i = (\tilde{X}_{i1}, \cdots, \tilde{X}_{ip})^\top \in \mathbb{R}^p$ 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

$$\varepsilon_i = Z_i^\top \alpha + \tilde{\varepsilon}_i,$$  \hspace{1cm} (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 $\tilde{X}_i$. We then should have $\text{var}(\tilde{\varepsilon}_i) = \tilde{\sigma}_\varepsilon^2 \leq \text{var}(Y_i) = 1.$
Factor Profiling

- Profiled Response: \( \tilde{Y}_i = Y_i - Z_i^T \gamma_0 \) with \( \gamma_0 = B^T \theta_0 + \alpha_0 \).

- Profiled Predictor and Noise: \( \tilde{X}_i \) and \( \tilde{\varepsilon}_i \).

- Profiled Regression Model: \( \tilde{Y}_i = \tilde{X}_i^T \theta_0 + \tilde{\varepsilon}_i \).
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 \(XX^T/(np) \in \mathbb{R}^{n \times n}\). 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

\[
\hat{d} = \arg \max_{0 \leq j \leq d_{\text{max}}} (\hat{\lambda}_j/\hat{\lambda}_{j+1}),
\]

where \(d_{\text{max}}\) is a pre-specified maximum factor dimension.
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

\[
\mathcal{O}(Z, B) = (np)^{-1} \sum_{j=1}^{p} ||X_j - Z\beta_j||^2
\]

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, \( S(Z) \) can be estimated by minimizing \( \mathcal{O}(Z, B) \) with respect to both \( Z \in \mathbb{R}^{n \times d} \) and \( B \in \mathbb{R}^{p \times d} \).
Estimation Accuracy

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

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

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

- With estimated $d_0$ and $S(Z)$, we can get factor profiled data as $\hat{Y} = Q(\hat{Z})Y \in \mathbb{R}^n$ and $\hat{X} = Q(\hat{Z})X$, with $\hat{X} = (\hat{X}_1, \ldots, \hat{X}_p) \in \mathbb{R}^{n \times p}$.

- Subsequently, the simple method of SIS can be applied to $\hat{Y}$ and $\hat{X}$ 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 $\hat{\theta}_j = (n^{-1}\hat{X}_j^T\hat{X}_j)^{-1}(n^{-1}\hat{Y}^T\hat{X}_j)$.

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

A BIC Criterion

Previous subsection proves that PIS is path consistent, which implies that \( P(\mathcal{M}_T = \mathcal{M}_T) \to 1 \) 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 \( \mathcal{M} \) is mostly plausible. To this end, we proposed here the following heuristic BIC-type selection criterion,

\[
\text{BIC}(\mathcal{M}) = \log \text{RSS}(\mathcal{M}) + |\mathcal{M}| \cdot \log n \cdot (\log p/n),
\]

(3.1)

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

Step (1) (Initialization). Set $\mathcal{M}^*_0 = \emptyset$ and $\hat{\mathbf{Y}}^{(0)} = \hat{\mathbf{Y}}$, i.e., the factor profiled response.

Step (2) (Sequential Screening).

2.1 (Estimation). In the $k$th step ($k \geq 1$), we are given $\mathcal{M}^*_k$ and also $\hat{\mathbf{Y}}^{(k-1)}$. Then, for every $j \in \mathcal{M}_F \setminus \mathcal{M}^*_k$, estimate its regression coefficient as $\hat{\delta}_j^{(k)} = (\hat{\mathbf{Y}}^{(k-1)\top} \hat{\mathbf{X}}_j) / \| \hat{\mathbf{X}}_j \|^2$ and its correlation coefficient with the response as $\hat{\gamma}_j^{(k)} = (\hat{\mathbf{Y}}^{(k-1)\top} \hat{\mathbf{X}}_j) / (\| \hat{\mathbf{Y}}^{(k-1)} \| \cdot \| \hat{\mathbf{X}}_j \|)$.

2.2 (Screening). We then find $a_k = \arg\max_{j \in \mathcal{M}_F \setminus \mathcal{S}^{(k-1)}} |\hat{\gamma}_j^{(k)}|$ and update $\mathcal{M}^*_k = \mathcal{M}^*_k \cup \{a_k\}$ accordingly.

2.3 (Elimination). According to $a_k$, we then get an updated response vector as $\hat{\mathbf{Y}}^{(k)} = \hat{\mathbf{Y}}^{(k-1)} - \hat{\delta}_{a_k}^{(k)} \hat{\mathbf{X}}_{a_k}$ 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 $\mathcal{M}^* = \{\mathcal{M}^*_k : 0 \leq k \leq n\}$ with $\mathcal{M}^*_k = \{a_1, \ldots, a_k\}$ for $k > 0$.

Step (4) (Model Selection). Select the best model as $\hat{\mathcal{M}}^* = \arg\min_{\mathcal{M} \in \mathcal{M}^*} \text{BIC}(\mathcal{M})$. 
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 $\tilde{X}_i$ 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., $\text{SNR} = \frac{\text{var}(X_i^T \theta_0)}{\sigma_\varepsilon^2}$, is given by 1, 2, or 5.
<table>
<thead>
<tr>
<th>Signal Noise Selection Ratio</th>
<th>Variable Selection Method</th>
<th>% of Correct Zeros</th>
<th>% of Incorrect Zeros</th>
<th>% of Correct Model Fit</th>
<th>Average Model Size</th>
<th>Absolute Estimation Error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SIS</td>
<td>100.0</td>
<td>97.2</td>
<td>0.0</td>
<td>1.0</td>
<td>25.4</td>
</tr>
<tr>
<td></td>
<td>PIS</td>
<td>100.0</td>
<td>95.8</td>
<td>0.5</td>
<td>0.1</td>
<td>14.6</td>
</tr>
<tr>
<td></td>
<td>PSS</td>
<td>100.0</td>
<td>95.8</td>
<td>0.5</td>
<td>0.1</td>
<td>14.6</td>
</tr>
<tr>
<td></td>
<td>SIS</td>
<td>100.0</td>
<td>70.3</td>
<td>0.0</td>
<td>1.0</td>
<td>21.3</td>
</tr>
<tr>
<td></td>
<td>PIS</td>
<td>100.0</td>
<td>46.3</td>
<td>40.0</td>
<td>1.6</td>
<td>7.9</td>
</tr>
<tr>
<td></td>
<td>PSS</td>
<td>100.0</td>
<td>43.3</td>
<td>45.5</td>
<td>1.7</td>
<td>7.4</td>
</tr>
<tr>
<td></td>
<td>SIS</td>
<td>100.0</td>
<td>67.0</td>
<td>0.0</td>
<td>1.0</td>
<td>18.4</td>
</tr>
<tr>
<td></td>
<td>PIS</td>
<td>100.0</td>
<td>0.2</td>
<td>99.5</td>
<td>3.0</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>PSS</td>
<td>100.0</td>
<td>0.0</td>
<td>100.0</td>
<td>3.0</td>
<td>0.9</td>
</tr>
</tbody>
</table>
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 $XX^T/(np)$ 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 $S(\hat{Z})$ can be estimated and the profiled data $(\hat{Y}, \hat{X})$ 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 $\mathcal{D} = \mathcal{D}_0 \cup \mathcal{D}_1$ 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 $X_0 = \{X_i : i \in \mathcal{D}_0\} \in \mathbb{R}^{n_0 \times p}$, $Y_0 = \{Y_i : i \in \mathcal{D}_0\} \in \mathbb{R}^{n_0}$, $X_1 = \{X_i : i \in \mathcal{D}_1\} \in \mathbb{R}^{n_1 \times p}$, and $Y_1 = \{Y_i : i \in \mathcal{D}_1\} \in \mathbb{R}^{n_1}$. Notations for $(\hat{X}_0, \hat{X}_1)$, $(\hat{Y}_0, \hat{Y}_1)$, and $(\hat{Z}_0, \hat{Z}_1)$ 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!