Robust high-dimensional principal component analysis and variable screening

Stefan Van Aelst and Yixin Wang

KU Leuven (University of Leuven)
High-dimensional linear regression

- Multiple linear regression model:
  \[ Y = X\beta + \varepsilon \]

  where \( X = (X_1, \ldots, X_p) \in \mathbb{R}^{n \times p}, Y \in \mathbb{R}^n, \varepsilon \in \mathbb{R}^n \) and \( \varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma_\varepsilon^2) \).

- High-dimensional: \( p \gg n \).
- Sparsity assumption: \( \beta \in \mathbb{R}^p \) contains many zeros.
- Goal: identify the subset \( M_J = \{1 \leq j \leq p : \beta_j \neq 0\} \).
- All variables are standardized
Factor profiling model

- $X_j$ may be correlated and also $\varepsilon$ may be correlated with $X$.
- Use a factor model to capture the correlation structure:

$$X = ZB^T + \tilde{X}$$
$$\varepsilon = Z\alpha + \tilde{\varepsilon}$$

with $Z^T Z = I_d$, $\text{Cov}(\tilde{X}_j, \tilde{X}_k) = 0$ $(j \neq k)$ and $\text{Cov}(\tilde{X}_j, \tilde{\varepsilon}) = 0$.
- The factor dimension $d$ is assumed to be small.
- The regression model becomes

$$Y - Z(B^T \beta + \alpha) = \tilde{X}^T \beta + \tilde{\varepsilon}$$
Factor profiling model

- Factor profiled regression model:

\[ \tilde{Y} = \tilde{X}^T \beta + \tilde{\epsilon} \]

with \( \text{Cov}(\tilde{X}_j, \tilde{X}_k) = 0 \) (\( j \neq k \)) and \( \text{Cov}(\tilde{X}_j, \tilde{\epsilon}) = 0 \).

- Factor profiled variables:

\[ \tilde{X} = X - ZB^T \]
\[ \tilde{Y} = Y - Z(B^T \beta + \alpha) \]

- Calculate LS coefficients: \( \hat{\beta}_j = (\tilde{X}_j^T \tilde{X}_j)^{-1} \tilde{X}_j^T \tilde{Y} \)

- FPSIS selects the set \( M_q = \{1 \leq j \leq p : |\hat{\beta}_j| \text{ is one of } q \text{ largest} \} \)
Estimating the factors

- FPSIS estimates the factors by least squares

\[
\arg\min_{Z,B} \|X - ZB^T\|^2
\]

under the constraint \(Z^TZ = I_d\) (Wang 2012).

- Consider eigen decomposition \((\lambda_j, U_j)\) of \(XX^T\), then a solution is \(\hat{Z} = (U_1, \ldots, U_{\hat{d}})^T\) with \(\hat{d} = \arg\max_{0 \leq j \leq d_{\text{max}}} (\lambda_j/\lambda_{j+1})\)
Robust Estimation of the Factor Subspace

- **Least squares:**

  \[
  \arg \min_{Z,B} \varnothing(Z,B) = \|X - ZB^T\|^2 = \sum_{i=1}^{n} \|x_i - Bz_i\|^2 = \sum_{i=1}^{n} \|r_i\|^2
  \]

- **Multivariate Least Trimmed Squares (Maronna 2005)**

  \[
  \varnothing(Z, B, \mu) = \sum_{i=1}^{h} \left( \|r_i\|^2 \right)_{i:n} = \sum_{i=1}^{h} \left( \|x_i - Bz_i - \mu\|^2 \right)_{i:n}
  \]
The MLTS subspace Algorithm

- Generate starting values $\hat{\mu}^{(0)}$, $\hat{Z}^{(0)}$, $\hat{B}^{(0)}$ and compute the corresponding $h$-subsets.
- (C-step) Iterative a few times:
  - Calculate updates $\hat{Z}^{(k)}$, $\hat{B}^{(k)}$ and $\hat{\mu}^{(j)}$ alternatingly by weighted least squares with other quantities fixed;
  - Update the $h$-subset based on $r_i(\hat{\mu}^{(k)}, \hat{Z}^{(k)}, \hat{B}^{(k)})$.
- Select $h$-subset(s) with smallest $\sum_{i=1}^{h}(||r_i||^2)_{i:n}$ and iterate further until convergence.
- Select $\hat{d}$ which minimizes a robustly modified information criterion proposed by Bai and Ng (2002).
Starting values

- Computationally efficient and robust initialization is obtained by using deterministic starting values (Hubert et al. 2012).
- Robustly standardize $X$ using the median and $Q_n$ estimates of the columns.
- Select the first $d$ PCs of the $h$ observations of $X$ with smallest norm of the robustly standardized variables in one of the following 5 transformed data sets:
  1. $T_1 = X$
  2. $T_2 = R$, where $R_j = \text{rank}(X_j)$
  3. $T_3 = S$, where $S_j = \tanh(X_j)$
  4. $T_4 = W$, where $W_j = \Phi^{-1}(V_j)$ and $V_j = (R_j - 1/3)/(n + 1/3)$;
  5. $T_5 = (x_1/\|x_1\|, \ldots, x_n/\|x_n\|)^T$. 
Outlier identification and reweighting

- **Orthogonal outliers**: Apply a robust Yeo-Johnson transformation (Van der Veeken 2010) to the $||r_i||$ and flag an observation as orthogonal outlier if its transformed residual $\geq \Phi^{-1}(0.975)$.

  $\rightarrow$ Recalculate factor subspace using the nonoutlying observations.

- **Score outliers**: Observations with robust distance $\geq \sqrt{\chi^2_{d,0.975}}$ are flagged as score outliers.
Robustly calculating the factor profiled variables

- FPSIS calculates the factor profiled variables as $\tilde{X} = Q(\tilde{Z})X$ and $\tilde{Y} = Q(\tilde{Z})Y$ with $Q(\tilde{Z}) = I_d - \tilde{Z}(\tilde{Z}^T\tilde{Z})^{-1}\tilde{Z}^T$.
- Robust FPSIS calculates the factor profiled predictors as $\tilde{X} = X - 1_n \hat{\mu}^T - \tilde{Z}\hat{B}^T$.
- To obtain the factor profiled response, robustly regress $Y$ on $\tilde{Z}$. 

Factor profiling
Outliers in the regression model

- Vertical outliers
- Leverage points corresponding to
  - Orthogonal outlier
  - Score outlier
  - Orthogonal + score outlier
- Effect on factor profiled regression model and marginal regression models
Effect of good leverage points: original model

(a) SO  (b) OO  (c) OO + SO
Effect of good leverage points: factor profiled model

(a) SO

(b) OO

(c) OO + SO

Factor profiling
Effect of good leverage points: marginal model for $X_1$

(a) SO  
(b) OO  
(c) OO + SO

Factor profiling
Effect of good leverage points: marginal model for $X_2$

(a) SO  
(b) OO  
(c) OO + SO
Effect of bad leverage points: original model

(a) SO  
(b) OO  
(c) OO + SO
Effect of bad leverage points: factor profiled model

(a) SO  (b) OO  (c) OO + SO
Effect of bad leverage points: marginal model for $X_1$

(a) SO

(b) OO

(c) OO + SO
Effect of bad leverage points: marginal model for $X_2$

(a) SO

(b) OO

(c) OO + SO

Factor profiling
Robust FPSIS

**Step 1.** Robustly standardize variables and estimate the factor model by MLTS. Identify orthogonal and score outliers.

**Step 2.** Calculate profiled predictors: \( \tilde{X}_j = X_j - \frac{1}{n} \hat{\mu}_j^T - \hat{Z} \hat{B}_j^T \).

**Step 3.** Robustly regress \( Y \) on \( \hat{Z} \), excluding the leverage points, to obtain initial factor profiled response \( \tilde{Y}^0 \).

**Step 4.** Check whether each score outlier is a good or a bad leverage point. Robustly regress \( Y \) on \( \hat{Z} \), excluding only bad leverage points, to obtain improved factor profiled response \( \tilde{Y} \).

**Step 5.** Robustly regress \( \tilde{Y} \) on each of the factor profiled predictors \( \tilde{X}_j \), excluding bad leverage points.

**Step 6.** Order the slope estimates: \( |\hat{\beta}_{(1)}| > \ldots > |\hat{\beta}_{(p)}| \) to obtain solution path.
Simulation study

- Simulation scheme (Wang 2012):
  - $N = 200$ datasets
  - $p = 1000, 10000$ and $n = 400$
  - $Y = \mathbf{X}\beta + \varepsilon$, $\mathbf{X} = \mathbf{Z}\mathbf{B}^\top + \tilde{\mathbf{X}}$, $\varepsilon = \mathbf{Z}\alpha + \tilde{\varepsilon}$
  - $\mathbf{B}, \tilde{\mathbf{X}}$ and $\mathbf{Z}$ are standard normal variables.
  - $d = 2$,
  - Coefficients: $|\mathcal{M}_T| = 8$;
     - $\beta_0j = (-1)^{R_{aj}}(4n^{-1/2}\log n + |R_{bj}|)$ for $j = 1, \ldots, |\mathcal{M}_T|$, where $R_{aj} \sim B(1,0.4)$ and $R_{bj} \sim N(0,1)$
     - $\beta_0j = 0$ for $j > |\mathcal{M}_T|$.
  - $\alpha = 0.8\sigma_\varepsilon(\sqrt{2}, \sqrt{2})^\top \in \mathbb{R}^2$, $\tilde{\varepsilon} \sim N(0, \tilde{\sigma}_\varepsilon^2)$, where $\tilde{\sigma}_\varepsilon = 0.6\sigma_\varepsilon$, with $\sigma_\varepsilon^2$ generated that the signal-to-noise ratio $c = \text{var}(\mathbf{X}_j^\top \beta)/\sigma_\varepsilon^2$ equals 1, 3 or 5.
Simulation study: outliers

- Contamination in predictors:
  - Orthogonal outliers: \( \mathbf{X}_{\text{OO}} \sim N(\mu_{\text{OO}}, \mathbf{I}_p) \) with 
    \( \mu_{\text{OO}} = 25(1, \ldots, 1, 0 \ldots, 0)^T \).
  - Score outliers: \( \mathbf{X}_{\text{SO}} = \mathbf{Z}_{\text{SO}} \mathbf{B}^T + \hat{\mathbf{X}} \) with \( \mathbf{Z}_{\text{SO}} \sim N(5 \cdot \mathbf{1}_d, \mathbf{I}_d) \)

- Contamination in response: \( y_{\text{OUT}} \sim (-1)^\delta N(300, 1) \) with 
  \( \delta \sim B(1, 0.5) \)

- Contamination levels:
  - 0%, no contamination.
  - 5% (good/bad) leverage points, no vertical outliers.
  - 5% (good/bad) leverage points + 5% extra vertical outliers.
  - 20% (good/bad) leverage points, no vertical outliers.
  - 20% (good/bad) leverage points + 10% extra vertical outliers.
## Simulation Results: SIS and FPSIS

Median of minimum model size to capture all important variables by SIS and FPSIS on clean data and contaminated data with 5% leverage points.

<table>
<thead>
<tr>
<th>$p$</th>
<th>method</th>
<th>$c$</th>
<th>clean</th>
<th>SO-leverage</th>
<th>OO-leverage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>good</td>
<td>bad</td>
</tr>
<tr>
<td>1000</td>
<td>SIS</td>
<td>1</td>
<td>902</td>
<td>916</td>
<td>924.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>905.5</td>
<td>905.5</td>
<td>917</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>888</td>
<td>904.5</td>
<td>913.5</td>
</tr>
<tr>
<td></td>
<td>FPSIS</td>
<td>1</td>
<td>13</td>
<td>12</td>
<td>916.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>8</td>
<td>8</td>
<td>919.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>8</td>
<td>8</td>
<td>921</td>
</tr>
<tr>
<td>10000</td>
<td>SIS</td>
<td>1</td>
<td>8994</td>
<td>9089.5</td>
<td>9138</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>9093.5</td>
<td>8989.5</td>
<td>9047.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>9077.5</td>
<td>9109</td>
<td>9019</td>
</tr>
<tr>
<td></td>
<td>FPSIS</td>
<td>1</td>
<td>97</td>
<td>94</td>
<td>8894</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>14</td>
<td>14</td>
<td>9024.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>11</td>
<td>11</td>
<td>8960.5</td>
</tr>
</tbody>
</table>
# Simulation Results: RFPSIS

Median of minimum model size to capture all important variables with RFPSIS in all the simulation settings.

<table>
<thead>
<tr>
<th>$p$</th>
<th>extra VO</th>
<th>$c$</th>
<th>clean</th>
<th>$\epsilon = 5%$</th>
<th>$\epsilon = 20%$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>SO-leverage</td>
<td>SO-leverage</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>good/bad</td>
<td>good/bad</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>good</td>
<td>good</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>bad</td>
<td>bad</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>good/bad</td>
<td>good/bad</td>
</tr>
<tr>
<td>1000</td>
<td>no</td>
<td>1</td>
<td>17</td>
<td>18.5</td>
<td>16.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>1</td>
<td>17</td>
<td>18.5</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>9</td>
<td>10</td>
<td>9.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>8</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>10000</td>
<td>no</td>
<td>1</td>
<td>120.5</td>
<td>149</td>
<td>188</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>18</td>
<td>26.5</td>
<td>24.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>13</td>
<td>17.5</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>1</td>
<td>120.5</td>
<td>225</td>
<td>247.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>18</td>
<td>31</td>
<td>44.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>13</td>
<td>20.5</td>
<td>25</td>
</tr>
</tbody>
</table>
Solution Path: SIS and FPSIS

\(p = 1000, c = 1\)

\(p = 1000, c = 3\)

\(p = 1000, c = 5\)

\(p = 10000, c = 1\)

\(p = 10000, c = 3\)

\(p = 10000, c = 5\)

Median of minimum model size to capture \(m\) important variables by SIS (solid) and FPSIS (dotted) with clean observations (●), good SO-leverage points (△), bad SO-leverage points (▲), good OO-leverage points (◆) or bad OO-leverage points (◇).
Solution Path: RFPSIS

Median of minimum model size to capture $m$ important variables by RFPSIS with good or bad SO-leverage points.
Median of minimum model size to capture $m$ important variables by RFPSIS with good or bad OO-leverage points.
Cellwise outliers

- Multivariate LTS subspace estimation (Maronna 2005)

\[
\mathcal{O}(\mathbf{Z}, \mathbf{B}, \mu) = \sum_{i=1}^{n} w_i \| \mathbf{r}_i \|^2 = \sum_{i=1}^{n} w_i \left[ \sum_{j=1}^{p} (x_{ij} - z_i^T \mathbf{b}_j - \mu_j)^2 \right]
\]

where \( w_i = 1 \) if \((\| \mathbf{r}_i \|^2)_{i:n} \leq (\| \mathbf{r}_i \|^2)_{h:n}\) and \( w_i = 0 \) otherwise.

- Componentwise LTS subspace estimation (Boente and Salibian-Barrera 2015)

\[
\mathcal{O}(\mathbf{Z}, \mathbf{B}, \mu) = \sum_{i=1}^{n} \sum_{j=1}^{p} w_{ij} r_{ij}^2 = \sum_{i=1}^{n} \sum_{j=1}^{p} w_{ij} (x_{ij} - z_i^T \mathbf{b}_j - \mu_j)^2
\]

where \( w_{ij} = 1 \) if \((r_{ij}^2)_{i:n} \leq (r_{ij}^2)_{h:n}\) and \( w_{ij} = 0 \) otherwise.
Cellwise robust FPSIS

**Step 1.** Robustly standardize variables and estimate the factor model by componentwise LTS. Identify in the predictor space:

- Outlying cells: \( \tilde{w}_{ij} = 0 \) if \( r_{ij} / Q_n(r_j) \leq \sqrt{\chi^2_{1,0.975}} \).
- Score outliers

**Step 2.** Calculate profiled predictors:

\[
\hat{\tilde{X}}_j = X_j - 1_n \hat{\mu}_j^T - \hat{Z} \hat{B}_j^T.
\]

**Step 3.** Robustly regress \( Y \) on \( \hat{Z} \), excluding score outliers, to obtain initial factor profiled response \( \hat{\tilde{Y}}^0 \).

**Step 4.** Check whether each score outlier is a good or a bad leverage point. Robustly regress \( Y \) on \( \hat{Z} \), excluding only bad leverage points, to obtain improved factor profiled response \( \hat{\tilde{Y}} \).

**Step 5.** Robustly regress \( \hat{\tilde{Y}} \) on each of the factor profiled predictors \( \hat{\tilde{X}}_j \), excluding bad leverage points and outlying cells (\( \tilde{w}_{ij} = 0 \)).

**Step 6.** Order the slope estimates: \( |\hat{\beta}_{(1)}| > \ldots > |\hat{\beta}_{(p)}| \) to obtain solution path.
Simulation Study: CRFPSIS

- Cellwise outlying predictors: $x_{ij}^c \sim N(25, 1)$
- Contaminated response: $y_{OUT} \sim (-1)^{\delta} \cdot N(300, 1)$
  with $\delta \sim B(1, 0.5)$
- Contamination level:
  - 5% contaminated cells in $X$ + no outliers in $Y$
  - 5% contaminated cells in $X$ + 5% outliers in $Y$

10% contaminated rows
$\epsilon \sim B(1, 0.5)$

25% contaminated rows
$\epsilon \sim B(1, 0.2)$

all the rows
$\epsilon \sim B(1, 0.05)$
## Simulation Results: CRFPSIS

Median of minimum model size to capture all important variables by RFPSIS and CRFPSIS on clean data and componentwise contaminated data.

<table>
<thead>
<tr>
<th>p</th>
<th>extra VO</th>
<th>c</th>
<th></th>
<th>sim a</th>
<th></th>
<th>sim b</th>
<th></th>
<th>sim c</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>clean</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RFPSIS</td>
<td>CRFPSIS</td>
<td>RFPSIS</td>
<td>CRFPSIS</td>
<td>RFPSIS</td>
<td>CRFPSIS</td>
</tr>
<tr>
<td>1000</td>
<td>no</td>
<td>1</td>
<td>17</td>
<td>41.5</td>
<td>24</td>
<td>29.5</td>
<td>36</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>9</td>
<td>13</td>
<td>10</td>
<td>11</td>
<td>13</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>8</td>
<td>10</td>
<td>9</td>
<td>9</td>
<td>11</td>
<td>9.5</td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>1</td>
<td>17</td>
<td>41.5</td>
<td>25</td>
<td>34</td>
<td>39</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>9</td>
<td>13</td>
<td>10</td>
<td>12</td>
<td>15</td>
<td>12.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>8</td>
<td>10</td>
<td>9</td>
<td>10</td>
<td>12</td>
<td>11</td>
</tr>
<tr>
<td>10000</td>
<td>no</td>
<td>1</td>
<td>120.5</td>
<td>393.5</td>
<td>217</td>
<td>353</td>
<td>369.5</td>
<td>341</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>18</td>
<td>74</td>
<td>31.5</td>
<td>64.5</td>
<td>70</td>
<td>63</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>13</td>
<td>40.5</td>
<td>21</td>
<td>38</td>
<td>45.5</td>
<td>34.5</td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>1</td>
<td>120.5</td>
<td>393.5</td>
<td>249.5</td>
<td>432.5</td>
<td>413.5</td>
<td>362</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>18</td>
<td>74</td>
<td>40.5</td>
<td>85</td>
<td>95.5</td>
<td>68.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>13</td>
<td>40.5</td>
<td>27</td>
<td>49.5</td>
<td>51</td>
<td>41</td>
</tr>
</tbody>
</table>
Conclusions

- Robust screening by RPFSIS works almost as good as FPSIS on clean data.
- SIS and FPSIS break down with a small fraction of outliers.
- RFPSIS can still discover the important predictors with a model size much smaller than the initial dimension when the data is contaminated.
- Important predictors with weak signals are harder to recover in presence of outliers.
- Robust subspace estimation can be adjusted to handle cellwise contamination.
- Robust BIC criteria have been developed to determine the final model size.
Thank you for your attention
References


