Prediction Performance of Single Index Principal Fitted Component Models

Jia-Ern Pai
&
Kofi Placid Adragni, PhD
Outline

• Introduction

• Inverse Reduction Models
  Principal Component model
  Principal Fitted Component model

• Principal Fitted Component Regression

• Prediction Comparisons

• Simulation Study
  - Simulation Setup
  - Simulation Results

• Conclusion
Introduction

• Frequently encountered problems in regression analyses
  o Large $p$ and small $n$ problems
    - When $n < p$, the OLS regression cannot provide stable parameter estimates
    - Example:
      Modeling the effect of upgraded Fuel System Integrity: age of vehicles, vehicle types, manufactures, model years, impact speeds, and driver’s ages…etc.
  o Collinearity problems
    - Inflated variances of estimates and predictions
    - Example:
      age of vehicles and model year

Dimension reduction is necessary
Introduction (cont.)

- \( \mathbf{X} \) is a \( p \)-vector random predictors, \( Y \) is an univariate response variable
  - Forward regression methods are usually adopted
    - Denoted as \( Y \mid \mathbf{X} \)
    - Examples:
      Partial Least Squares regression, LASSO regression, and principal component analysis, ..., etc.
- Utilizing the randomness property of \( \mathbf{X} \)
  - Inverse regression models would be another options.
    - Denoted as \( \mathbf{X} \mid Y \)
- Inverse reduction methods often provide more regression information between \( \mathbf{X} \) and \( Y \)

Research interest is on prediction performances of the Signal-Index-Isotropic Principal Fitted Component models
Introduction (cont.)

- Why are the inverse reduction methods more informative?
- Example: Principal Component Analysis
- Given a design matrix $X \in \mathbb{R}^{n \times p}$
  - From the sample $\text{cov}(X)$, we obtain:
    - Eigenvalues: $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_p$
    - Eigenvectors: $\hat{\gamma}_1, \hat{\gamma}_2, \ldots, \hat{\gamma}_p$
    - Principal components: $\{\hat{\gamma}_1^T X, \hat{\gamma}_2^T X, \ldots, \hat{\gamma}_p^T X\}$

- Main drawbacks of principal component analysis:
  - The response variable is not involved
  - The dimension reduction cannot be conducted when $\hat{\lambda}_i \approx \hat{\lambda}_j$, $\forall i, j$, such that $i \neq j$
Introduction (cont.)

• Goal of Dimension Reduction
  - \( \mathbf{X} \in \mathbb{R}^p \) and \( \mathbf{R}(\mathbf{X}) \in \mathbb{R}^d \), such that \( d \leq p \)

• What do we expect?
  - \( \mathbf{R}(\mathbf{X}) \) carries as much regression information as \( \mathbf{X} \) on \( Y \)

• Original regression model:
  \[
  Y \mid \mathbf{X} = \alpha^T \mathbf{X} + \varepsilon \quad (1)
  \]
  - Replace \( \mathbf{X} \) by \( \mathbf{R}(\mathbf{X}) \) without losing any regression information

  \[
  Y \mid \mathbf{X} = \beta^T \mathbf{R}(\mathbf{X}) + e \quad (2)
  \]
• Definition of the sufficient reduction (Cook, 2007)

• $X \in \mathbb{R}^p$ and $R(X) \in \mathbb{R}^d$, such that $d \leq p$

  - Inverse reduction, $X \mid (Y, R(X)) \sim X \mid R(X)$

  - Forward reduction, $Y \mid X \sim Y \mid R(X)$

  - Joint reduction, $X$ is independent of $Y \mid R(X)$

• If any condition holds, then $R(X)$ is a sufficient reduction
Inverse Reduction Models

- Principal Component models
- Suppose \( \mathbf{X} \in \mathbb{R}^p \), we regress \( \mathbf{X} \) on \( Y \)
  \[
  \mathbf{X} \mid y = \mathbf{\mu} + \Gamma \mathbf{v}_y + \mathbf{\epsilon} \quad (3)
  \]
  - \( \Gamma \) is a semi-orthogonal matrix:
    \[
    \Gamma^T \Gamma = \mathbf{I}_d, \text{ such that } d \leq p
    \]
  - \( \mathbf{v}_y \) is an unknown function of \( y \)
- \( \Gamma^T \mathbf{X} \) is a sufficient reduction in the Principal Component model
Inverse Reduction Models (cont.)

• Cook (2007) assumed $\mathbf{v}_y = \mathbf{f}_y \mathbf{y}$
  
  o $\mathbf{f}_y$ is a known flexible basis function of $y$
  
  o In practice, $\mathbf{f}_y$ can be determined by polynomial basis functions or piecewise polynomial basis functions

• Principal Fitted Component (PFC) models

  $\mathbf{X} | \mathbf{y} = \mathbf{\mu} + \mathbf{\Gamma} \mathbf{f}_y + \mathbf{\epsilon}$  \quad (4)

  o $\mathbf{\Gamma}^T \mathbf{X}$ is still a sufficient reduction
  
  o PFC model is model-based in this research
  
  o We assume $\mathbf{\epsilon} \sim N(0, \sigma^2 I_p)$
  
  - $\text{var}(\mathbf{\epsilon}) = \sigma^2 I_p$ : Isotropic error term
Inverse Reduction Models (cont.)

- To have fair and straightforward prediction performance comparisons with the OLS and LASSO regressions
  - Only one principal fitted component is used
  - Set $f_y = y$

- Single-Index-Isotropic Principal Fitted Component model
  \[ X|y = \mu + \Gamma \beta y + \varepsilon \]  
  - $\Gamma \in \mathbb{R}^{p \times 1}$ and $\Gamma^T \Gamma = 1$
  - $\beta \in \mathbb{R}$ and $E[Y] = 0$
  - $\varepsilon \sim N(0, \sigma^2 I_p)$

We only concern the isotropic PFC model in this research, so the term “single index PFC model” is adopted in the rest presentation content.
Inverse Reduction Models (cont.)

- \( X \mid y = \mu + \Gamma \beta y + \varepsilon \) \hspace{1cm} (6)

- A sufficient reduction in the single index PFC model is not unique
  - Example:
    - \( \Gamma^T X \equiv a \Gamma^T X \) in the single index PFC model, if \( a \neq 0 \)

- However, \( \text{span}(\Gamma) \) is unique
  - \( \text{span}(\Gamma) = \text{span}(a \Gamma) \)

- We should estimate \( \text{span}(\Gamma) \) instead of \( \Gamma \)

- We still need to have a \( \Gamma \) before finding \( \text{span}(\Gamma) \)

Parameter space in the single index PFC model:

- \( (\mu, \Gamma, \beta, \sigma^2) \)
  - Estimated by MLE
Principal Fitted Component Regression

• Consider a forward linear regression model
  \[ Y \mid X = \alpha^T X + e \]  \hspace{1cm} (6)

• \( \Gamma^T X \) is a sufficient reduction in the single index PFC model
  \[ Y \mid X = \beta(\Gamma^T X) + \varepsilon \]  \hspace{1cm} (7)

• \( \hat{\Gamma} \) is obtained from the single index PFC model

• Denote \( Z \) as \( \hat{\Gamma}^T X \)
  \[ Y \mid Z = \gamma Z + \varepsilon^* \]  \hspace{1cm} (8)

• Like model (6), model (7) a simple linear regression model
  - \( \hat{\Gamma}^T X \) is proxy of \( X \)

The procedure of replacing \( X \) by a sufficient reduction in a forward regression is called the Principal Fitted Component Regression (PFCR)
Prediction Comparisons

- How to make predictions with PFCR?
  - Given two data set \((X, Y)\) and \((X^*, Y^*)\)
    - \((X, Y)\) is used for the model building
    - \((X^*, Y^*)\) is used for making predictions
  - \((X, Y)\) and \((X^*, Y^*)\) are generated in the same way
- How to assess the prediction performance?
  - The sample mean squared prediction error (PE) is adopted

\[
PE = \frac{1}{n} \sum_{i} (Y_i^* - \hat{E}(Y | \hat{\Gamma}^T X_i^*))^2 \quad (9)
\]
Simulation Study

• Purpose
  Compare the prediction performances of the single index PFC model with other forward methods, such as the OLS, Ridge, LASSO, and Partial Least Square (PLS) regressions

• We use single index PLS model to make fair comparisons

• Scenarios
  o $n > p$ problem
    - Large $n$ case: All the predictors are response-related
  o $n < p$ problems
    - Dense case: All the predictors are response-related
    - Sparse case: Only some of predictors are response-related
Simulation Study (cont.)

• Data generation:

• We make $X$ as a linear function of $Y$

$$X = \beta(\Gamma Y)^T + \varepsilon \quad \text{(10)}$$

  - $\Gamma \in \mathbb{R}^{p \times 1}$, $\beta \in \mathbb{R}$
  - $Y=(y_1, y_2, \ldots, y_n)$, such that $y_i \sim N(0, \sigma_y^2)$
  - $\varepsilon \sim N(0, \sigma^2 I_p)$

• $\beta$ determines the strength of association between $X$ and $Y$

• $\beta$ is large, $X$ and $Y$ can reveal sufficient regression information to each other

  - Inverse and forward dimension reduction models should be able to find sufficient reductions more easily

• Changing different values of $\beta$, $p$, and $n$, distinct scenarios are created
Simulation Study (cont.)

• Iterating 100 times data generations, model buildings, PE calculations for every model in each distinct scenario

• From 100 PE’s
  
  - PE’s and SE(PE)’s can be calculated

• We present PE’s and SE(PE)’s as simulation results
Simulation Study (cont.)

- Large $n$ case
- Simulation set up

<table>
<thead>
<tr>
<th></th>
<th>level 1</th>
<th>level 2</th>
<th>level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n$</td>
<td>100</td>
<td>300</td>
<td>600</td>
</tr>
<tr>
<td>$\beta$</td>
<td>0.1</td>
<td>0.4</td>
<td>1</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma^2_y$</td>
<td></td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

- $\Gamma \in \mathbb{R}^{25 \times 1}$
- $\Gamma = \left(\frac{1}{\sqrt{25}}, \frac{1}{\sqrt{25}}, ..., \frac{1}{\sqrt{25}}\right)^T$
• Large $n$ case

<table>
<thead>
<tr>
<th></th>
<th>Beta = 0.1</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OLS</td>
<td>LASSO</td>
<td>Ridge</td>
<td>PLS</td>
<td>PFC</td>
</tr>
<tr>
<td>$n=100$</td>
<td>1.32(0.020)</td>
<td>1.00(0.014)</td>
<td>1.31(0.020)</td>
<td>1.17(0.016)</td>
<td>1.15(0.015)</td>
</tr>
<tr>
<td>$n=300$</td>
<td>1.08(0.010)</td>
<td>1.00(0.009)</td>
<td>1.08(0.010)</td>
<td>1.06(0.009)</td>
<td>1.05(0.008)</td>
</tr>
<tr>
<td>$n=600$</td>
<td>1.04(0.006)</td>
<td>1.00(0.004)</td>
<td>1.03(0.006)</td>
<td>1.03(0.005)</td>
<td>1.03(0.006)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Beta = 0.4</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OLS</td>
<td>LASSO</td>
<td>Ridge</td>
<td>PLS</td>
<td>PFC</td>
</tr>
<tr>
<td>$n=100$</td>
<td>1.15(0.018)</td>
<td>1.00(0.013)</td>
<td>1.14(0.018)</td>
<td>1.02 (0.019)</td>
<td>1.03(0.021)</td>
</tr>
<tr>
<td>$n=300$</td>
<td>0.97(0.008)</td>
<td>0.94(0.008)</td>
<td>0.97(0.008)</td>
<td>0.94(0.009)</td>
<td>0.95(0.009)</td>
</tr>
<tr>
<td>$n=600$</td>
<td>0.93(0.006)</td>
<td>0.91(0.006)</td>
<td>0.92(0.005)</td>
<td>0.90(0.006)</td>
<td>0.91(0.005)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Beta = 1</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OLS</td>
<td>LASSO</td>
<td>Ridge</td>
<td>PLS</td>
<td>PFC</td>
</tr>
<tr>
<td>$n=100$</td>
<td>0.66(0.018)</td>
<td>0.63(0.007)</td>
<td>0.65(0.015)</td>
<td>0.60(0.010)</td>
<td>0.60(0.009)</td>
</tr>
<tr>
<td>$n=300$</td>
<td>0.56(0.006)</td>
<td>0.53(0.006)</td>
<td>0.55(0.005)</td>
<td>0.53(0.006)</td>
<td>0.53(0.005)</td>
</tr>
<tr>
<td>$n=600$</td>
<td>0.53(0.003)</td>
<td>0.52(0.003)</td>
<td>0.53(0.003)</td>
<td>0.52(0.003)</td>
<td>0.52(0.003)</td>
</tr>
</tbody>
</table>
Simulation Study (cont.)

- Dense case

- Simulation set up

<table>
<thead>
<tr>
<th></th>
<th>level 1</th>
<th>level 2</th>
<th>level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p )</td>
<td>100</td>
<td>200</td>
<td>400</td>
</tr>
<tr>
<td>( \beta )</td>
<td>0.1</td>
<td>0.4</td>
<td>1</td>
</tr>
<tr>
<td>( n )</td>
<td></td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>( \sigma^2 )</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>( \sigma^2_y )</td>
<td></td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

- \( \Gamma \in \mathbb{R}^{p \times 1} \)

- \( \Gamma = (1,1,...,1)^T \)
Dense case

<table>
<thead>
<tr>
<th>p</th>
<th>Beta=0.1</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LASSO</td>
<td>PLS</td>
<td>PFC</td>
</tr>
<tr>
<td>p = 100</td>
<td>5.60 (2.000)</td>
<td>0.73 (0.010)</td>
<td>0.74 (0.011)</td>
</tr>
<tr>
<td>p = 200</td>
<td>0.82 (0.015)</td>
<td>0.61 (0.010)</td>
<td>0.61 (0.011)</td>
</tr>
<tr>
<td>p = 400</td>
<td>0.72 (0.012)</td>
<td>0.49 (0.009)</td>
<td>0.50 (0.009)</td>
</tr>
</tbody>
</table>

Beta=0.4

<table>
<thead>
<tr>
<th>p</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LASSO</td>
<td>PLS</td>
<td>PFC</td>
</tr>
<tr>
<td>p = 100</td>
<td>133.82 (130.234)</td>
<td>0.07 (0.001)</td>
<td>0.07 (0.001)</td>
</tr>
<tr>
<td>p = 200</td>
<td>0.10 (0.002)</td>
<td>0.04 (0.001)</td>
<td>0.04 (0.001)</td>
</tr>
<tr>
<td>p = 400</td>
<td>0.11 (0.002)</td>
<td>0.02 (0.000)</td>
<td>0.02 (0.000)</td>
</tr>
</tbody>
</table>

Beta=1

<table>
<thead>
<tr>
<th>p</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LASSO</td>
<td>PLS</td>
<td>PFC</td>
</tr>
<tr>
<td>p = 100</td>
<td>0.70 (0.152)</td>
<td>0.01 (0.000)</td>
<td>0.01 (0.000)</td>
</tr>
<tr>
<td>p = 200</td>
<td>0.02 (0.000)</td>
<td>0.01 (0.000)</td>
<td>0.01 (0.000)</td>
</tr>
<tr>
<td>p = 400</td>
<td>0.02 (0.000)</td>
<td>0.003 (0.000)</td>
<td>0.003 (0.000)</td>
</tr>
</tbody>
</table>
Simulation Study (cont.)

- Sparse case
- We only compare the prediction performances of the sparse single index PFCR to LASSO regression
  - PLS regression does not have coefficient shrinkage procedure
- Simulation set up

<table>
<thead>
<tr>
<th></th>
<th>level 1</th>
<th>level 2</th>
<th>level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p$</td>
<td>100</td>
<td>200</td>
<td>400</td>
</tr>
<tr>
<td>$\beta$</td>
<td>0.1</td>
<td>0.4</td>
<td>1</td>
</tr>
<tr>
<td>$p_0$</td>
<td></td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>$n$</td>
<td></td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>$\sigma_y^2$</td>
<td></td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

- $P_0$ is the number of active predictors
- $\Gamma \in \mathbb{R}^{p \times 1}$
- $\Gamma = (1, \ldots, 1, 0, \ldots, 0)^T$
- Sparse case
- PLS model is not considered, because it does not have a threshold procedure

<table>
<thead>
<tr>
<th></th>
<th>Beta=0.1</th>
<th></th>
<th>Beta=0.4</th>
<th></th>
<th>Beta=1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>LASSO</td>
<td>Sparse PFC</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p = 100$</td>
<td>1.41(0.281)</td>
<td>1.15(0.020)</td>
<td>1.92(0.595)</td>
<td>0.48(0.008)</td>
<td>0.61(0.100)</td>
</tr>
<tr>
<td>$p = 200$</td>
<td>1.05(0.020)</td>
<td>1.20(0.019)</td>
<td>0.54(0.010)</td>
<td>0.53(0.009)</td>
<td>0.12(0.002)</td>
</tr>
<tr>
<td>$p = 400$</td>
<td>1.04(0.017)</td>
<td>1.16(0.017)</td>
<td>0.57(0.010)</td>
<td>0.62(0.010)</td>
<td>0.12(0.002)</td>
</tr>
</tbody>
</table>

(7) (8) (9)
Conclusion

- In some large $n$ case, not all predictors are active
  - The PFCR is preferred

- The prediction performances of the signal index PFCR and signal index PLS regression are almost the same
  - Only under the assumption $f_y = y$
  - The PFCR is more flexible

- It seems that the LASSO regression provides unstable prediction performances when $p$ is close to $n$
  - Sparse PFCR is recommended
  - “lars” is used when simulating the prediction performance of the LASSO regression


Thank you

Jia-ern.pai@dot.org