Federal Committee on
Statistical Methodology Research Conference
November, 2013

Prediction Performance of Single Index Principal Fitted Component Models

Jia-Ern Pai
&
Kofi Placid Adragni, PhD

DEPARTMENT OF TRANSPORTATION
UNITED STATES OF AMERICA

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
  - 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.
  - Collinearity problems
    - Inflated variances of estimates and predictions
    - Example:
      age of vehicles and model year

Dimension reduction is necessary

Introduction (cont.)

- **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 **X**
  - Inverse regression models would be another options.
    - Denoted as $\mathbf{X} \mid Y$
- Inverse reduction methods often provide more regression information between **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 $\mathbb{X} \in \mathbf{R}^{n \times p}$
  - From the sample cov($\mathbb{X}$) , we obtain:
    - Eigenvalues: $\hat{\lambda}_1, \hat{\lambda}_2, ..., \hat{\lambda}_p$
    - Eigenvectors: $\hat{\boldsymbol{\gamma}}_1, \hat{\boldsymbol{\gamma}}_2, ..., \hat{\boldsymbol{\gamma}}_p$
    - Principal components: $\{\hat{\boldsymbol{\gamma}}_1^T \mathbb{X}, \hat{\boldsymbol{\gamma}}_2^T \mathbb{X}, ..., \hat{\boldsymbol{\gamma}}_p^T \mathbb{X}\}$

- 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
  - $\mathbf{X} \in \mathbf{R}^p$ and $\mathrm{R}(\mathbf{X}) \in \mathbf{R}^d$, such that $d \leq p$
- What do we expect?
  - $\mathrm{R}(\mathbf{X})$ carries as much regression information as $\mathbf{X}$ on Y
- Original regression model:

$$Y \mid \mathbf{X} = \boldsymbol{\alpha}^T \mathbf{X} + \boldsymbol{\varepsilon} \qquad (1)$$

- Replace $\mathbf{X}$ by $\mathrm{R}(\mathbf{X})$ without losing any regression information

$$Y \mid \mathbf{X} = \boldsymbol{\beta}^T \mathrm{R}(\mathbf{X}) + \boldsymbol{e} \qquad (2)$$

Introduction (cont.)

- Definition of the sufficient reduction (Cook, 2007)
- $\mathbf{X} \in \mathbf{R}^p$ and $\mathrm{R}(\mathbf{X}) \in \mathbf{R}^d$, such that $d \leq p$

  - Inverse reduction, $\mathbf{X} \mid (Y, \mathrm{R}(\mathbf{X})) \sim \mathbf{X} \mid \mathrm{R}(\mathbf{X})$

  - Forward reduction, $Y \mid \mathbf{X} \sim Y \mid \mathrm{R}(\mathbf{X})$

  - Joint reduction, $\mathbf{X}$ is independent of $Y \mid \mathrm{R}(\mathbf{X})$

- If any condition holds, then $\mathrm{R}(\mathbf{X})$ is a sufficient reduction

DEPARTMENT OF TRANSPORTATION UNITED STATES OF AMERICA

Inverse Reduction Models

- Principal Component models
- Suppose $\mathbf{X} \in \mathbf{R}^p$, we regress $\mathbf{X}$ on Y

$$\mathbf{X} \mid y = \boldsymbol{\mu} + \boldsymbol{\Gamma} \boldsymbol{\nu}_y + \boldsymbol{\varepsilon} \qquad (3)$$

  - $\boldsymbol{\Gamma}$ is a semi-orthogonal matrix:

    $\boldsymbol{\Gamma}^T \boldsymbol{\Gamma} = \mathbf{I}_d$, such that $d \leq p$

  - $\boldsymbol{\nu}_y$ is an unknown function of y

- $\boldsymbol{\Gamma}^T \mathbf{X}$ is a sufficient reduction in the Principal Component model

Inverse Reduction Models (cont.)

- Cook (2007) assumed $\boldsymbol{\nu}_y = \boldsymbol{\beta}\boldsymbol{f}_y$
  - $\boldsymbol{f}_y$ is a known flexible basis function of y
  - In practice, $\boldsymbol{f}_y$ can be determined by polynomial basis functions or piecewise polynomial basis functions

- Principal Fitted Component (PFC) models

$$\mathbf{X} \mid y = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\beta}\boldsymbol{f}_y + \boldsymbol{\varepsilon} \qquad (4)$$

  - $\boldsymbol{\Gamma}^T\mathbf{X}$ is still a sufficient reduction
  - PFC model is model-based in this research
  - We assume $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I}_p)$

    - $\mathrm{var}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}_p$: 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 $\boldsymbol{f}_y = y$
- Single-Index-Isotropic Principal Fitted Component model

$$\mathbf{X} \mid y = \boldsymbol{\mu} + \boldsymbol{\Gamma} \beta y + \boldsymbol{\varepsilon} \qquad (6)$$

  - $\boldsymbol{\Gamma} \in \mathbf{R}^{p \times 1}$ and $\boldsymbol{\Gamma}^T \boldsymbol{\Gamma} = 1$
  - $\beta \in \mathbf{R}$ and $\mathrm{E}[Y] = 0$
  - $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I}_p)$

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.)

- $\mathbf{X} \mid y = \boldsymbol{\mu} + \boldsymbol{\Gamma}\beta y + \boldsymbol{\varepsilon}$ (6)
- A sufficient reduction in the single index PFC model is not unique
  - Example:
    - $\boldsymbol{\Gamma}^T \mathbf{X} \equiv a\boldsymbol{\Gamma}^T \mathbf{X}$ in the single index PFC model, if $a \neq 0$
- However, $\text{span}(\boldsymbol{\Gamma})$ is unique
  - $\text{span}(\boldsymbol{\Gamma}) = \text{span}(a\boldsymbol{\Gamma})$
- We should estimate $\text{span}(\boldsymbol{\Gamma})$ instead of $\boldsymbol{\Gamma}$
- We still need to have a $\boldsymbol{\Gamma}$ before finding $\text{span}(\boldsymbol{\Gamma})$

Parameter space in the single index PFC model: $(\boldsymbol{\mu}, \boldsymbol{\Gamma}, \beta, \sigma^2)$

- Estimated by MLE

Principal Fitted Component Regression

- Consider a forward linear regression model

$$Y \mid \mathbf{X} = \boldsymbol{\alpha}^T \mathbf{X} + \boldsymbol{e} \qquad (6)$$

- $\boldsymbol{\Gamma}^T \mathbf{X}$ is a sufficient reduction in the single index PFC model

$$Y \mid \mathbf{X} = \beta(\boldsymbol{\Gamma}^T \mathbf{X}) + \boldsymbol{\varepsilon} \qquad (7)$$

- $\hat{\boldsymbol{\Gamma}}$ is obtained from the single index PFC model
- Denote Z as $\hat{\boldsymbol{\Gamma}}^T \mathbf{X}$

$$Y \mid \mathrm{Z} = \gamma \mathrm{Z} + \varepsilon^* \qquad (8)$$

- Like model (6), model (7) a simple linear regression model

  - $\hat{\boldsymbol{\Gamma}}^T \mathbf{X}$ 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 $(\mathbf{X}, Y)$ and $(\mathbf{X}^*, Y^*)$
  - $(\mathbf{X}, Y)$ is used for the model building
  - $(\mathbf{X}^*, Y^*)$ is used for making predictions
- $(\mathbf{X}, Y)$ and $(\mathbf{X}^*, Y^*)$ are generated in the same way
- How to assess the prediction performance?
- The sample mean squared prediction error (PE) is adopted
  - $$\mathrm{PE} = \frac{1}{n}\sum_i (Y_i^* - \hat{E}(Y \mid \hat{\boldsymbol{\Gamma}}^T \mathbf{X}_i^*))^2 \qquad (9)$$

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

  ○ $n > p$ problem

  - Large n case: All the predictors are response-related

  ○ $n < p$ 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 $\mathbb{X}$ as a linear function of **Y**

$$\mathbb{X} = \beta(\boldsymbol{\Gamma}\mathbf{Y})^T + \boldsymbol{\varepsilon} \qquad (10)$$

  - $\boldsymbol{\Gamma} \in \mathbf{R}^{p\times 1}$, $\beta \in \mathbf{R}$
  - $\mathbf{Y}=(y_1, y_2, \ldots, y_n)$, such that $y_i \sim N(0, \sigma_y^2)$
  - $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I}_p)$
- β determines the strength of association between $\mathbb{X}$ and **Y**
- β is large, $\mathbb{X}$ 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 β , *p*, and *n*, distinct scenarios are created

DEPARTMENT OF TRANSPORTATION UNITED STATES OF AMERICA

Simulation Study (cont.)

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

- From 100 PE's

  - $\overline{\text{PE}}$'s and SE(PE)'s can be calculated

- We present $\overline{\text{PE}}$'s and SE(PE)'s as simulation results

Simulation Study (cont.)

- Large n case

- Simulation set up

	level 1	level 2	level 3
n	100	300	600
β	0.1	0.4	1
p	25		
σ^2	1		
σ_y^2	1		

- $\mathbf{\Gamma} \in \mathbf{R}^{25\times 1}$

- $\mathbf{\Gamma} = (\frac{1}{\sqrt{25}}, \frac{1}{\sqrt{25}}, \ldots, \frac{1}{\sqrt{25}})^T$

- Large *n* case

(1)

Beta = 0.1					
	OLS	LASSO	Ridge	PLS	PFC
n = 100	1.32(0.020)	1.00(0.014)	1.31(0.020)	1.17(0.016)	1.15(0.015)
n = 300	1.08(0.010)	1.00(0.009)	1.08(0.010)	1.06(0.009)	1.05(0.008)
n = 600	1.04(0.006)	1.00(0.004)	1.03(0.006)	1.03(0.005)	1.03(0.006)

(2)

Beta = 0.4					
	OLS	LASSO	Ridge	PLS	PFC
n = 100	1.15(0.018)	1.00(0.013)	1.14(0.018)	1.02 (0.019)	1.03(0.021)
n = 300	0.97(0.008)	0.94(0.008)	0.97(0.008)	0.94(0.009)	0.95(0.009)
n = 600	0.93(0.006)	0.91(0.006)	0.92(0.005)	0.90(0.006)	0.91(0.005)

(3)

Beta = 1					
	OLS	LASSO	Ridge	PLS	PFC
n = 100	0.66(0.018)	0.63(0.007)	0.65(0.015)	0.60(0.010)	0.60(0.009)
n = 300	0.56(0.006)	0.53(0.006)	0.55(0.005)	0.53(0.006)	0.53(0.005)
n = 600	0.53(0.003)	0.52(0.003)	0.53(0.003)	0.52(0.003)	0.52(0.003)

DEPARTMENT OF TRANSPORTATION UNITED STATES OF AMERICA

Simulation Study (cont.)

- Dense case

- Simulation set up

	level 1	level 2	level 3
p	100	200	400
β	0.1	0.4	1
n	100		
σ^2	1		
σ_y^2	1		

- $\mathbf{\Gamma} \in \mathbf{R}^{p\times 1}$

- $\mathbf{\Gamma} = (1,1,\ldots,1)^T$

- Dense case

	Beta=0.1		
	LASSO	PLS	PFC
p = 100	5.60(2.000)	0.73(0.010)	0.74(0.011)
p = 200	0.82(0.015)	0.61(0.010)	0.61(0.011)
p = 400	0.72(0.012)	0.49(0.009)	0.50(0.009)

(4)

	Beta=0.4		
	LASSO	PLS	PFC
p = 100	133.82(130.234)	0.07(0.001)	0.07(0.001)
p = 200	0.10(0.002)	0.04(0.001)	0.04(0.001)
p = 400	0.11(0.002)	0.02(0.000)	0.02(0.000)

(5)

	Beta=1		
	LASSO	PLS	PFC
p = 100	0.70(0.152)	0.01(0.000)	0.01(0.000)
p = 200	0.02(0.000)	0.01(0.000)	0.01(0.000)
p = 400	0.02(0.000)	0.003(0.000)	0.003(0.000)

(6)

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

	level 1	level 2	level 3
p	100	200	400
β	0.1	0.4	1
p_0	10		
n	100		
σ^2	1		
σ_y^2	1		

- P_0 is the number of active predictors
- $\mathbf{\Gamma} \in \mathbf{R}^{p \times 1}$
- $\mathbf{\Gamma} = (1, \ldots, 1, 0, \ldots, 0)^T$

- Sparse case
- PLS model is not considered , because it does not have a threshold procedure

	Beta=0.1	
	LASSO	Sparse PFC
p = 100	1.41(0.281)	1.15(0.020)
p = 200	1.05(0.020)	1.20(0.019)
p = 400	1.04(0.017)	1.16(0.017)

(7)

	Beta=0.4	
	LASSO	Sparse PFC
p = 100	1.92(0.595)	0.48(0.008)
p = 200	0.54(0.010)	0.53(0.009)
p = 400	0.57(0.010)	0.62(0.010)

(8)

	Beta=1	
	LASSO	Sparse PFC
p = 100	0.61(0.100)	0.10(0.002)
p = 200	0.12(0.002)	0.11(0.002)
p = 400	0.12(0.002)	0.13(0.003)

(9)

DEPARTMENT OF TRANSPORTATION
UNITED STATES OF AMERICA

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

Reference

- Cook, R. D. (2007). Fisher Lecture: Dimension Reduction in Regression. Statistical Science, Vol. 22, No. 1, 1-26.

- Cook, R. D. & Forzani, L. (2008). Principal Fitted Components for Dimension Reduction in Regression. Statistical Science, Vol. 23, No. 4, 485-501.

- Adragni, K. P. & Cook, R. D. (2009). Sufficient dimension reduction and prediction in regression. Philosophical Transactions of the Royal Society A, Vol. 367, No. 1906, 4385-4405.

- Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society Series B (Methodological), Vol. 58, No. 1, 267-288.

- Hoerl, A. E. & Kennard, R. W. (1970). Ridge Regression: Applications to Nonorthogonal Problems. Technometrics, Vol. 12, No. 1, 69-82.

- Refaeilzadeh, P., Tang L., & Liu H. (2009) Cross-Validation. Encyclopedia of Database Systems, Part 3, 532-538.

- Anderson T. W. (2003). Principal Components. An Introduction to Multivariate Statistical Analysis, Chapter 11, 495-482.

Thank you

Jia-ern.pai@dot.org