PC Selection for Sparse FPCA

Hyunsung Kim

October 1, 2019

Department of Statistics
Chung-Ang University

Outline

1. Methods to Choose the Number of PCs

2. Simulation

Methods to Choose the Number of PCs

PVE

PVE(Proportion of Variance Explained)

$$PVE_i = \frac{\lambda_i}{\sum_{j=1}^{\infty} \lambda_j}$$

$$PVE = \frac{\sum_{j=1}^{K} \lambda_j}{\sum_{j=1}^{\infty} \lambda_j}$$

Select K, the number of PCs, where

$$PVE \geq 0.95$$

Leave-one-curve-out cross validation

$$LOOCV_i(K) = \frac{1}{N}\sum_{i=1}^{N}\|\mathbf{Y}_i - \widehat{\mathbf{Y}}_i^{-i}\|^2$$

where

$$\widehat{Y_i}^{-i}(t) = \hat{\mu}(t) + \sum_{k=1}^{K}\hat{\phi}_k^{-i}(t)\hat{\xi}_{ik}^{-i}$$

Select K, the number of FPCs, by minimizing the $LOOCV$ score.

LOOCV with Squared Loss

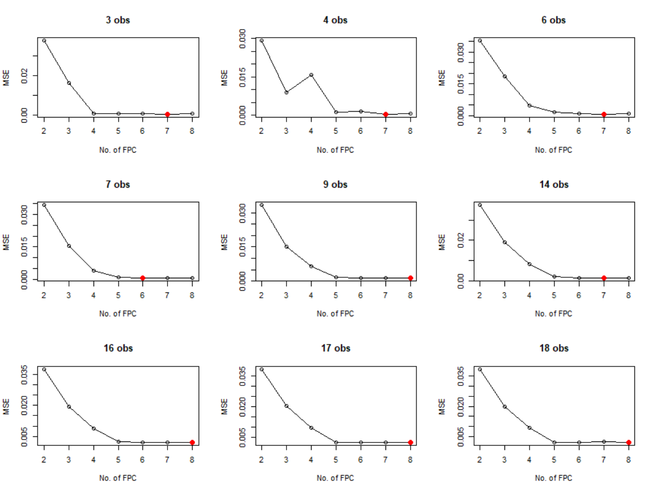

Figure 1: Estimated MSE for 1st training data

LOOCV with Kullback–Leibler Loss(Peng and Paul, 2009)

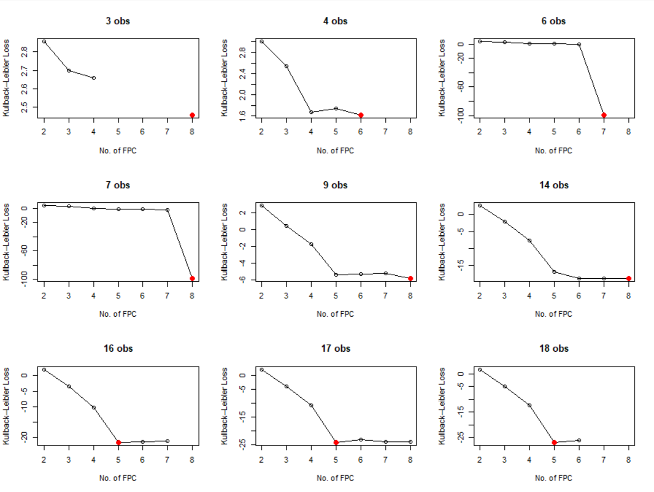

Figure 2: Estimated Kullback–Leibler divergence for 1st training data

Simulation

The Procedure of the Simulation

- Generate the 100 datasets from the temporal gene expression data and split the each dataset with training and test set.
- "Sparsify" the each dataset.
- Estimate the FPC functions and scores using the sparse FPCA method with 7 knots.
- Perform the 5 classification methods for the training sets, and predict for the test sets with the different number of FPCs.
- Choose the number of FPCs satisfied $PVE \geq 0.95$

Simulation Results

Table 1: Accuracy using FPCs selected by PVE

No. of obs	Logistic	SVM(Linear)	SVM(Gaussian)	SVM(Sigmoid)	SVM(Poly)	K	PVE
2	0.645	0.653	0.623	0.622	0.607	3.16	0.92
3	0.783	0.784	0.745	0.751	0.739	3.78	0.94
4	0.848	0.846	0.800	0.834	0.803	4.15	0.97
5	0.899	0.898	0.857	0.883	0.856	4.62	0.98
6	0.894	0.895	0.854	0.879	0.859	4.97	0.99
7	0.915	0.913	0.879	0.899	0.879	4.99	0.99
8	0.910	0.912	0.876	0.893	0.885	5.03	0.99
9	0.916	0.917	0.880	0.905	0.894	5.03	0.99
10	0.917	0.919	0.884	0.904	0.886	5.00	0.99
11	0.922	0.923	0.887	0.909	0.892	5.00	0.99
12	0.921	0.925	0.889	0.906	0.891	5.00	0.99
13	0.919	0.921	0.888	0.908	0.892	5.00	0.99
14	0.922	0.924	0.891	0.908	0.892	5.00	0.99
15	0.921	0.923	0.886	0.906	0.894	5.00	0.99
16	0.923	0.923	0.889	0.906	0.894	5.00	0.99
17	0.922	0.924	0.888	0.905	0.888	5.00	0.99
18	0.923	0.926	0.891	0.908	0.893	5.00	0.99
Average	0.888	0.890	0.853	0.872	0.855	4.75	0.98

Summary of Results

- Using PVE, almost 5 FPCs are selected.
- The selected FPCs explained about 99% of total variability except $N_i \leq 5$.
- The linear SVM perform well than other kernel SVM methods.
- If there are about 7 out of 18 observations, the model answered more than 90% correctly.

Conclusion

- LOOCV with squared loss doesn't look like the reliable method.
- Also, LOOCV's computation time is very slow, even though used parallel computing.
- LOOCV with Kullback–Leibler loss looks a better measure than squared loss, but `fpca.mle` function in `fpca` package is very unstable.
- PVE is the more useful method than LOOCV with squared loss in terms of dimension reduction.

Reference

Reference

Peng, J. and Paul, D.
A Geometric Approach to Maximum Likelihood Estimation of the Functional Principal Components From Sparse Longitudinal Data
Journal of Computational and Graphical Statistics, 18(4):995–1015, 2009.

Yao, F. *et al.*
Functional data analysis for sparse longitudinal data
Journal of the American Statistical Association, 100(470):577–590, 2005.