

Functional Logistic Regression with Sparse FPCA Method

Hyunsung Kim

August 21, 2019

Department of Statistics
Chung-Ang University

1. Introduction
2. Methodology
3. Simulation

Introduction

Temporal gene expression data

- The data was measured at same time points $(0, 7, \dots, 119)$.
- 6178 genes in

Methodology

The sparse FPCA method (James et al., 2001)

- The reduced rank model is used to solve the functional PC problem.
- To fit this model, EM algorithm was used.
- It can be applied the curves measured at irregular or sparse time points.

Functional logistic regression

$$Y_i = g^{-1}\left(\alpha + \sum_{k=1}^K \beta_k \xi_{ik}\right) + e_i, \quad i = 1, \dots, N$$

where $Y_i = 1$, if the curve $\in G_1$ and $Y_i = 0$, if the curve $\in G_2$,

$g(x) = \log\left(\frac{x}{1-x}\right)$ and ξ_{ik} is k th FPC score for the i th individual

- It is similar to standard logistic regression.
- The goal of functional logistic regression is estimating unknown parameters, $(\alpha, \beta_1, \dots, \beta_K)$.

Simulation

The simulated datasets

- The 100 datasets are simulated from the first 5 estimated FPCs from the temporal gene expression data.
- Each dataset has 200 curves with 2 groups. (G_1, G_2)
- Each dataset is randomly divided to 100 training and 100 test sets randomly.
- We apply the functional logistic regression to the 100 training sets, and predict to the 100 test sets.

Simulation

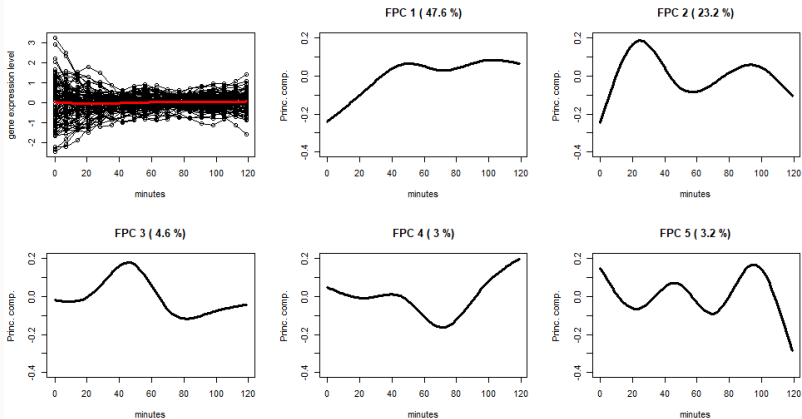


Figure 1: The mean curve and 5 FPC functions for 1st training set

Simulation

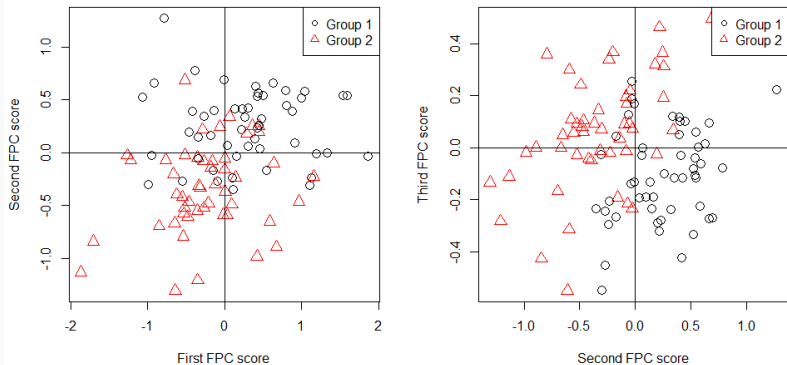


Figure 2: Scatterplot of pairwise FPC scores for 1st training dataset

Simulation

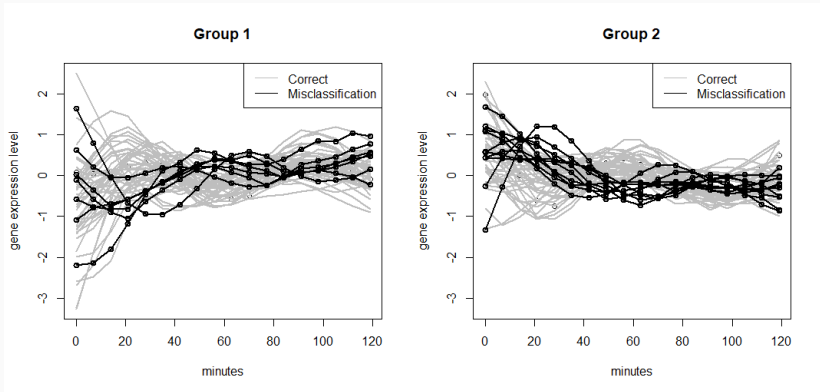


Figure 3: The curves classified by functional logistic regression for 1st simulated dataset

Table 1: Classification error rates between Dense and Sparse method

No. of FPCs	Group 1		Group 2		Overall	
	Dense	Sparse	Dense	Sparse	Dense	Sparse
1	32.72 (8.41)	31.67 (0.08)	32.70 (8.31)	33.65 (0.08)	32.71 (5.26)	32.68 (0.06)
2	22.16 (6.65)	22.20 (0.08)	22.06 (6.15)	22.00 (0.07)	22.11 (4.33)	22.10 (0.05)
3	7.58 (4.58)	12.45 (0.11)	8.26 (5.34)	12.47 (0.09)	7.92 (3.35)	12.46 (0.09)
4	7.14 (4.14)	11.81 (0.09)	7.62 (5.10)	11.37 (0.08)	7.38 (3.11)	11.59 (0.08)
5	7.40 (4.07)	12.22 (0.11)	7.86 (5.26)	11.45 (0.08)	7.63 (3.06)	11.83 (0.09)

Comparison between Dense and Sparse FPCA method

- The sparse method shows higher misclassification rate than the dense one.
- The Monte Carlo standard errors are much lower on the sparse method.
-

Reference

Reference



James G.M., Hastie T.J., Sugar C.A.

Principal component models for sparse functional data

Biometrika, 87(3):587–602, 2000.



Zhou L. *et al.*

Joint modeling of paired sparse functional data using principal components

Biometrika, 95(3):601–619, 2008.



Leng. X. and Müller. HG.

Classification using functional data analysis for temporal gene expression data

Bioinformatics, 22(1):68–76, 2006.