Functional Logistic Regression with Sparse Functional PCA Method

Hyunsung Kim

August 21, 2019

Department of Statistics
Chung-Ang University

Outline

1. Introduction
2. Methods
3. Simulation

Introduction

Introduction

Temporal gene expression data

- The data was measured at 18 equal time points$(0, 7, ..., 119)$.
- From this dense data, We generate 100 datasets based on the functional PCs.

Classification for the simulated data

- We compute FPC scores using sparse functional PCA method.
- Using the above FPC scores, we perform classification using the functional logistic regression.

Methods

Methods

The Sparse Functional PCA

- It can be applied the curves measured at irregular or sparse time points.
- James *et al.* (2001) used the reduced rank model to solve the functional PC problem.
- To fit above model, `EM` algorithm was used.

Functional Logistic Regression

$$Y_i = \pi_i + \epsilon_i, \quad i = 1, \ldots, n$$

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

$$\begin{aligned}\pi_i &= P(Y = 1|X = \boldsymbol{x}_i) \\ &= \frac{\exp[\alpha + \int_T \beta(t)x_i(t)dt]}{1 + \exp[\alpha + \int_T \beta(t)x_i(t)dt]}\end{aligned}$$

with $X : T \to \mathbb{R}$ is the predictor, α is an intercept parameter, $\boldsymbol{\beta} : T \to \mathbb{R}$ is a coefficient function, and ϵ_i is the independent errors with zero mean.

Functional Logistic Regression with functional PC approach

$$Y_i = \pi_i + \epsilon_i,\ i = 1, \ldots, n$$

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

$$\pi_i = \frac{\exp[\alpha + \sum_{k=1}^{K} \beta_k \xi_{ik}]}{1 + \exp[\alpha + \sum_{k=1}^{K} \beta_k \xi_{ik}]},\ i = 1, \ldots, n$$

with α is an intercept parameter and $\boldsymbol{\beta}$ is a coefficient function, ξ_{ik} is kth FPC score for the ith individual, and ϵ_i is the independent errors with zero mean.

Simulation

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) and is randomly divided to 100 training and test sets for each.
- We perform the functional logistic regression for the training sets, and predict for the test sets.

Simulation

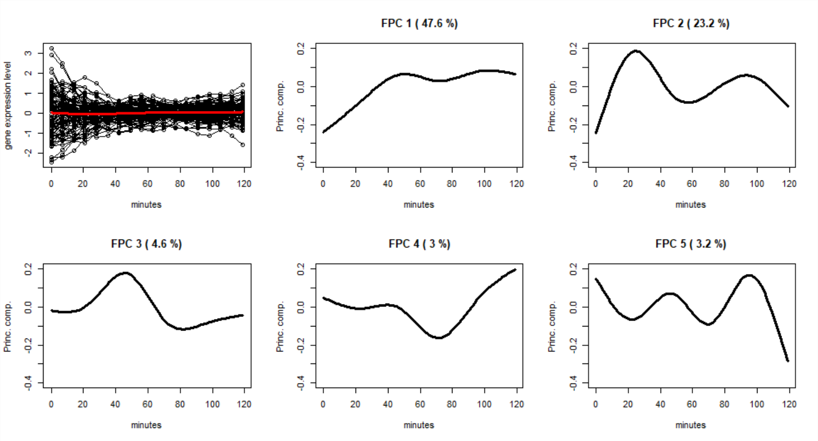

Figure 1: The mean curve and 5 FPC functions for 1st training set

Simulation

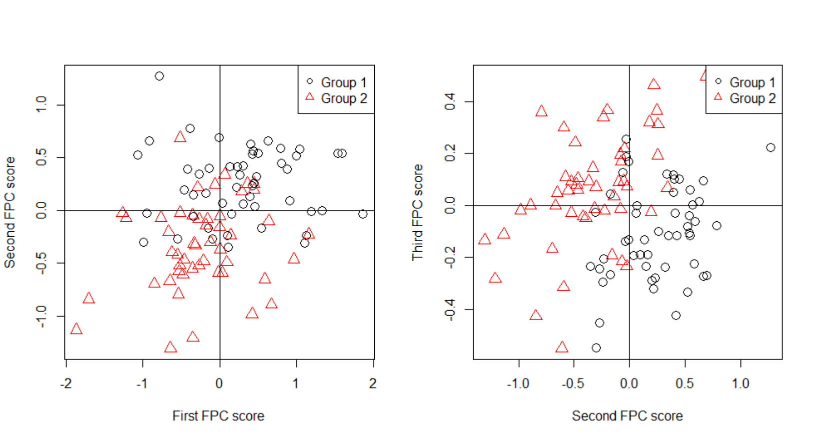

Figure 2: Scatterplot of pairwise FPC scores for 1st training dataset

Simulation

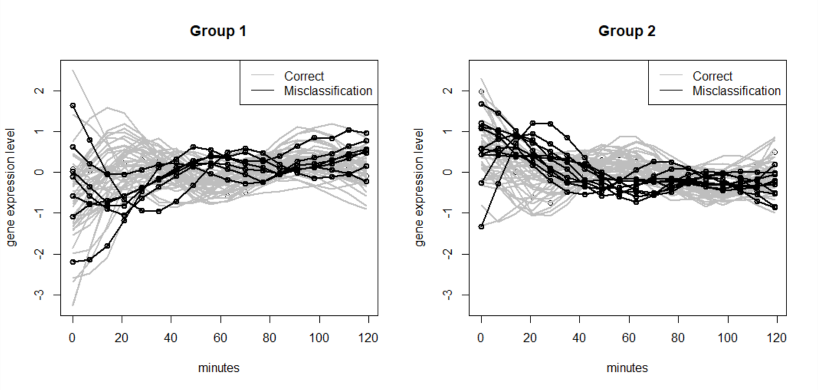

Figure 3: The curves classified by functional logistic regression for 1st simulated dataset

Simulation

Table 1: Classification error rates between Dense and Sparse method

No. of	Group 1		Group 2		Overall	
FPCs	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.
- For the data measured at all time points, the dense functional PCA method perform well than 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.