Functional Classification for Sparse Data

Hyunsung Kim

September 9, 2019

Department of Statistics
Chung-Ang University

Outline

1. `fpca` package

2. Simulation

`fpca` **package**

`fpca` package

- The package to obtain functional PC for sparsely and irregularly observed data.
- Using the `EM` option, it solves the reduced rank model(James *et al.*) to obtain FPC functions.
- It uses PACE method(Yao *et al.*) not the numerical integration to estimate FPC scores.

Comparison for Karhunen-Loève Expansion

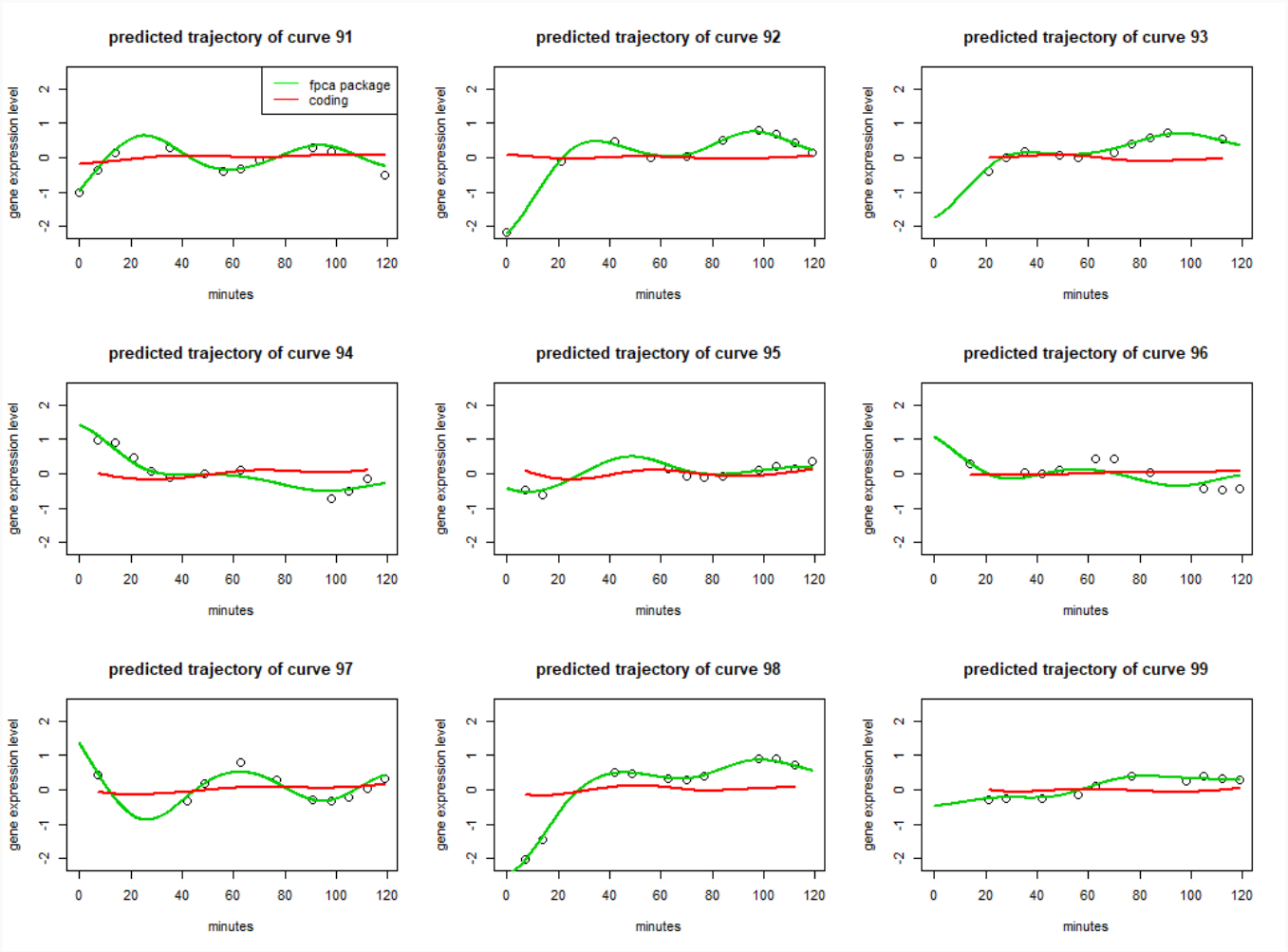

Figure 1: Predicted curves using 3 FPCs for 1st training set

Simulation

Simulation

The Classification Methods

- Functional logistic regression
- SVM using FPCs with 4 different kernels
  - Linear
  - Gaussian(Radial basis function)
  - Sigmoid
  - Polynomial

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.

Simulation Results

Table 1: Accuracy using 2 FPCs

No. of obs	Logistic	SVM(Linear)	SVM(Gaussian)	SVM(Sigmoid)	SVM(Polynomial)
2	0.560	0.553	0.551	0.554	0.530
3	0.685	0.684	0.652	0.669	0.643
4	0.743	0.743	0.729	0.719	0.709
5	0.777	0.775	0.757	0.762	0.735
6	0.788	0.787	0.770	0.771	0.747
7	0.793	0.791	0.779	0.779	0.754
8	0.789	0.786	0.771	0.767	0.755
9	0.784	0.783	0.767	0.767	0.745
10	0.784	0.782	0.769	0.772	0.751
11	0.775	0.775	0.761	0.759	0.744
12	0.789	0.787	0.770	0.775	0.759
13	0.781	0.783	0.768	0.762	0.752
14	0.784	0.784	0.769	0.770	0.752
15	0.779	0.778	0.768	0.764	0.747
16	0.783	0.783	0.768	0.771	0.748
17	0.779	0.780	0.764	0.757	0.746
18	0.781	0.778	0.764	0.766	0.748
Average	0.762	0.761	0.746	0.746	0.727

Simulation Results

Table 2: Accuracy using 3 FPCs

No. of obs	Logistic	SVM(Linear)	SVM(Gaussian)	SVM(Sigmoid)	SVM(Polynomial)
2	0.549	0.547	0.551	0.533	0.556
3	0.750	0.742	0.712	0.731	0.707
4	0.801	0.804	0.790	0.789	0.783
5	0.852	0.849	0.831	0.844	0.825
6	0.861	0.862	0.845	0.854	0.842
7	0.885	0.885	0.865	0.876	0.860
8	0.879	0.881	0.867	0.874	0.862
9	0.887	0.888	0.872	0.876	0.860
10	0.894	0.894	0.878	0.885	0.876
11	0.881	0.886	0.868	0.877	0.859
12	0.892	0.893	0.874	0.883	0.872
13	0.893	0.891	0.879	0.885	0.869
14	0.894	0.897	0.881	0.888	0.875
15	0.893	0.896	0.882	0.887	0.874
16	0.896	0.899	0.884	0.885	0.877
17	0.896	0.898	0.885	0.889	0.878
18	0.897	0.898	0.885	0.889	0.877
Average	0.853	0.854	0.838	0.844	0.832

Simulation Results

Table 3: Accuracy using 4 FPCs

No. of obs	Logistic	SVM(Linear)	SVM(Gaussian)	SVM(Sigmoid)	SVM(Polynomial)
2	0.554	0.568	0.552	0.566	0.598
3	0.699	0.711	0.678	0.690	0.656
4	0.812	0.811	0.771	0.796	0.769
5	0.863	0.860	0.834	0.847	0.821
6	0.876	0.874	0.856	0.863	0.845
7	0.883	0.885	0.859	0.874	0.855
8	0.886	0.890	0.861	0.881	0.860
9	0.897	0.900	0.876	0.886	0.875
10	0.903	0.902	0.875	0.893	0.871
11	0.896	0.897	0.879	0.889	0.872
12	0.898	0.900	0.877	0.887	0.874
13	0.895	0.899	0.879	0.888	0.872
14	0.902	0.905	0.880	0.890	0.875
15	0.900	0.902	0.881	0.891	0.875
16	0.901	0.905	0.885	0.894	0.879
17	0.900	0.904	0.885	0.897	0.878
18	0.902	0.905	0.883	0.892	0.877
Average	0.857	0.860	0.836	0.849	0.833

Simulation Results

Table 4: Accuracy using 5 FPCs

No. of obs	Logistic	SVM(Linear)	SVM(Gaussian)	SVM(Sigmoid)	SVM(Polynomial)
2	0.518	0.537	0.523	0.537	0.502
3	0.623	0.632	0.622	0.603	0.596
4	0.678	0.680	0.641	0.661	0.645
5	0.854	0.855	0.815	0.843	0.823
6	0.865	0.866	0.834	0.851	0.835
7	0.894	0.896	0.868	0.882	0.866
8	0.889	0.889	0.866	0.882	0.865
9	0.898	0.895	0.867	0.887	0.873
10	0.897	0.896	0.869	0.890	0.872
11	0.900	0.901	0.872	0.891	0.877
12	0.900	0.901	0.876	0.889	0.874
13	0.896	0.899	0.876	0.891	0.873
14	0.898	0.902	0.877	0.890	0.875
15	0.899	0.903	0.874	0.890	0.876
16	0.900	0.904	0.877	0.894	0.879
17	0.898	0.902	0.878	0.892	0.877
18	0.900	0.905	0.875	0.889	0.879
Average	0.842	0.845	0.818	0.833	0.817

Simulation Results

Comparison between 5 Classification methods

- The logistic regression and linear SVM perform well than other kernel SVM methods.
- When the number of FPCs is greater than or equal to 3, the accuracy is very similar.
- If there are about 7 out of 18 observations, the accuracy is almost the same.

Reference

Reference

James G.M. *et al.*
Principal component models for sparse functional data
Biometrika, 87(3):587–602, 2000.

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

Rossi F. and Villa N. *et al.*
Support vector machine for functional data classification
Neurocomputing, 69:730–742, 2006.

Leng. X. and Müller. HG.
Classification using functional data analysis for temporal gene expression data
Bioinformatics, 22(1):68–76, 2006.