Classification Using Functional Data Analysis for Temporal Gene Expression Data

Xiaoyan Leng, Hans-Georg Müller

2006

Contents

1. Introduction
2. Models and Methods
3. Results
4. Discussion and Conclusions

1. Introduction

- Recently microarray experiments have been widely used to collect large-scale temporal data to monitor gene expression

- Temporal gene expression data can be useful for genetic analysis such as cell type determination

- Classifying genes into different functional groups is a first step to understand different biological functions

- Many classification analyses have been performed with hierarchical clustering, k-means clustering, principal component analysis(PCA), and singular value decomposition(SVD)

1. Introduction

- In these methods, data are treated as vectors of discrete samples and permutation of components will not affect the analysis results
- Hence the timing of the biological processes is irrelevant in these analyses
- A more efficient way to look at such data is to incorporate the information that is inherent in time order and smoothness of processes over time
- The tools for such an approach are provided by the recently developed methodology of functional data analysis(FDA)

2. Models and Methods : Functional PCA

- model the sample curve *X(t)* on [*0, T*], with mean $E[X(t)] = \mu(t)$ and covariance function cov[$X(s)$, $X(t)$]=$G(s, t)$

Covariance function *G* by *Mercer*'s theorem

$$G(s,t) = \sum_m \lambda_m \rho_m(s)\rho_m(t) \quad m = 1, 2, \ldots$$

$$\rho_m : \text{eigenfunctions}, \;\; \lambda_m : \text{eigenvalues and } \lambda_1 \geq \lambda_2 \geq \ldots$$

random curve : Karhunen-Loe've representation

$$X(t) = \mu(t) + \sum_m \epsilon_m \rho_m(t) \quad 0 \leq t \leq T \text{ where, FPC scores}$$

$$\epsilon_m = \int_0^T (X(t) - \mu(t))\rho_m(t)dt \quad \text{are uncorrelated}$$

$$\text{random variables with } E(\epsilon_m) = 0, \;\; E(\epsilon_m^2) = \lambda_m, \;\; \text{and } \sum \lambda_m < \infty$$

2. Models and Methods : Functional PCA

- An estimate of the $\mu(t)$, $\hat{\mu}(t)$ can be obtained by any linear scatterplot smoother
- Forming a dense grid s_k of $[0, T](k = 1, 2, ..., S,\ s_S = T)$,

$$\hat{G} = [C_n(s_k, s_l)]_{S \times S},\ \text{where}$$

$$C_n(s_k, s_l) = \frac{1}{n} \sum_{i=1}^{n} [(X_i(s_k) - \hat{\mu}(s_k))(X_i(s_l) - \hat{\mu}(s_l))].$$

- This $\hat{G}$ yields m-th eigenvector $(\hat{\rho}_m(s_1), ..., \hat{\rho}_m(s_S))'$ with the corresponding eigenvalue $\hat{\lambda}_m$, for $m = 1, ..., M$

2. Models and Methods : Functional PCA

- Then, FPC scores for the i-th gene are obtained numerically by

$$\hat{\epsilon}_{im} = \sum_{k=1}^{S} ((X_i(s_k) - \hat{\mu}(s_k))\hat{\rho}_m(s_k).$$

- Individual temporal gene expression can be predicted using their FPC scores, by

$$\hat{X}_i(t) = \hat{\mu}(t) + \sum_{m=1}^{M} \hat{\epsilon}_{im}\hat{\rho}_m(t) \quad 0 \leq t \leq T.$$

- $\hat{\epsilon}_{im}$ can then be used to describe both between-group variability and between-group mean differences that may be relevant to classification

2. Models and Methods : Functional Logistic Regression

- For an i.i.d. sample $X_i(t)$(assumed to have mean zero), for $i = 1, ..., n$ the linear predictors are defined by, $\eta_i = \alpha + \int \beta(t) X_i(t) dt$
- Then functional generalized linear model is

$$Y_i = g^{-1}(\eta_i) + e_i, \quad i = 1, ..., n \text{ and } E(e_i) = 0, \ \mathrm{var}(e_i) < C < \infty$$

and by the M-truncated model,

$$Y_i = g^{-1}\left(\alpha + \sum_{m=1}^{M} \beta_m \epsilon_m\right) + e_i, \quad i = 1, 2, ..., n.$$

2. Models and Methods : Functional Logistic Regression

- For fixed M, the estimator of the parameter vector $\hat{\beta}' = (\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_M)$ can be obtained by solving the score equation,

$$U(\beta) = \sum_{i=1}^{n} (Y_i - \mu_i) g'(\eta_i) \epsilon_i / \sigma^2(\mu_i).$$

- For functional binomial regression, set $\pi_i = P(Y_i = 1)$ and prior probabilities p_1 and p_0 for the groups G_1 and G_0

- Then estimate $\hat{\pi}_i = \hat{P}(Y_i = 1 \mid X_i(t)) = g^{-1}(\hat{\alpha} + \sum_{m=1}^{M} \hat{\beta}_m \hat{\epsilon}_m)$, and classify the i-th observation into G_1 if $\hat{\pi}_i \geq p_1$, otherwise into G_0

2. Models and Methods : B-spline Based Method

- Rice and Wu(2001) and Shi *et al.*(1996) proposed a mixed effects model for unequally sampled noisy curves
- Let $X_i = (X_i(t_{i1}), X_i(t_{i2}), ..., X_i(t_{in_i}))'$ be the vector of observations for the i-th curve for $i = 1, 2, ..., n$
- The approximating model of Rice and Wu is,

$$X_{ij} = X_i(t_{ij}) = \sum_{k=1}^{p} \beta_k \bar{B}_k(t_{ij}) + \sum_{l=1}^{q} \gamma_{il} B_l(t_{ij}) + \epsilon_{ij},$$

where $E(X_i(t)) = \mu(t) = \sum_{k=1}^{p} \beta_k \bar{B}_k(t)$ and $\bar{B}_k(\cdot), B_l(\cdot)$ are possibly different B-spline bases on $[0, T]$, and γ_{il} are random effects

- The estimates $\hat{\beta}_k, \hat{\gamma}_{il}$ are obtained by least squares, and classification can be based on the random coefficients γ_i

3. Results : Temporal gene expression data for cell cycle

- There are 6178 genes in total, and each gene expression consists of 18 data points, measured every 7 min between 0 and 119 min
- Of 90 genes, 44 are known to be related to G_1 phase regulation and 46 to non-G_1 phase regulation of the yeast cell cycle → training set
- The expression profiles for these 90 genes are depicted in Fig.1

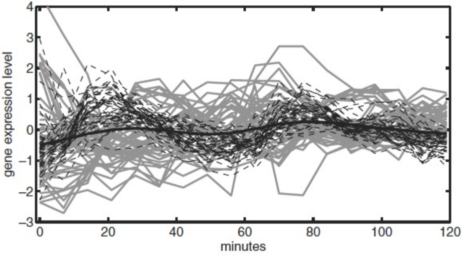

Fig. 1. Temporal gene expression profiles of yeast cell cycle. Dashed lines: G_1 phase; Gray solid lines: non-G_1 phases; Black solid line: overall mean curve.

3. Results : Temporal gene expression data for cell cycle

- Differentiated gene expression profiles into phase-specific groups

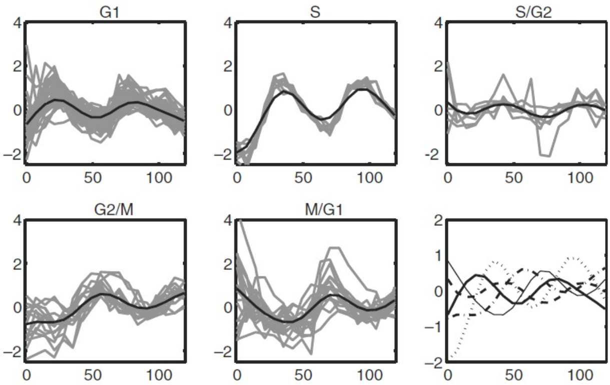

Fig. 2. Yeast cell-cycle gene expression profiles sorted by phases. The first five panels provide the expression profiles for G$_1$, S, S/G$_2$, G$_2$/M, and M/G$_1$ pha with mean functions indicated by the black solid lines. The lower right panel contains the mean curves of all phases overlaid: G$_1$: thick solid line; S: dotted l S/G$_2$: dashed line; G$_2$/M: dash–dot line; and M/G$_1$: thin solid line.

3. Results : Temporal gene expression data for cell cycle

- The number M of FPCs is chosen by minimizing the leave-one-gene-out cross-validation classification error rate
- Fig.4 shows the overall misclassification error rates
- The first five FPCs for the gene expression curves in the training set are depicted in Fig.5

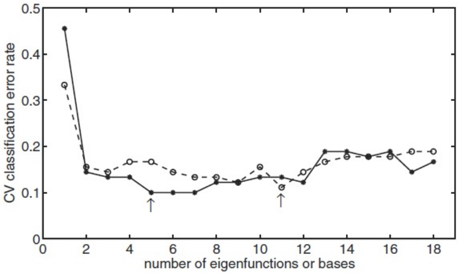

Fig. 4. Choosing M, the number of eigenfuntion for yeast cell-cycle data. Leave-one-out cross-validation estimates of the misclassification rate, in dependency on M: FPCA: solid line; *B*-spline: dashed line.

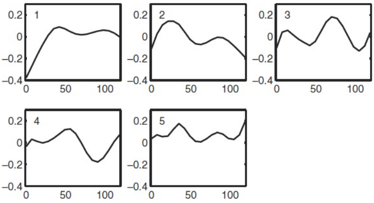

Fig. 5. The first five FPCs for the known 90 genes for yeast cell-cycle data. These five FPCs account for 98.9% of the total variation, with the first FPC accounting for 66.5%, the second for 21.9%, the third for 4.7%, the fourth for 3.1% and the fifth for 2.7%.

3. Results : Temporal gene expression data for cell cycle

- For FPCA, when the first five FPCs are used, the overall cross-validation classification error rate is at a minimum 10.00% (for G1 genes 11.36%, non-G1 genes 8.70%)

- On the other hand, using B-splines, the misclassification error rate attains its lowest value 11.1% for 11 bases, whereas FPCA used only 5 FPCs

- Thus FPCA is seen to be advantageous in this application

3. Results : Temporal gene expression data for cell cycle

- Plotting the FPC scores for the each two FPCs reveals interesting patterns for genes of different phases
- The order of the phases appears to be most evident in the scatterplot of the third vs second FPC scores

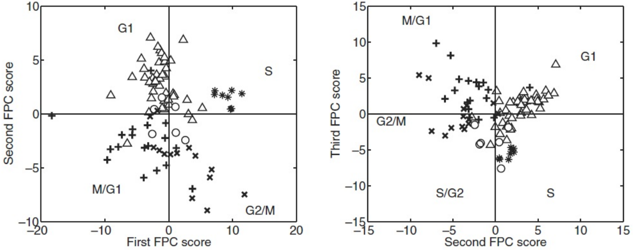

Fig. 6. Scatterplot of pairwise FPC scores for genes in the five cell-cycle phases. G_1: triangles; S: stars; S/G_2: circles; G_2/M: x-marks; M/G_1: plus signs. Left panel: Second versus first FPC score; Right panel: Third versus second FPC score, for yeast cell-cycle data.

3. Results : Temporal gene expression data for cell cycle

- It is found that five G_1 genes were misclassified into the non-G_1 group
- Left panel: the trajectories of four misclassified G_1 genes are close to those of the S genes(thus misclassfied into S)
- Right panel: One misclassified G_1 gene is seen to be close to the M/G_1 trajectories

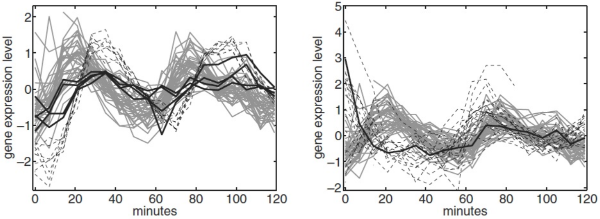

Fig. 7. Profiles of misclassified yeast cell cycle genes. Left panel: black solid lines: misclassified G_1 genes; gray sold lines: G_1 genes; dashed lines: S genes; note that the misclassified G1 genes (YDL055C, YDR113C, YDR356W and YJL092W) are actually close to the trajectories of the S genes. Right panel: black solid line: misclassified G_1 gene; gray solid lines: G_1 genes; dashed lines: M/G_1 genes; note that the misclassified G_1 gene (YCL055W) is actually close to the trajectories of the M/G_1 genes. The genes YJL092W and YCL0055W were also pointed out by Spellman *et al.* (1998).

3. Results : Expression patterns of cell-type specific genes

- Iranfar *et al.*(2001) studied expression patterns of cell-type specific gene fragments in *Dictyostelium discoideum*

- Fitting a biologically based kinetic equation, the authors recognized 35 cell-type specific genes

- Setting these 35 genes(14 prestalk genes and 21 prespore genens) as training set, explore other potential cell-type specific genes by using functional discriminant analysis

3. Results : Expression patterns of cell-type specific genes

- A considerable number of prestalk genes peaked between 8 and 10 h of development and then decreased significantly
- most prespore genes were not expressed until 10 h of development and continued to be expressed thereafter

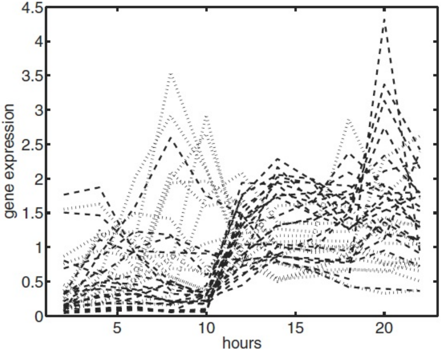

Fig. 8. Developmental temporal patterns of *Dictyostelium* gene expression. Dotted lines: prestalk-specific genes; dashed lines: prespore-specific genes.

3. Results : Expression patterns of cell-type specific genes

- Cross-validation error rates indicated that using the first three FPCs yields the lowest overall misclassification rate of 22.86% (28.57% for prestalk genes, 19.05% for prespore genes)

- Misclassified prestalk and prespore genes are highlighted in the Fig.9

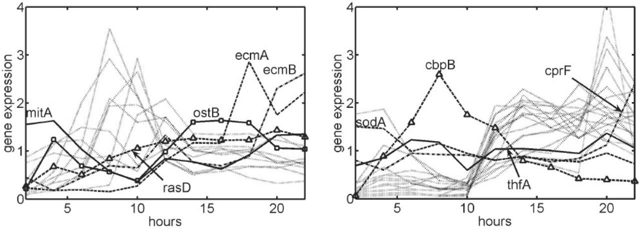

Fig. 9. Misclassified cell-type specific genes for *Dictyostelium*. Left panel: Prestalk genes are shown in light gray. The four misclassified prestalk-specific genes are labelled and highlighted in thick lines; gene *ras*D is also indicated (see text). Right panel: Prespore genes are shown in light gray. The four misclassified prespore-specific genes are lablelled and highlighted in thick lines.

3. Results : Expression patterns of cell-type specific genes

- Since two genes ostB and mitA show no cell-type specific features, they might not be correctly classified as prestalk genes
- For another gene rasD, the estimated probability for classification into prespore genes with 0.4376 only slightly exceeds the prior probability, 0.4
- Gene cbpB shows an early peak at 8 h, and follows the pattern of prestalk genes
- Genes sodA and thfA did not show obvious cell-specific features in their expression patterns
- Gene cprF did not start to express until 20 h, which may have contributed to its misclassification

3. Results : Expression patterns of cell-type specific genes

- Now use the fitted model to classify the rest of the genes
- Choose ranges of estimated probabilities for a gene to be classified into the prestalk and prespore group in order to identify subgroups
- Each identified subgroup displays reasonably homogeneous patterns

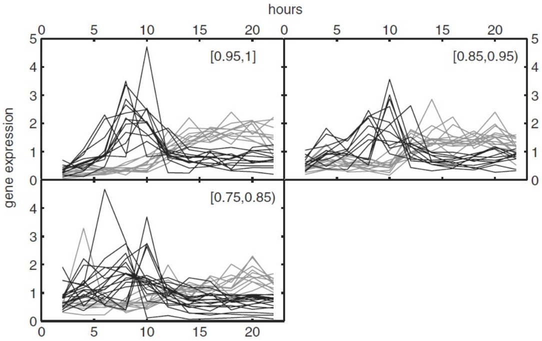

Fig. 10. Subgroups of *Dictyostelium* cell-type specific genes according to different posterior probabilities. Black solid lines: prestalk-specific genes corresponding to the indicated range of posterior probabilities; gray solid lines: prespore-specific genes corresponding to posterior probabilities of [0.0,05] (upper left panel), (0.05, 0.15] (upper right panel) and (0.15, 0.25] (lower panel).

3. Results : Simulation study

- A simulation study was performed based on the first five estimated FPCs from the yeast cell-cycle data

- Five random coefficients $\epsilon_m (m = 1, ..., 5)$ were generated for each subject from normal distributions with means 0.6, 0.5, 0.4, 0.3, and 0.2 for group 1 and -0.6, -0.5, -0.4, -0.3, and -0.2 for group 2
- The variances of these random coefficients correspond to the estimated eigenvalues from the yeast cell-cycle data

- The priors for the two groups were chosen equal *i.e.* $\pi_1 = \pi_2 = \frac{1}{2}$

- 100 training and test datasets are generated, and each dataset was composed of 100 training samples and 100 test samples

3. Results : Simulation study

- The simulation classification error rates based on FPCA and B-splines are compared in Table 1(Monte Carlo standard errors are in parentheses)
- The overall classification error rates based on FPCA are always slightly lower than those observed for B-splines

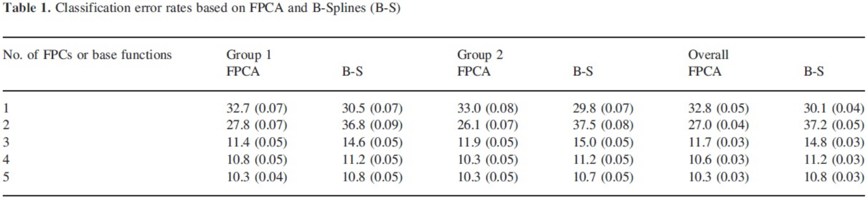

Table 1. Classification error rates based on FPCA and B-Splines (B-S)

No. of FPCs or base functions	Group 1 FPCA	B-S	Group 2 FPCA	B-S	Overall FPCA	B-S
1	32.7 (0.07)	30.5 (0.07)	33.0 (0.08)	29.8 (0.07)	32.8 (0.05)	30.1 (0.04)
2	27.8 (0.07)	36.8 (0.09)	26.1 (0.07)	37.5 (0.08)	27.0 (0.04)	37.2 (0.05)
3	11.4 (0.05)	14.6 (0.05)	11.9 (0.05)	15.0 (0.05)	11.7 (0.03)	14.8 (0.03)
4	10.8 (0.05)	11.2 (0.05)	10.3 (0.05)	11.2 (0.05)	10.6 (0.03)	11.2 (0.03)
5	10.3 (0.04)	10.8 (0.05)	10.3 (0.05)	10.7 (0.05)	10.3 (0.03)	10.8 (0.03)

4. Discussion and Conclusion

- Temporal gene expression data play a critical role in exploring the regulation of gene expression
- Temporal gene expression data provide valuable functional information about temporal patterns of gene expression and also interactions between genes
- Since most biological processes are in fact continuous, temporal gene expression data can be viewed as discretized samples from smooth random trajectories over time
- This feature leads to a functional data analysis approach

4. Discussion and Conclusion

- In this paper, a functional discriminant analysis method is proposed, using a functional version of logistic regression and functional principal components

- The proposed method provides low-error rate (10%) classification for the yeast cell-cycle gene expression data and also in simulations
- In comparisons with the B-spline approach, the FPCA method demonstrate overall lower-error rates with fewer eigenfunctions/base functions in both data analysis and simulations

- the FPCA methods allow the identification of genes that were probably misclassified by traditional classification methods

- Extending the proposed algorithm to functional cluster analysis is feasible and useful in the common situation where group membership is unknown

References I

Hans-Georg Müller Xiaoyan Leng.
Classification using functional data analysis for temporal gene expression data.
BIOINFORMATICS, 22(1):68–76, 2006.

Stadmüller-U. Müller, H.G.
Generalized functional linear models.
Annals Stat., 33:774–805, 2005.

Colin O. Wu John A. Rice.
Nonparametric mixed effects models for unequally sampled noisy curves.
BIORMETRICS, 57:253–259, 2001.

Thank You!