

# FUNCTIONAL DISCRIMINANT ANALYSIS FOR MICROARRAY GENE EXPRESSION DATA VIA RADIAL BASIS FUNCTION NETWORKS

Yuko Araki, Sadanori Konishi and Seiya Imoto

*Key words:* Classification, Functional data analysis, Maximum penalised likelihood, Microarray, Radial basis function.

*COMPSTAT 2004 section:* Functional data analysis.

**Abstract:** We introduce functional logistic discriminant analysis (FLDA) which is an extension of the classical method of logistic discriminant analysis to data where predictor variables are functions or curves. FLDA approach can effectively classify functions into two distinct classes by imposing smoothness constraint on the predictor functions and coefficient function by radial basis function expansion and regularization. In order to select the value of a smoothing parameter, we derive an information criterion which enables us to evaluate model estimated by regularization. The proposed method is illustrated through the analysis of yeast cell cycle microarray data. It is shown that FLDA performs well especially in terms of prediction ability.

## 1 Introduction

Classification or discrimination technique is one of the most widely used statistical tools in various fields of natural and social sciences. In recent years, several techniques have been proposed for analyzing multivariate observations with complex structure (see, for example, Hastie et al. (2001)).

The focus in the present paper will be on the problem of classifying functions, where each observation can be interpreted as a discretized realization of a function evaluated at possibly differing time points. Recently, Cardot et al. (2003) used functional approaches for estimating land use based on the temporal evolution of remote sensing data.

Our motivation arises from the analysis of yeast cell cycle gene expression data which provide inference about how gene expression levels evolve in time and how genes are dependent during a given biological process (Spellman et al. (1998) and Luan and Li (2003)). Classification of genes enables us to predict functions of unknown genes and to identify the set of co-regulated genes. In the yeast cell cycle data analysis, one wish to classify genes based on the cDNA microarray time series data.

We introduce functional discriminant analysis using Gaussian radial basis function networks with help of regularization. It is designed to construct a decision rule based on data given as a set of functions. We first transfer the

vector valued observations to a set of functions. Secondly, functional logistic model is constructed by using Gaussian radial basis functions and then estimation is by regularized maximum likelihood method. In order to select smoothing parameters, we derive model selection criterion within the framework of functional data analysis by developing the generalized information criterion due to Konishi and Kitagawa (1996).

The paper is organized as follows. In Section 2 we describe the radial basis expansion smoothing technique which converts discrete raw data into underlying smooth functional form. In Section 3 the new method, functional logistic discriminant analysis, is set out and the details of its implementation are described. Section 4 presents an application of the proposed method to yeast cell cycle gene expression data collected by Spellman et al. (1998).

## 2 Radial basis smoothing techniques

In the context of functional data analysis (Ramsay and Silverman (1997)), individual data should be considered to have a functional form in nature even though observed data are usually recorded discretely. In addition, those discrete raw data which are supposed to have functional form may contain observational error. Therefore, converting raw data into underlying smooth functional form requires efficient smoothing techniques.

The typical functional data analysis approach is to fit each curve individually using expansion in basis functions. Common basis functions for smoothing functional data are  $B$ -spline basis and Fourier expansions. In our model, we use Gaussian radial basis function with hyperparameter (Ando et al. (2002)). An advantage of this basis expansion is that it controls the amount of overlapping among basis functions and adopts the information of the desired outputs. For background about radial basis function networks, we refer to Moody and Darken (1989), Poggio and Girosi (1990).

Suppose we have  $N$  independent observations  $\{(x_i, t_i); t_i \in \mathcal{T}, i = 1, 2, \dots, N\}$ , where  $x_i$  are random response variables and  $t_i$  are explanatory variables, assuming that they are drawn from the Gaussian nonlinear regression model

$$x_i = u(t_i) + \epsilon_i, \quad i = 1, \dots, N, \quad (1)$$

where  $u(t)$  is a smooth function to be estimated, and the errors  $\epsilon_i$  are independently, normally distributed with mean zero and variance  $\sigma^2$ . We consider the function  $u(t)$  that can be expanded in the form of the radial basis function network taking the following form;

$$u(t; \boldsymbol{\omega}) = \sum_{k=1}^m \omega_k \phi_k(t) + \omega_0, \quad (2)$$

where  $\boldsymbol{\omega} = (\omega_0, \omega_1, \dots, \omega_m)^T$  and  $\phi_k(t)$  are a set of Gaussian radial basis

functions with hyperparameter  $\nu$  given as

$$\phi_k(t; \mu_k, \sigma_k^2) = \exp \left\{ -\frac{(t - \mu_k)^2}{2\nu\sigma_k^2} \right\}, \quad k = 1, \dots, m, \quad (3)$$

where  $\mu_k$  is a scalar determining the location of the  $k$ th basis function,  $\sigma_k$  is the width,  $\nu$  is a hyperparameter. The function  $\hat{u}(t) \equiv x(t)$  which is estimated from the observed data  $\{(x_i, t_i); i = 1, \dots, N\}$  is called ‘functional data’, and is proceeded to further analysis.

The nonlinear function  $u(t)$  is estimated in two-stage procedure; position the centers and determine the dispersions first, then calculate the weights using an appropriate optimization schemes. This two stage learning is reported to solve the problem of convergence and the identification problem. Among several strategies,  $k$ -means clustering method algorithm is used to determine the centers  $\mu_k$  and the dispersion parameters  $\sigma_k^2$  of the basis functions. More precisely, observation points  $\{t_1, \dots, t_N\}$  are grouped into  $m$  clusters  $\{C_1, \dots, C_m\}$ , where  $m$  is a given number of radial basis functions. Then the centers and the dispersion parameters are determined by

$$c_k = \frac{1}{n_k} \sum_{t_i \in C_k} t_i, \quad s_k^2 = \frac{1}{n_k} \sum_{t_i \in C_k} (t_i - c_k)^2,$$

where  $n_k$  represents the number of data which belong to the cluster  $C_k$ . We define the basis function  $\phi_k(t; c_k, s_k^2)$  using those estimates as  $\phi_k(t)$ . Hence it follows that the nonlinear regression model based on the radial basis function network can be written as

$$f(x_i|t_i; \boldsymbol{\omega}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{\{x_i - \boldsymbol{\omega}^T \boldsymbol{\phi}(t_i)\}^2}{2\sigma^2} \right], \quad (4)$$

where  $\boldsymbol{\phi}(t_i) = (1, \phi_1(t_i), \dots, \phi_m(t_i))^T$ .

In fitting data with complex structure, the maximum likelihood method does not yield satisfactory results, since it often occurs overfitting and yields unstable parameter estimates. Moreover, in smoothing functional data, all individual data should be fitted by using the common basis functions in our model. In other words, the number of basis functions is fixed even though the amount of smoothness imposed on a set of discrete data will be differ from each other. Therefore the unknown weights and the error variances are estimated by regularization method. Regularization allows us to adjust individual differences by a smoothing parameter. In addition, implementing the hyperparameter and adjusting the smoothing parameter capture the structure in the data flexibly.

The regularization method maximizes the penalized log-likelihood function

$$l_\gamma(\boldsymbol{\omega}, \sigma^2) = \sum_{i=1}^N \log f(x_i|t_i; \boldsymbol{\omega}, \sigma^2) - \frac{N\gamma}{2} \boldsymbol{\omega}^T D_2^T D_2 \boldsymbol{\omega}, \quad (5)$$

where  $D_2^T D_2$  is the second order difference matrix and  $\gamma$  is called a smoothing parameter which adjusts the amount of smoothness and also avoids ill-posed problem. The maximum penalized likelihood estimates are

$$\hat{\omega} = (\Phi^T \Phi + N\beta D_2^T D_2)^{-1} \Phi^T \mathbf{x}, \quad \hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^N \{x_i - \hat{\omega}^T \phi(t_i)\}^2, \quad (6)$$

where  $\Phi = (\phi(t_1), \phi(t_2), \dots, \phi(t_N))^T$ ,  $\beta = \gamma\sigma^2$  and  $\mathbf{x} = (x_1, \dots, x_N)^T$ .

The number of basis functions  $m$ , the adjusted parameters  $\nu$  and  $\gamma$  are determined by using an information criterion given by Ando et al. (2002). Thus the observed discrete data  $\{(x_i, t_i); t_i \in \mathcal{T}, i = 1, \dots, N\}$  are smoothed by the method described above and we have a functional data given by  $x(t)$ ;

$$\hat{u}(t) = \sum_{k=1}^m \hat{\omega}_k \phi_k(t) + \hat{\omega}_0 \equiv x(t), \quad t \in \mathcal{T}. \quad (7)$$

Marx and Eilers (1999) used  $B$ -splines expansion with regularization, called  $P$ -splines, and applied the procedure to medical diagnosis and phoneme recognition.

### 3 Functional logistic discrimination

Suppose we have  $n$  independent observations  $\{(x_\alpha(t), g_\alpha); \alpha = 1, \dots, n\}$ , where  $x_\alpha(t)$  are functional predictor variables and  $g_\alpha$  are indicators of the group membership. For example, we consider two-class classification, i.e.  $k = 1$  or  $2$ ,  $g_\alpha = k$  implies that it belongs to class  $G_k$ . A set of functions smoothed by the Gaussian radial basis function smoothing method are given by

$$x_\alpha(t) = \mathbf{w}_\alpha^T \phi(t), \quad \alpha = 1, \dots, n, \quad (8)$$

where  $\mathbf{w}_\alpha$  are estimated parameter vectors and  $\phi(t)$  is a vector of Gaussian basis functions given in equation (3).

A Bayes rule of allocation is to assign  $x_\alpha(t)$  to group  $G_k (k = 1, 2)$  with the maximum posterior probability  $\Pr(g = k | x_\alpha(t))$ . We consider the log-odds of the posterior probability given in the following form;

$$\log \left\{ \frac{\Pr(g = 1 | x_\alpha(t))}{\Pr(g = 2 | x_\alpha(t))} \right\} = \beta_\alpha + \int_{\mathcal{T}} \beta(t) x_\alpha(t) dt. \quad (9)$$

By making use of the same Gaussian radial basis function  $\phi(t)$  as in (8), we expand the functional parameter as  $\beta(t) = \beta_0 + \sum_{i=1}^m \beta_i \phi_i(t) = \boldsymbol{\beta}^T \phi(t) (\in \mathcal{T})$ , where  $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_m)^T$ . We denote the posterior probability  $\Pr(g = 1 | x_\alpha(t)) = \pi(x_\alpha(t))$ , so that  $\Pr(g = 2 | x_\alpha(t)) = 1 - \pi(x_\alpha(t))$ . Then the log-odds model (9) can be expressed as

$$\log \left\{ \frac{\pi(x_\alpha(t))}{1 - \pi(x_\alpha(t))} \right\} = \mathbf{Z}_\alpha^T \boldsymbol{\beta}, \quad (10)$$

where  $Z$  is an  $n \times (m + 2)$  matrix given by

$$Z^T = \begin{bmatrix} 1 & 1 & \dots & 1 \\ \Phi^T \mathbf{w}_1 & \Phi^T \mathbf{w}_2 & \dots & \Phi^T \mathbf{w}_n \end{bmatrix} \quad (11)$$

with  $(m + 1) \times (m + 1)$  matrix  $\Phi$  having  $\phi_{jk} = \int \phi_j(t)\phi_k(t)dt$  as the  $(j, k)$  th element.

We define the binary variable  $y_\alpha$  coded as either 0 or 1 to indicate the group membership of a sample, where  $y_\alpha = 1$  if  $g_\alpha = 1$  and  $y_\alpha = 0$  if  $g_\alpha = 2$ . The log-likelihood function is

$$l(\boldsymbol{\beta}) = \sum_{\alpha=1}^n [y_\alpha \log \pi(x_\alpha(t)) + (1 - y_\alpha) \log\{1 - \pi(x_\alpha(t))\}], \quad (12)$$

where  $\pi(x_\alpha(t)) = \exp(\mathbf{Z}_\alpha^T \boldsymbol{\beta}) / \{1 + \exp(\mathbf{Z}_\alpha^T \boldsymbol{\beta})\}$ . We estimate the parameter vector  $\boldsymbol{\beta}$  by maximizing the penalized log-likelihood function

$$l(\boldsymbol{\beta}) - \frac{n\lambda}{2} \boldsymbol{\beta}^T D_2^T D_2 \boldsymbol{\beta}. \quad (13)$$

This is because regularization method yields estimates with lower variances, even though they are biased. Therefore we obtain the solution  $\hat{\boldsymbol{\beta}}_\lambda$  by the iterative algorithm like Newton-Raphson algorithm.

The crucial issue on regularization method is the choice of the optimal value of smoothing parameter  $\lambda$ . We obtain an information-theoretic criterion within the framework of functional data analysis. An information criterion for evaluating functional logistic discrimination model estimated by regularization is of the form

$$\text{GIC}_F = -2 \log l(\hat{\boldsymbol{\beta}}_\lambda) + 2 \text{tr} Q^{-1} R, \quad (14)$$

where  $Q$  and  $R$  are  $(m + 2) \times (m + 2)$  matrices given by the first and second derivatives of equation (13). We choose the smoothing parameter  $\lambda$  to minimize  $\text{GIC}_F$ .

## 4 Real data example

In this section we show the effectiveness of the proposed method through the analysis of the yeast cell cycle gene expression data collected by Spellman et al. (1998). Gene expressions for all 6,178 genes in the yeast genome were measured by cDNA microarrays over time during about two cell cycles. These data contain 77 microarrays and consist of two short time-courses (two time points) and four medium time-courses (18, 24, 17 and 14 time points). Spellman et al. (1998) identified 800 genes as cell cycle related genes based on the clustering analysis, and also grouped these genes into five classes, G1, S, G2, M, and M/G1, by considering peaks in the expression patterns. Figure 1 shows the expression patterns of the 800 genes in the five classes.

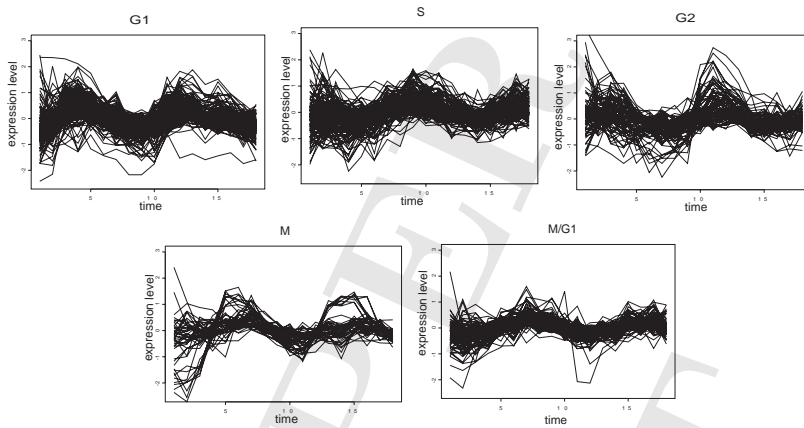


Fig. 1: Raw gene expression patterns during the yeast synchronization experiment.

In our analysis, we concentrate on the time-course “ $\alpha$  factor-based synchronization experiment data” (18 time points), for simplicity excluding the genes containing missing values. That is, the expression patterns of 612 genes out of 800 cell cycle related genes are used in our analysis, and those expression data are considered as a discretized realization of 612 expression curves evaluated at 18 time points. Note that microarray data usually contain observational noise. Therefore, the smoothing that we will first perform has an important role to remove the observational noise from expression data. In addition, since the gene expression pattern of each cell cycle related gene can be considered as a function of time, the proposed method is appropriate for analyzing time course gene expression data.

We carried out two-class logistic discrimination for all possible combinations. In order to evaluate the effectiveness of our FLDA model, the genes in each class were randomly assigned into training data and test data. That is, the FLDA model is estimated by using the training data, and the predictive ability of the estimated model is evaluated by the test data.

We first performed the Gaussian radial basis smoothing method described in Section 2 to the time-course expression data  $\{(t_i, x_{ij}); i = 1, \dots, 18; j = 1, \dots, 612\}$ , where  $t_i$  is the  $i$ th time point and  $x_{ij}$  is the expression value of  $j$ th gene at time  $t_i$ . In the functional discrimination analysis, each smoothing step has to be carried out by using the same number of basis functions. Hence the differences of the degree of smoothness between different gene expression patterns can be adjusted by the smoothing parameters. However, since there are various gene expression patterns in the same group, adjusting their smoothness by using only the smoothing parameter might not be enough. In such a case, the proposed radial basis function smoothing method with the hyperparameter works efficiently in practice.

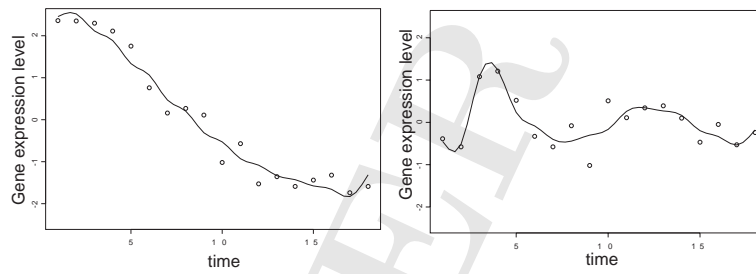


Fig. 2: Smoothed gene expression patterns by the Gaussian radial basis function networks with hyperparameter.

Figure 2 shows examples of two G1-grouped genes with the Gaussian radial basis smoothing curves. Although there are various types of expression patterns in the same class, we succeeded in extracting the effective expression curves that are possibly close to the real expression patterns. We observe that the hyperparameter allows flexible curve fitting, and the smoothing parameter adjusts the differences of gene expression patterns effectively.

The linear discriminant analysis (LDA) and the quadratic discriminant analysis (QDA) are the most popular classical method for discriminant analysis. We compare FLDA evaluated by the criterion  $GIC_F$  with LDA and QDA which analyze discretized data directly. For almost all combinations of the classes, the proposed method yields a lower test error. We suggest investigating genes that were classified in the opposite group with high posterior probability, since they may have been misclassified by Spellman et al. (1998).

## 5 Conclusion

The functional logistic discriminant analysis proposed in this paper appears to be a useful tool for classifying functions or curves. An advantage of our method is that one could treat the samples as a set of functions, hence the problems of the observational point difference and highly correlated data are overcome. Also the model selection criterion enables us to evaluate models subjectively. Potential research would be extending our modeling strategy to the case of sampled surface for multi-group classification.

### Acknowledgement

The authors would like to thank two anonymous reviewers for pointing out related references.

### References

- [1] Ando, T. and Konishi, S. (2002). Nonlinear regression modeling via regularized radial basis function networks. The Institute of Statistical Math-

- ematics, Research Memorandum. **845**.
- [2] Cardot, H., Faivre, R. and Goulard, M. (2003). Functional approaches for predicting land use with the temporal evolution of coarse resolution remote sensing data. *J. of Applied Statist.* **30**, 1185–1199.
  - [3] Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer-Verlag, New York.
  - [4] Konishi, S. and Kitagawa, G. (1996). Generalised information criteria in model selection. *Biometrika* **83**, 875–890.
  - [5] Luan, Y. and Li, H. (2003). Clustering of time-course gene expression data using a mixed-effects model with  $B$ -splines. *Bioinformatics* **19**(4), 474–482.
  - [6] Marx, B.D. and Eilers, P.H.C. (1999). Generalized linear regression on sampled signals and curves : A  $P$ -spline approach. *Technometrics* **41**, 1–13.
  - [7] Moody, J. and Darken, C. J. (1989). Fast learning in networks of locally-tuned processing units. *Neural Comp.* **1**, 281–294.
  - [8] Poggio, T. and Girosi, F. (1990). Networks for approximation and learning. *Proc. IEEE* **78**, 1484–1487.
  - [9] Ramsay, J. O. and Silverman, B. W. (1997). *Functional Data Analysis*. Springer-Verlag, New York.
  - [10] Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Bostein, D. and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell.* **9**, 3273–3297.

*Address:*

Yuko Araki and Sadanori Konishi  
Graduate School of Mathematics, Kyushu University  
6-10-1 Hakozaki, Higashi-Ku, Fukuoka 812-8581, Japan.

Seiya Imoto  
Institute of Medical Science, University of Tokyo  
4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan.

*E-mail:*

yuko@math.kyushu-u.ac.jp (Araki, Y.),  
konishi@math.kyushu-u.ac.jp (Konishi, S.),  
imoto@ims.u-tokyo.ac.jp (Imoto, S.)