SEMI-SUPERVISED LOGISTIC DISCRIMINATION FOR FUNCTIONAL DATA

Kawano, Shuichi
Department of Mathematical Sciences, Graduate School of Engineering, Osaka Prefecture University

Konishi, Sadanori
Department of Mathematics, Faculty of Science and Engineering, Chuo University

https://doi.org/10.5109/1495407

出版情報：Bulletin of informatics and cybernetics. 44, pp.1-15, 2012-12. Research Association of Statistical Sciences
バージョン：

SEMI-SUPERVISED LOGISTIC DISCRIMINATION FOR FUNCTIONAL DATA

by

Shuichi Kawano and **Sadanori Konishi**

Reprinted from the Bulletin of Informatics and Cybernetics
Research Association of Statistical Sciences, Vol.44

FUKUOKA, JAPAN
2012

Bulletin of Informatics and Cybernetics, Vol. 44, 2012

SEMI-SUPERVISED LOGISTIC DISCRIMINATION FOR FUNCTIONAL DATA

By

Shuichi Kawano* and **Sadanori Konishi†**

Abstract

Multi-class classification methods based on both labeled and unlabeled functional data sets are discussed. We present a semi-supervised logistic model for classification in the context of functional data analysis. Unknown parameters in our proposed model are estimated by regularization with the help of EM algorithm. A crucial point in the modeling procedure is the choice of a regularization parameter involved in the semi-supervised functional logistic model. In order to select the adjusted parameter, we introduce model selection criteria from information-theoretic and Bayesian viewpoints. Monte Carlo simulations and a real data analysis are given to examine the effectiveness of our proposed modeling strategy.

Key Words and Phrases: EM algorithm, Functional data analysis, Model selection, Regularization, Semi-supervised learning.

1. Introduction

In recent years, functional data analysis has been used in various fields of study such as chemometrics and meteorology (e.g., we refer to Ramsay and Silverman, 2002; 2005, Ferraty and Vieu, 2006). The basic idea behind functional data analysis is to express a discrete data set as a smooth function data set, and then exploit information obtained from the set of functional data using the functional analogs of classical multivariate statistical tools. Till this day, several researchers have studied a variety of functional versions of traditional supervised and unsupervised statistical methods; e.g., functional regression analysis (James and Silverman, 2005; Yao *et al.*, 2005; Araki *et al.*, 2009a), functional discriminant analysis (Ferraty and Vieu, 2003; Rossi and Villa, 2006; Araki *et al.*, 2009b), functional principal component analysis (Rice and Silverman, 1991; Siverman, 1996; Yao and Lee, 2006) and functional clustering (Abraham *et al.*, 2003; Rossi *et al.*, 2004; Chiou and Li, 2007).

Meanwhile, a semi-supervised learning, which is a modeling procedure based on both labeled and unlabeled data, has received considerable attention in the contemporary statistics, machine learning and computer science (see, e.g., Chapelle *et al.*, 2006; Liang *et al.*, 2007; Zhu, 2008). In particular, it is known that the semi-supervised learning is useful in the application areas including text mining and bioinformatics, in which obtaining labeled data is difficult while unlabeled data can be easily obtained. Many of

* Department of Mathematical Sciences, Graduate School of Engineering, Osaka Prefecture University, 1-1 Gakuen-cho, Sakai, Osaka 599-8531, Japan. skawano@ms.osakafu-u.ac.jp

† Department of Mathematics, Faculty of Science and Engineering, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan. konishi@math.chuo-u.ac.jp

ordinary statistical multivariate analysis have been extended into the semi-supervised resemblances by earlier researchers; e.g., semi-supervised regression analysis (Verbeek and Vlassis, 2006; Lafferty and Wasserman, 2007; Ng *et al.*, 2007), semi-supervised discriminant analysis (Miller and Uyer, 1997; Yu *et al.*, 2004; Zhou *et al.*, 2004; Dean *et al.*, 2006; Kawano and Konishi, 2011) and semi-supervised clustering (Basu *et al.*, 2004; Zhong, 2006; Kulis *et al.*, 2009).

In this paper, our aim is to extend the supervised modeling procedures for functional data into semi-supervised counterparts. We, in particular, focus on a multi-class classification or discriminant problem, and develop a semi-supervised logistic model for functional classification problem. Unknown parameters in the model are estimated by the regularization method along with the technique of EM algorithm. A crucial issue for the modeling procedure is to choose a value of a regularization parameter involved in the semi-supervised functional logistic model. In order to select the optimal value of the regularization parameter, we then introduce model selection criteria based on information-theoretic and Bayesian approaches that evaluate semi-supervised functional logistic models estimated by the regularization method. Some numerical examples including a microarray data analysis are illustrated to investigate the effectiveness of our modeling strategy.

This paper is organized as follows. In Section 2, we consider a functionalization method that converts the discrete data into the functional form using basis expansions. Section 3 proposes a functional logistic model in the context of the semi-supervised multi-class classification problem. In this section, we also present an estimation procedure based on the regularization method with the help of EM algorithm. Section 4 derives model selection criteria to select a regularization parameter in the functional logistic models. In Section 5, Monte Carlo simulations and a real data analysis are given to assess the performances of the proposed semi-supervised functional logistic discrimination. Some concluding remarks are given in Section 6.

2. Functionalization

Suppose that we have n independent observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, where $\boldsymbol{x}_\alpha$ consist of the N_α observed values $x_{\alpha 1}, \ldots, x_{\alpha N_\alpha}$ at discrete times $t_{\alpha 1}, \ldots, t_{\alpha N_\alpha}$, respectively. Our aim in this section is to express a data set $\{(x_{\alpha i},\, t_{\alpha i});\ i = 1, \ldots, N_\alpha,\ \ t_{\alpha i} \in \mathcal{T} \subset \mathbb{R}\}$ $(\alpha = 1, \ldots, n)$ as a set of smooth functions $\{x_\alpha(t); \alpha = 1, \ldots, n,\, t \in \mathcal{T}\}$ by a smoothing technique. In this section we drop the notation on the subject $\boldsymbol{x}_\alpha$, and hence consider a functionalization procedure of the data set $\{(x_i, t_i);\ i = 1, \ldots, N\}$.

It is assumed that the observed values $\{(x_i, t_i);\ i = 1, \ldots, N\}$ for a subject are drawn from a regression model as follows:

$$x_i = u(t_i) + \varepsilon_i, \quad i = 1, \ldots, N, \tag{1}$$

where $u(t)$ is a smooth function to be estimated and the errors ε_i are independently, normally distributed with mean zero and variance σ^2. We also assume that the function $u(t)$ can be represented by a linear combination of pre-prepared basis functions in the form

$$u(t) = \sum_{k=1}^{m} \omega_k \phi_k(t; \mu_k, \eta_k^2), \tag{2}$$

where ω_k are coefficient parameters, m is the number of basis functions and $\phi_k(t;\mu_k,\eta_k^2)$ are Gaussian basis functions given by

$$\phi_k(t;\mu_k,\eta_k^2) = \exp\left\{-\frac{(t-\mu_k)^2}{2\eta_k^2}\right\}, \quad k=1,\ldots,m. \tag{3}$$

Here μ_k are the centers of the basis functions and η_k are the dispersion parameters. In particular, we use Gaussian basis functions proposed by Kawano and Konishi (2007), and hence the centers μ_k and the dispersion parameters η_k are determined as follows: for equally spaced knots τ_k so that $\tau_1 < \cdots < \tau_4 = \min(t) < \cdots < \tau_{m+1} = \max(t) < \cdots < \tau_{m+4}$, we set the centers and the dispersion parameters as $\hat{\mu}_k = \tau_{k+2}$ and $\hat{\eta} \equiv \hat{\eta}_k = (\tau_{k+2}-\tau_k)/3$ for $k=1,\ldots,m$, respectively. For details of the procedure, we refer to Kawano and Konishi (2007).

It follows that the nonlinear regression model based on the Gaussian basis functions can be written as

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

where $\boldsymbol{\omega} = (\omega_1,\ldots,\omega_m)^T$ and $\boldsymbol{\phi}(t) = (\phi_1(t),\ldots,\phi_m(t))^T$. The parameters $\boldsymbol{\omega}$ and σ^2 are estimated by maximizing the regularized log-likelihood function in the form

$$\begin{aligned}\ell_\zeta(\boldsymbol{\omega},\sigma^2) &= \sum_{i=1}^N \log f(x_i|t_i;\boldsymbol{\omega},\sigma^2) - \frac{N\zeta}{2}\boldsymbol{\omega}^T\mathcal{K}\boldsymbol{\omega} \\ &= -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{x}-\Phi\boldsymbol{\omega})^T(\boldsymbol{x}-\Phi\boldsymbol{\omega}) - \frac{N\zeta}{2}\boldsymbol{\omega}^T\mathcal{K}\boldsymbol{\omega},\end{aligned} \tag{5}$$

where $\boldsymbol{x} = (x_1,\ldots,x_N)^T$, $\Phi = (\boldsymbol{\phi}(t_1),\ldots,\boldsymbol{\phi}(t_N))^T$, ζ (>0) is a smoothing parameter and $\mathcal{K}$ is a positive semi-definite matrix defined by $\mathcal{K} = D_2^T D_2$, where D_2 is a second-order difference term. The regularized maximum likelihood estimates are given by

$$\hat{\boldsymbol{\omega}} = (\Phi^T\Phi + N\zeta\hat{\sigma}^2\mathcal{K})^{-1}\Phi^T\boldsymbol{x}, \qquad \hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^N\left\{x_i - \hat{\boldsymbol{\omega}}^T\boldsymbol{\phi}(t_i)\right\}^2. \tag{6}$$

We obtain the optimal number of basis functions m and the value of the smoothing parameter ζ by using a model selection criterion GIC (Ando *et al.*, 2008) for each smooth curve as the minimizer of the form

$$\mathrm{GIC}(\zeta) = N\log(2\pi\hat{\sigma}^2) + N + 2\mathrm{tr}\{QR^{-1}\}, \tag{7}$$

where $\hat{\sigma}^2$ is given in Equation (6) and the $m\times m$ matrices Q and R are, respectively, given by

$$Q = \frac{1}{N\hat{\sigma}^2}\begin{pmatrix} \frac{1}{\hat{\sigma}^2}\Phi^T\Lambda^2\Phi - \zeta\mathcal{K}\hat{\boldsymbol{\omega}}\mathbf{1}_N^T\Lambda\Phi & \frac{1}{2\hat{\sigma}^4}\Phi^T\Lambda^3\mathbf{1}_N - \frac{1}{2\hat{\sigma}^2}\Phi^T\Lambda\mathbf{1}_N \\ \frac{1}{2\hat{\sigma}^4}\mathbf{1}_N^T\Lambda^3\Phi - \frac{1}{2\hat{\sigma}^2}\mathbf{1}_N^T\Lambda\Phi & \frac{1}{4\hat{\sigma}^6}\mathbf{1}_N^T\Lambda^4\mathbf{1}_N - \frac{N}{4\hat{\sigma}^2}\end{pmatrix}, \tag{8}$$

$$R = \frac{1}{N\hat{\sigma}^2}\begin{pmatrix} \Phi^T\Phi + N\zeta\hat{\sigma}^2\mathcal{K} & \frac{1}{\hat{\sigma}^2}\Phi^T\Lambda\mathbf{1}_N \\ \frac{1}{\hat{\sigma}^2}\mathbf{1}_N^T\Lambda\Phi & \frac{N}{2\hat{\sigma}^2}\end{pmatrix}, \tag{9}$$

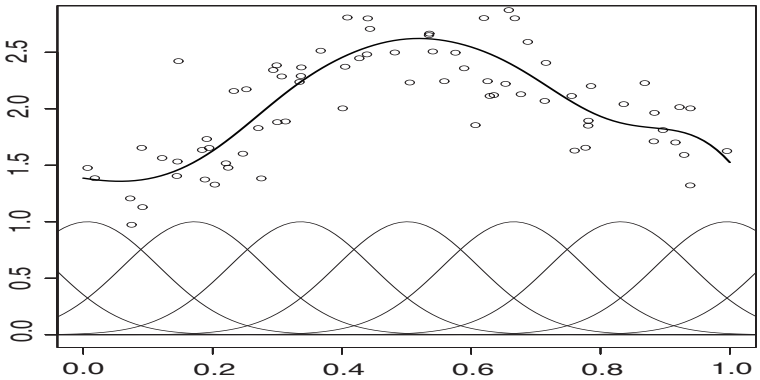


Figure 1: Functionalization by Gaussian basis expansions

where $\mathbf{1}_N = (1, \ldots, 1)^T$ and $\Lambda = \mathrm{diag}\left[x_1 - \hat{\boldsymbol{\omega}}^T \boldsymbol{\phi}(t_1), \ldots, x_N - \hat{\boldsymbol{\omega}}^T \boldsymbol{\phi}(t_N)\right]$.

Hence, the observed discrete data $\{(x_{\alpha i}, t_{\alpha i}); t_{\alpha i} \in \mathcal{T}, i = 1, \ldots, N_\alpha\}$ $(\alpha = 1, \ldots, n)$ are smoothed by the methodology described above, and we obtain a functional data set $\{x_\alpha(t);\ \alpha = 1, \ldots, n\}$ given by

$$\hat{u}(t) = \sum_{k=1}^{m} \hat{\omega}_{\alpha k} \phi_k(t) \equiv x_\alpha(t), \qquad t \in \mathcal{T}. \tag{10}$$

Figure 1 shows a sketch of the functionalization using Gaussian basis functions. Circles represent observed discrete data, the below solid curves basis functions pre-prepared and the above solid line the estimated smooth curve. For details of the functionalization step in functional data analysis, we refer to Ramsay and Silverman (2005) or Araki *et al.* (2009a).

3. Semi-supervised functional logistic discrimination

3.1. Semi-supervised logistic model for functional data

In the framework of semi-supervised functional data analysis, we are given n_1 labeled functional data $\{(x_\alpha(t), g_\alpha); \alpha = 1, \ldots, n_1,\ t \in \mathcal{T}\}$ and $(n - n_1)$ unlabeled functional data $\{x_\alpha(t); \alpha = n_1 + 1, \ldots, n,\ t \in \mathcal{T}\}$. Here $x_\alpha(t)$ are functional predictors given in the previous section and $g_\alpha \in \{1, \ldots, L\}$ are group indicator variables in which $g = k$ implies that the functional predictor $x_\alpha(t)$ belongs to group k. First, a functional logistic model is constructed by using only labeled functional data $\{(x_\alpha(t), g_\alpha); \alpha = 1, \ldots, n_1,\ t \in \mathcal{T}\}$.

We consider the posterior probabilities for group k $(k = 1, \ldots, L)$ given in a functional data $x_\alpha(t)$ as follows: $\Pr(g_\alpha = k | x_\alpha)$. Under these posterior probabilities, Araki *et al.* (2009b) introduced a functional logistic model in the form

$$\log\left\{\frac{\Pr(g_\alpha = k|x_\alpha)}{\Pr(g_\alpha = L|x_\alpha)}\right\} = \beta_{kf} + \int x_\alpha(t)\beta_k(t)dt, \qquad k = 1, \ldots, L-1. \tag{11}$$

By using the same Gaussian basis function $\phi_j(t)$ as in Equation (2), $\beta_k(t)$ is assumed to

be expanded as

$$\beta_k(t) = \sum_{j=1}^{m} \beta_{kj}\phi_j(t). \tag{12}$$

Then we can rewrite the functional logistic model in Equation (11) using the expansion in Equation (12) as follows:

$$\log\left\{\frac{\Pr(g_\alpha = k|x_\alpha)}{\Pr(g_\alpha = L|x_\alpha)}\right\} = \beta_{kf} + \int x_\alpha(t)\beta_k(t)dt = \boldsymbol{\beta}_k^T \boldsymbol{z}_\alpha, \tag{13}$$

where $\boldsymbol{\beta}_k = (\beta_{kf}, \beta_{k1}, \ldots, \beta_{km})^T$ and $\boldsymbol{z}_\alpha = (1, \boldsymbol{w}_\alpha^T J)^T$. Here J is an $m \times m$ matrix with the (i,j)-th element

$$J_{ij} = \sqrt{\pi\hat{\eta}^2}\exp\left\{-\frac{(\hat{\mu}_i - \hat{\mu}_j)^2}{4\hat{\eta}^2}\right\}, \qquad i,j = 1,\ldots,m, \tag{14}$$

where $\hat{\mu}_i$ and $\hat{\eta}$ are estimated centers and width parameters included in Gaussian basis functions in Section 2, respectively.

Thus the conditional probabilities can be rewritten as

$$\begin{aligned}
\Pr(g_\alpha = k|x_\alpha) &= \frac{\exp\{\boldsymbol{\beta}_k^T \boldsymbol{z}_\alpha\}}{1 + \sum_{j=1}^{L-1}\exp\{\boldsymbol{\beta}_j^T \boldsymbol{z}_\alpha\}}, \quad k = 1,\ldots,L-1, \\
\Pr(g_\alpha = L|x_\alpha) &= \frac{1}{1 + \sum_{j=1}^{L-1}\exp\{\boldsymbol{\beta}_j^T \boldsymbol{z}_\alpha\}}.
\end{aligned} \tag{15}$$

We describe $\Pr(g_\alpha = k|x_\alpha)$ as $\pi_k(x_\alpha;\boldsymbol{\beta})$, since the probabilities depend on a parameter vector $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_{L-1}^T)^T$.

We introduce an $(L-1)$-dimensional response variable $\boldsymbol{y}_\alpha = (y_1^{(\alpha)}, \ldots, y_{L-1}^{(\alpha)})^T$ ($\alpha = 1,\ldots,n_1$), which indicates that the k-th element of $\boldsymbol{y}_\alpha$ is set to 1 if the corresponding $x_\alpha(t)$ belongs to the k-th class, for n_1 labeled functional data $\{(x_\alpha(t), g_\alpha); \alpha = 1,\ldots,n_1\}$. Hence we obtain a multinomial distribution with the posterior probabilities $\pi_k(x_\alpha;\boldsymbol{\beta})$ as follows:

$$f(\boldsymbol{y}_\alpha|x_\alpha;\boldsymbol{\beta}) = \prod_{k=1}^{L-1}\pi_k(x_\alpha;\boldsymbol{\beta})^{y_k^{(\alpha)}}\{\pi_L(x_\alpha;\boldsymbol{\beta})\}^{1-\sum_{j=1}^{L-1}y_j^{(\alpha)}}. \tag{16}$$

By introducing a dummy class label variable $\boldsymbol{t}_\alpha$ for unlabeled functional data $\{x_\alpha(t); \alpha = n_1+1,\ldots,n\}$ given by

$$\boldsymbol{t}_\alpha = (t_1^{(\alpha)}, \ldots, t_{L-1}^{(\alpha)})^T = \begin{cases} (0,\ldots,0,\overset{(k)}{1},0,\ldots,0)^T & \text{if } x_\alpha(t) \text{ belongs to } k\text{-th class}, \\ (0,\ldots,0)^T & \text{if } x_\alpha(t) \text{ belongs to } L\text{-th class}, \end{cases}$$

it is assumed that $\boldsymbol{t}_\alpha$ is distributed as the same multinomial distribution with the posterior probabilities $\pi_k(x_\alpha;\boldsymbol{\beta})$ as in Equation (16). Also, for unlabeled functional data, we

assume $\beta_{kf} + \int x_\alpha(t)\beta_k(t) = \boldsymbol{\beta}_k^T \boldsymbol{z}_\alpha$ ($\alpha = n_1+1, \ldots, n;\ k = 1, \ldots, L-1$) similar to Equation (13). The log-likelihood function based on both labeled and unlabeled functional data is then obtained by

$$\ell(\boldsymbol{\beta}) = \sum_{\alpha=1}^{n_1} \left[\sum_{k=1}^{L-1} y_k^{(\alpha)} \boldsymbol{\beta}_k^T \boldsymbol{z}_\alpha - \log\left(1 + \sum_{l=1}^{L-1} \exp\{\boldsymbol{\beta}_l^T \boldsymbol{z}_\alpha\}\right)\right] + \sum_{\alpha=n_1+1}^{n} \left[\sum_{k=1}^{L-1} t_k^{(\alpha)} \boldsymbol{\beta}_k^T \boldsymbol{z}_\alpha - \log\left(1 + \sum_{l=1}^{L-1} \exp\{\boldsymbol{\beta}_l^T \boldsymbol{z}_\alpha\}\right)\right]. \tag{17}$$

3.2. Estimation via regularization

As mentioned in Araki *et al.* (2009b), the maximum likelihood method often causes some ill-posed problems for a functional logistic model; i.e., unstable or infinite parameter estimates. Then we employ a regularization method to obtain the estimator of the parameters included in the functional logistic model. A regularization method achieves to maximize a regularized log-likelihood function

$$\ell_\lambda(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) - \frac{n_1 \lambda}{2} \sum_{k=1}^{L-1} \boldsymbol{\beta}_k^T K \boldsymbol{\beta}_k, \tag{18}$$

where λ (> 0) is a regularization parameter and K is an $(m+1) \times (m+1)$ matrix given by

$$K = \begin{pmatrix} 0 & \mathbf{0}^T \\ \mathbf{0} & K^* \end{pmatrix}. \tag{19}$$

Here $\mathbf{0}$ is an m-dimensional zero vector and K^* is an $m \times m$ positive semi-definite matrix. In the section of numerical examples, we use an identity matrix as the matrix K^*.

In maximizing the regularized log-likelihood function in Equation (18), it is difficult to obtain the estimator of the parameters, since the values of dummy class labels $\boldsymbol{t}$ are unknown and $\partial \ell_\lambda(\boldsymbol{\beta}) / \partial \boldsymbol{\beta} = \mathbf{0}$ does not have an explicit solution with respect to the parameter vector $\boldsymbol{\beta}$. Hence, we employ a following EM-based algorithm to obtain the estimator $\hat{\boldsymbol{\beta}}$.

Step1 Initializing the parameter vector $\boldsymbol{\beta}$ by maximizing the regularized log-likelihood function via only labeled functional data $\{(x_\alpha(t), g_\alpha); \alpha = 1, \ldots, n_1\}$ with the help of Fisher's scoring method.

Step2 Construct a classification rule $\pi_k(x_\alpha; \hat{\boldsymbol{\beta}})$.

Step3 (E-step) By the use of the classification rule in Step2, compute the posterior probabilities $\pi_k(x_\alpha; \hat{\boldsymbol{\beta}})$ ($k = 1, \ldots, L$) for unlabeled functional data $x_\alpha(t)$ ($\alpha = n_1 + 1, \ldots, n$). According to the posterior probabilities, estimate $\boldsymbol{t}_\alpha$ as follows:

$$\hat{\boldsymbol{t}}_\alpha = (\hat{t}_1^{(\alpha)}, \ldots, \hat{t}_{L-1}^{(\alpha)})^T = (\pi_1(x_\alpha; \hat{\boldsymbol{\beta}}), \ldots, \pi_{L-1}(x_\alpha; \hat{\boldsymbol{\beta}}))^T. \tag{20}$$

Note that $\hat{t}_k^{(\alpha)}$ is the conditional expectation of $t_k^{(\alpha)} (k = 1, \ldots, L-1)$.

Step4 (M-step) Replace $t_k^{(\alpha)}$ into $\hat{t}_k^{(\alpha)}$ in the regularized log-likelihood function. Then estimate the parameter vector $\boldsymbol{\beta}$ using Fisher's scoring method.

Step5 Repeat the Step2 to the Step4 until the convergence condition

$$|\ell_\lambda(\hat{\boldsymbol{\beta}}^{(k+1)}) - \ell_\lambda(\hat{\boldsymbol{\beta}}^{(k)})| < 10^{-5} \tag{21}$$

is satisfied, where $\hat{\boldsymbol{\beta}}^{(k)}$ is the value of $\boldsymbol{\beta}$ after the k-th EM iteration.

Therefore, we derive a statistical model $f(\boldsymbol{y}|x;\hat{\boldsymbol{\beta}})$ which is constructed by using both labeled and unlabeled functional data. The statistical model includes a tuning parameter; i.e., the regularization parameter λ. Since the selection of this parameter is regarded as the selection of candidate models, we introduce model selection criteria to choose the constructed models.

4. Model selection criteria

In this section, we derive two types of model selection criteria to evaluate semi-supervised functional logistic models from the viewpoints of information-theoretic and Bayesian approaches.

4.1. Generalized information criterion

Akaike (1974) proposed the Akaike information criterion (AIC), which enables us to evaluate statistical models estimated by the maximum likelihood method. While the AIC is very useful for various fields of research, the criterion cannot be directly applied into models constructed by other estimation procedures.

Konishi and Kitagawa (1996) introduced an information criterion, which can evaluate models constructed by various estimation procedures including robust, Bayesian and regularization methods. Using the result of Konishi and Kitagawa (1996), we propose a generalized information criterion (GIC) in the context of the semi-supervised functional logistic model. The model selection criterion is given as follows:

$$\mathrm{GIC} = -2\sum_{\alpha=1}^{n_1} \log f(\boldsymbol{y}_\alpha|x_\alpha;\hat{\boldsymbol{\beta}}) + 2\mathrm{tr}\left\{Q(\hat{\boldsymbol{\beta}})R^{-1}(\hat{\boldsymbol{\beta}})\right\}, \tag{22}$$

where the matrices $Q(\hat{\boldsymbol{\beta}})$ and $R(\hat{\boldsymbol{\beta}})$ are

$$Q(\hat{\boldsymbol{\beta}}) = \frac{1}{n_1}\left[\{(B-C)\odot A\}^T - \lambda E\hat{\boldsymbol{\beta}}\mathbf{1}_{n_1}^T\right]\{(B-C)\odot A\}, \tag{23}$$

$$R(\hat{\boldsymbol{\beta}}) = -\frac{1}{n_1}(C\odot A)^T(C\odot A) + \frac{1}{n_1}D + \lambda E, \tag{24}$$

with

$$\begin{aligned}
A &= (Z,\ldots,Z), \qquad n_1\times(m+1)(L-1),\\
B &= (\boldsymbol{y}_{(1)}\mathbf{1}_{m+1}^T,\ldots,\boldsymbol{y}_{(L-1)}\mathbf{1}_{m+1}^T)^T,\\
C &= (\boldsymbol{\pi}_{(1)}\mathbf{1}_{m+1}^T,\ldots,\boldsymbol{\pi}_{(L-1)}\mathbf{1}_{m+1}^T)^T,\\
D &= \text{block diag}\{Z^T\mathrm{diag}(\boldsymbol{\pi}_{(1)})Z,\ldots,Z^T\mathrm{diag}(\boldsymbol{\pi}_{(L-1)})Z\},\\
E &= \text{block diag}(K,\ldots,K), \qquad (m+1)(L-1)\times(m+1)(L-1),\\
Z &= (\boldsymbol{z}_1,\ldots,\boldsymbol{z}_{n_1})^T,\\
\boldsymbol{y}_{(k)} &= (y_k^{(1)},\ldots,y_k^{(n_1)})^T,\\
\boldsymbol{\pi}_{(k)} &= (\pi_k(x_1;\hat{\boldsymbol{\beta}}),\ldots,\pi_k(x_{n_1};\hat{\boldsymbol{\beta}}))^T.
\end{aligned}$$

Here the operator $\odot$ denotes the Hadamard product, which means the elementwise product of matrices; that is, $A_{ij} \odot B_{ij} = (a_{ij} b_{ij})$ for matrices $A_{ij} = (a_{ij})$ and $B_{ij} = (b_{ij})$.

4.2. Generalized Bayesian information criterion

In Bayesian inference, Schwarz (1978) presented the Bayesian information criterion (BIC) from the viewpoint of maximizing a marginal likelihood. However, the BIC covers only models estimated by the maximum likelihood method.

By extending the Schwarz's (1978) idea, Konishi *et al.* (2004) derived a novel Bayesian information criterion to evaluate models estimated by regularization in the framework of generalized linear models. Hence, by using the result given in Konishi *et al.* (2004), we present a generalized Bayesian information criterion (GBIC) for evaluating the statistical model constructed by the semi-supervised functional logistic modeling procedure in the form

$$\begin{aligned}\mathrm{GBIC} = -2\sum_{\alpha=1}^{n_1} \log f(\boldsymbol{y}_\alpha | x_\alpha; \hat{\boldsymbol{\beta}}) + n_1 \lambda \sum_{k=1}^{L-1} \hat{\boldsymbol{\beta}}_k^T K \hat{\boldsymbol{\beta}}_k - (L-1)\log |K|_+ \\ + \log |R(\hat{\boldsymbol{\beta}})| - (L-1)(m+1-d)\log\lambda - (L-1)d\log\left(\frac{2\pi}{n_1}\right), \qquad (25)\end{aligned}$$

where $R(\hat{\boldsymbol{\beta}})$ is given by Equation (24) and $|K|_+$ is the product of the positive eigenvalues of K with the rank d.

We thus select a tuning parameter λ by minimizing either the model selection criterion GIC or GBIC. For more details of derivations about the model selection criteria, we refer to Konishi and Kitagawa (2008).

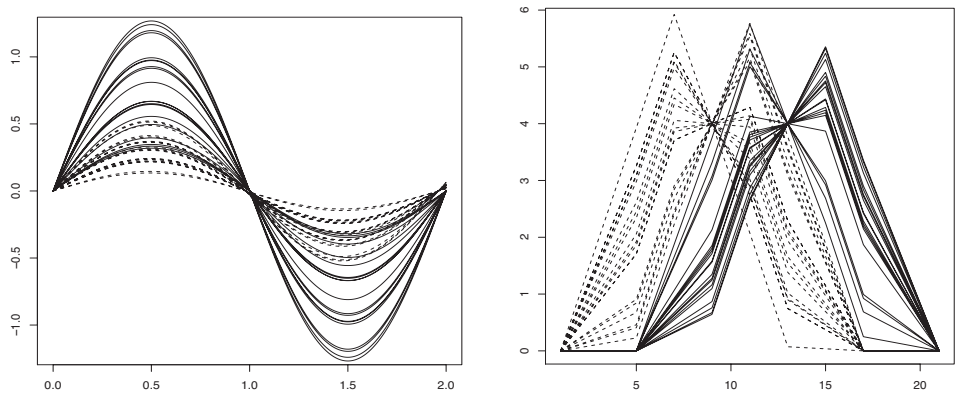
Note that the GIC in (22) and the GBIC in (25) are proposed based on the log-likelihood function from only labeled functional data. The reason why we employ the log-likelihood function based on only labeled functional data is according to Hirose *et al.* (2008) and Kawano and Konishi (2011). It may be possible to introduce model selection criteria based on the log-likelihood function from both labeled and unlabeled functional data in Equation (17). We consider this as our future research topic.

5. Numerical studies

We conducted some numerical examples to investigate the effectiveness of the proposed modeling procedure. Monte Carlo simulations and a real data analysis are given to illustrate our proposed semi-supervised functional logistic modeling strategy.

5.1. Monte Carlo simulations

We demonstrated the efficiency of the proposed functional logistic modeling procedure through Monte Carlo simulations. In the simulation study, we generated n discrete samples $\{(x_{\alpha t_i}, g_\alpha); \alpha = 1, \ldots, n,\ i = 1, \ldots, l\}$, where predictors $x_{\alpha t_i}$ are assumed to be obtained by $x_{\alpha t_i} = h_\alpha(t_i) + \varepsilon_{\alpha t_i}$ and the class label g_α indicates 1 or 2 which is the



(a) (b)

Figure 2: True functions for (a) Case 1 and (b) Case 2. In each case, there are 10 subjects. Solid lines represent the group 1, while dashed lines represent the group 2.

group number. We considered two settings as follows:

Case 1

$$h_\alpha(t_i) = \sin(c_\alpha t_i \pi) u_\alpha,\ \varepsilon_{\alpha t_i} \sim N(0, 0.1),\ t_i = \frac{2i-2}{49},\ n = 600,\ l = 50,$$
$$g_\alpha = 1 : c_\alpha = 1,\ u_\alpha \sim U[0.3, 1.3],$$
$$g_\alpha = 2 : c_\alpha = 1.02,\ u_\alpha \sim U[0.1, 0.6],$$

Case 2

$$h_\alpha(t_i) = u_\alpha w(t_i) + (1 - u_\alpha) v(t_i),\ \varepsilon_{\alpha t_i} \sim N(0, 1),\ t_i = \frac{i+4}{5},\ n = 600,\ l = 101,$$
$$g_\alpha = 1 : u_\alpha \sim U[0, 1],\ w(t_i) = \max(6 - |t_i - 11|, 0),\ v(t_i) = \max(6 - |t_i - 11|, 0) - 4,$$
$$g_\alpha = 2 : u_\alpha \sim U[0, 1],\ w(t_i) = \max(6 - |t_i - 11|, 0),\ v(t_i) = \max(6 - |t_i - 11|, 0) + 4.$$

Figure 2 denotes the true functions $h(t)$ for the Cases 1 and the Case 2, respectively. We divided the data set into 300 training data and 300 test data with an equal prior probability for each class. In order to implement the semi-supervised method, the training data were randomly divided into two halves with labeled functional data and unlabeled functional data, where the labeled functional data were assigned as 5%, 10%, 20%, 30%, 40%, 50% and 60% of the training data, respectively.

We compared the performances of semi-supervised functional logistic model (SFLDA) with those of supervised functional logistic model (FLDA) proposed by Araki *et al.* (2009b), support vector machine with the RBF kernel (SVM), k-nearest neighbor classification (KNN), functional support vector machine with the RBF kernel (FSVM) proposed by Rossi and Villa (2006), and semi-supervised methods proposed by Zhou *et al.* (2004) (LLGC: learning with local and global consistency) and Yu *et al.* (2004) (ILLGC: inductive learning with local and global consistency). The discrete data set

Table 1: Comparison of test errors with different percentages of labeled functional data in the training data set for the Case 1. Figures in parentheses indicate the model selection criteria used in the simulation study.

Method \ %	5	10	20	30	40	50	60
SFLAD (GIC)	0.269	0.210	0.202	0.192	0.189	0.186	0.185
FLDA (GIC)	0.248	0.216	0.204	0.193	0.187	0.185	0.184
SFLAD (GBIC)	0.271	0.210	0.202	0.193	0.188	0.185	0.185
FLDA (GBIC)	0.359	0.237	0.200	0.188	0.185	0.183	0.182
SVM	0.278	0.221	0.203	0.195	0.194	0.183	0.185
KNN	0.268	0.244	0.236	0.228	0.225	0.220	0.215
FSVM	0.322	0.266	0.253	0.231	0.229	0.218	0.215
LLGC	0.313	0.255	0.227	0.204	0.197	0.192	0.187
ILLGC	0.335	0.255	0.221	0.200	0.193	0.189	0.185

was transformed into a functional data set using the smoothing technique described in Section 2. Semi-supervised and supervised functional modeling strategies (i.e., SFLDA, FLDA and FSVM) were applied into the functional data set. The regularization parameter in the SFLDA and the FLDA was selected by using the GIC or the GBIC. For the GIC or the GBIC of the FLDA, we refer to Araki *et al.* (2009a; 2009b). Adjusted parameters included in the SVM, the FSVM, the LLGC and the ILLGC were optimized by the five-fold cross validation, respectively. The number of neighbors k in the KNN was selected by the leave-one-out cross validation.

Tables 1 and 2 show comparisons of the test error rates for the simulated data. These values were averaged over 50 repetitions. The average values of the tuning parameter λ for 50 runs of the Case 1 were $\lambda = 5.96 \times 10^{-5}$ for the GIC and $\lambda = 9.48 \times 10^{-5}$ for the GBIC, while those of the Case 2 were $\lambda = 1.00 \times 10^{-2}$ for the GIC and $\lambda = 2.28 \times 10^{-2}$ for the GBIC. For the Case 1, we observe that the SFLDA methods evaluated by the GIC and the GBIC are superior to other methods except for the FLDA methods in almost all cases. Also, our proposed methods SFLDA seem to provide lower misclassification errors than the FLDA methods, when the size of labeled functional data is small (e.g., 10% of training data). In the case of the Case 2, the SFLDA methods outperform the SVM, the KNN, the FSVM, the LLGC and the ILLGC in all situations with respect to minimizing the test errors. In addition, the proposed procedures SFLDA may be competitive or slightly superior to the FLDA methods.

5.2. Microarray data analysis

We describe an application of the semi-supervised functional discriminant analysis to yeast gene expression data given in Spellman *et al.* (1998). This data set contains 77 microarrays and consists of two short time-courses (i.e., two time points) and four medium time-courses (18, 24, 17 and 14 time points). About 800 genes were classified into five different cell-cycle phases, namely, M/G1, G1, S, S/G2 and G2/M phases, while the other 5,378 genes were not classified. For more details of this data set, we refer to Spellman *et al.* (1998).

Table 2: Comparison of test errors with different percentages of labeled functional data in the training data set for the Case 2. Figures in parentheses indicate the model selection criteria used in the simulation study.

Method \ %	5	10	20	30	40	50	60
SFLAD (GIC)	0.056	0.040	0.032	0.031	0.029	0.028	0.027
FLDA (GIC)	0.056	0.043	0.035	0.029	0.029	0.029	0.027
SFLAD (GBIC)	0.056	0.040	0.032	0.029	0.029	0.028	0.026
FLDA (GBIC)	0.056	0.043	0.035	0.029	0.029	0.028	0.026
SVM	0.075	0.056	0.040	0.037	0.034	0.030	0.031
KNN	0.068	0.062	0.052	0.051	0.050	0.047	0.048
FSVM	0.107	0.081	0.068	0.057	0.057	0.053	0.054
LLGC	0.124	0.082	0.062	0.049	0.043	0.040	0.040
ILLGC	0.111	0.049	0.040	0.035	0.031	0.030	0.030

In our analysis, we used the "cdc15-based experiment data" sampled over 24 points after synchronization. For simplicity, any genes that contain missing values across any of the 24 time points were discarded. These expression data were considered to be a discretized realization of 632 expression curves evaluated at 24 time points. We functionalized the data using the smoothing methodology given in Section 2. A total of 300 genes were used as the training data set, and the remaining 332 genes were used as the test data set. We compared the SFLDA, which is our proposed semi-supervised functional method, with the FLDA, which is the supervised functional method.

First, we demonstrated the effectiveness of our semi-supervised methodology by setting functional data with known class labels as unlabeled functional data. We randomly split the training data set into labeled functional data and unlabeled functional data, where 15%, 20%, 30%, 40% and 50% of training data are allocated as labeled functional data, respectively, and we repeated the procedures 10 times. The values of the selected regularization parameter for 10 runs were $\lambda = 2.80 \times 10^{-5}$ for the GIC and $\lambda = 7.78 \times 10^{-4}$ for the GBIC. Figure 3 shows the average precisions of the test data set for different ratios of labeled-unlabeled functional data in the training data set. On the x-axis, 15 means that 15% of the training data was assigned as labeled functional data, and the remaining 85% was used as unlabeled functional data. From the left panel of Figure 3, we observe that the SFLDA with the GIC seems to extract useful information from unlabeled functional data, since the SFLDA performs better than the FLDA in all cases. In contrast, the right panel of Figure 3 shows that the SFLDA is superior to the FLDA until 30% labeled functional data, whereas the SFLDA is comparable to the FLDA in the range from 30% to 50% labeled functional data.

Second, we examined the performances of our methods by using real unlabeled functional data which were not classified by Spellman *et al.* (1998). We prepared labeled functional data which consist of 20%, 25%, 30%, 40%, 50% and 60% of the training data, while unlabeled functional data are set to 500 samples randomly selected from 5,378 real unlabeled examples. Our proposed models and the supervised functional models were applied into the data set. We repeated these procedures 10 times. We obtained the averaged optimal values of the regularization parameter for 10 repetitions

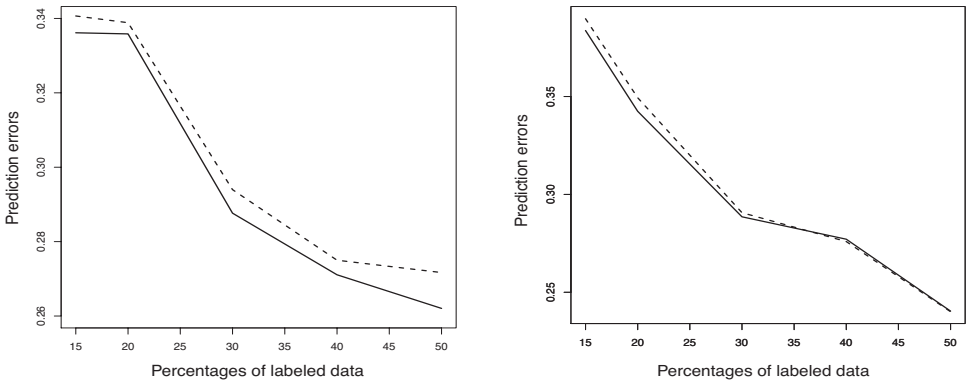


Figure 3: Average prediction errors for several ratios of labeled functional data in the training data set. Solid line shows the result of the SFLDA while dashed line shows that of the FLDA. The left-hand panel indicates the results for the methods evaluated by the GIC, whereas the right-hand panel indicates those by the GBIC.

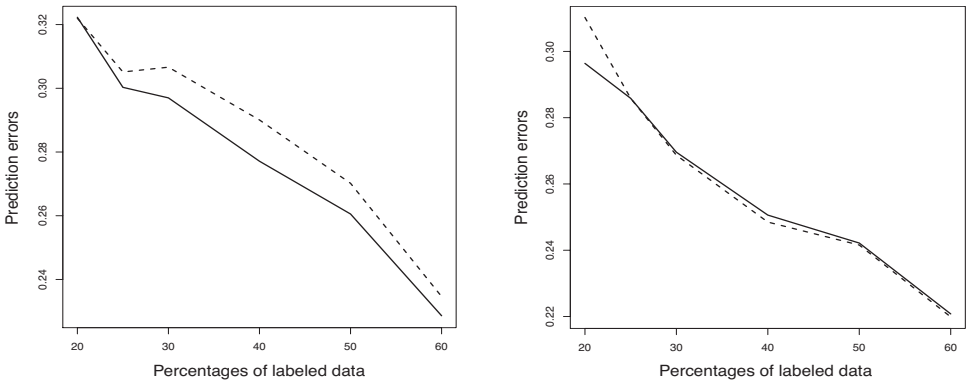


Figure 4: Average prediction errors for several ratios of labeled functional data in the training data set, where we use real unlabeled functional data. Solid line shows the result of the SFLDA while dashed line shows that of the FLDA. The left-hand panel indicates the results for the methods evaluated by the GIC, whereas the right-hand panel indicates those by the GBIC.

as $\lambda = 1.00 \times 10^{-5}$ for the GIC and $\lambda = 7.85 \times 10^{-5}$ for the GBIC. Figure 4 shows the average test error rates for various ratios of labeled functional data in the training data set. For the left-hand panel of Figure 4, the SFLDA outperforms the FLDA without 20% labeled functional data, while the SFLDA gives lower prediction errors than the FLDA on 20% labeled functional data. Hence, these results suggest that real unlabeled functional data included in Spellman's *et al.* (1998) data set may have a potential for improving a prediction accuracy of our functional logistic procedures.

6. Concluding remarks

We proposed a semi-supervised functional logistic modeling procedure for the multi-class classification problem with the help of regularization. On the step of functionalization, a smoothing method using Gaussian basis expansions was applied to the observed discrete data set. A crucial issue for our semi-supervised modeling process is the choice of the regularization parameter λ. In order to select the value of the parameter, we introduced model selection criteria from the viewpoints of information-theoretic and Bayesian approaches. Monte Carlo simulations and a microarray data analysis showed that our modeling strategy yields relatively lower prediction error rates than previously developed methods. A further research should be to construct a semi-supervised functional regression modeling or clustering.

Acknowledgement

The authors would like to thank the anonymous reviewer for his helpful comments. This work was supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Young Scientists (B), #24700280, 2012–2015. The computational resource was also provided by the Super Computer System, Human Genome Center, Institute of Medical Science, University of Tokyo.

References

Abraham, C., Cornillon, P. A., Matzner-Lober, E. and Molinari, N. (2003). Unsupervised curve clustering using B-splines. *Scandinavian Journal of Statistics*, **30**, 581–595.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **AC-19**, 716–723.

Ando, T., Konishi, S. and Imoto, S. (2008). Nonlinear regression modeling via regularized radial basis function networks. *Journal of Statistical Planning and Inference*, **138**, 3616–3633.

Araki, Y., Konishi, S., Kawano, S. and Matsui, H. (2009a). Functional regression modeling via regularized Gaussian basis expansions. *Annals of the Institute of Statistical Mathematics*, **61**, 811–833.

Araki, Y., Konishi, S., Kawano, S. and Matsui, H. (2009b). Functional logistic discrimination via regularized basis expansions. *Communications in Statistics - Theory and Methods*, **38**, 2944–2957.

Basu, S., Bilenko, M. and Mooney, R. J. (2004). A probabilistic framework for semi-supervised clustering. *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM Press, 59–68.

Chapelle, O., Schölkopf, B. and Zien, A. (2006). *Semi-Supervised Learning.* Cambridge, MA: MIT Press.

Chiou, J. M. and Li, P. L. (2007). Functional clustering and identifying substructures of longitudinal data. *Journal of the Royal Statistical Society Series B*, **69**, 679–699.

Dean, N., Murphy, T. B. and Downey, G. (2006). Using unlabelled data to update classification rules with applications in food authenticity studies. *Journal of the Royal Statistical Society Series C*, **55**, 1–14.

Ferraty, F. and Vieu, P. (2003). Curves discrimination: a nonparametric functional approach. *Computational Statistics and Data Analysis*, **44**, 161–173.

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis.* New York: Springer.

Hirose, K., Kawano, S. and Konishi, S. (2008). Bayesian factor analysis and information criterion. *Bulletin of Informatics and Cybernetics*, **40**, 75–87.

James, G. M. and Silverman, B. W. (2005). Functional adaptive model estimation. *Journal of the American Statistical Association*, **100**, 565–576.

Kawano, S. and Konishi, S. (2007). Nonlinear regression modeling via regularized Gaussian basis functions. *Bulletin of Informatics and Cybernetics*, **39**, 83–96.

Kawano, S. and Konishi, S. (2011). Semi-supervised logistic discrimination via regularized Gaussian basis expansions. *Communications in Statistics - Theory and Methods*, **40**, 2412–2423

Konishi, S., Ando, T. and Imoto, S. (2004). Bayesian information criteria and smoothing parameter selection in radial basis function networks. *Biometrika*, **91**, 27–43.

Konishi, S. and Kitagawa, G. (1996). Generalised information criteria in model selection. *Biometrika*, **83**, 875–890.

Konishi, S. and Kitagawa, G. (2008). *Information Criteria and Statistical Modeling.* New York: Springer.

Kulis, B., Basu, S., Dhillon, I. and Mooney, R. (2009). Semi-supervised graph clustering: a kernel approach. *Machine Learning*, **74**, 1–22.

Lafferty, J. and Wasserman, L. (2007). Statistical analysis of semi-supervised regression. *Advances in Neural Information Processing Systems*, **21**, 801–808.

Liang, F., Mukherjee, S. and West, M. (2007). The use of unlabeled data in predictive modeling. *Statistical Science*, **22**, 189–205.

Miller, D. and Uyar, H. S. (1997). A mixture of experts classifier with learning based on both labelled and unlabelled data. *Advances in Neural Information Processing Systems*, **9**, 571–577.

Ng, M. K., Chan, E. Y., So, M. M. C. and Ching, W. K. (2006). A semi-supervised regression model for mixed numerical and categorical variables. *Pattern Recognition*, **40**, 1745–1752.

Ramsay, J. O. and Silverman, B. W. (2002). *Applied Functional Data Analysis.* New

York: Springer.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis.* Second Edition. New York: Springer.

Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society Series B*, **53**, 233–243.

Rossi, F., Conan-Guez, B. and Goli, A. E. (2004). Clustering functional data with the SOM algorithm. *Proceedings of XIIth European Symposium on Artificial Neural Networks*, Bruges, 305–312.

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

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

Silverman, B. W. (1996). Smoothed functional principal components analysis by choice of norm. *Annals of Statistics*, **24**, 1–24.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K. *et al.* (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell*, **9**, 3273–3297.

Verbeek, J. J. and Vlassis, N. (2006). Gaussian fields for semi-supervised regression and correspondence learning. *Pattern Recognition*, **39**, 1864–1875.

Yao, F. and Lee, T. C. M. (2006). Penalized spline models for functional principal component analysis. *Journal of the Royal Statistical Society Series B*, **68**, 3–25.

Yao, F., Müller, H. G. and Wang, J. L. (2005). Functional linear regression analysis for longitudinal data. *Annals of Statistics*, **33**, 2873–2903.

Yu, K., Tresp, V. and Zhou, D. (2004). Semi-supervised induction with basis functions. Max Planck Institute Technical Report 141, Max Planck Institute for Biological Cybernetics, Tübingen, Germany.

Zhong, S. (2006). Semi-supervised model-based document clustering: A comparative study. *Machine Learning*, **65**, 3–29.

Zhou, D., Bousquet, O., Lal, T. N., Weston, J. and Schölkopf, B. (2004). Learning with local and global consistency. *Advances in Neural Information Processing Systems*, **16**, 321–328.

Zhu, X. (2008). Semi-supervised learning literature survey. Computer Sciences Technical Report 1530, University of Wisconsin-Madison.

Received May 29, 2012
Revised July 13, 2012