Multi-class logistic discrimination via wavelet-based functionalization and model selection criteria

Fujii, Toru
Graduate School of Mathematics, Kyushu University

Konishi, Sadanori
Graduate School of Mathematics, Kyushu University

出版情報：MHF Preprint Series. 2006-25, 2006-07-27. 九州大学大学院数理学研究院
バージョン：
権利関係：

MHF Preprint Series

Kyushu University
21st Century COE Program
Development of Dynamic Mathematics with High Functionality

Multi-class Logistic Discrimination via Wavelet-based Functionalization and Model Selection Criteria

T. Fujii & S. Konishi

MHF 2006-25

(Received July 27, 2006)

Faculty of Mathematics
Kyushu University
Fukuoka, JAPAN

Multi-class logistic discrimination via wavelet-based functionalization and model selection criteria

Toru Fujii[1] **and Sadanori Konishi**[1]

[1] *Graduate School of Mathematics, Kyushu University,*
6-10-1 Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan.

Abstract

We consider a multi-class logistic discrimination for functional data. We use a wavelet-based smoothing technique in obtaining a set of functional data, from irregularly sampled time-dependent covariates of a number of individuals. A method of estimating discriminant model is based on a regularized log-likelihood, where we apply model selection criteria derived from Kullback-Leibler information and Bayes' analysis.

Keywords: Functional data analysis, Model selection, Multi-class logistic discrimination, Wavelets.

1. Introduction

Classification or discrimination have been important statistical problem areas in various fields of natural and social sciences. Several techniques have been proposed for analyzing multivariate data such as Fisher's linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) (see e.g. Hastie *et al.* (2003)).

It is often the case that dimension of covariates is quite high, while the whole population of training set is relatively small. In such cases, variance-covariance matrix becomes singular, and the Mahalanobis distance cannot be calculated. Besides the above problems caused by high dimensionality, this paper consider the case that covariates of data are given by temporal (and/or spatial) observations whose observational points may vary among individuals. The above discriminant methods based on multivariate vectors must ignore the time order of temporal observations, further, there may be problems in application for the case of un-uniform observational points.

To solve these ill-posed problems, we introduce a functional discriminant approach, which fits a curve (or a function) to the temporal observations of each individual, and then discriminates individuals based on the functionalized covariates. This approach is based on a framework of functional data analysis (FDA) proposed by Ramsay and Silverman (2002, 2005) and has been applied in various fields such as biomechanics, chemometrics, meteorology, and so on. Basis expansions for functionalization, such as Fourier bases and splines, have been very popular in FDA, while more recently, radial basis expansions have also been considered by Araki *et al.* (2004). The above basis expansions are known to be effective for analyzing temporal observations when underlying curve is sufficiently smooth. Here, however, we believe that the local adaptivity of wavelet-based curve estimation may yield favorable results when data have irregular and complex structures.

Wavelets form an orthonormal basis and enable multi-resolution analysis by localizing a function in different phases of both time and frequency domains simultaneously, and thus offer some advantages over traditional Fourier analysis for analyzing data with intrinsically local properties, such as discontinuities and sharp spikes. Wavelet-based methods have been predominantly applied in sound and image analysis due to their ability to detect edges and singularities. In statistics, applications of wavelet-based methods have been frequently reported by Donoho *et al.* (1995), Hall and Patil (1996) among others.

We apply a wavelet-based method of Fujii and Konishi (2005) for constructing functional covariates from temporal observations, and we then conduct a multi-class logistic discriminant analysis. A crucial issue in model building process is choice of smoothing parameter. We present an information-theoretic and Bayesian type criteria for evaluating models estimated by a method of regularization in the frame work of wavelet-based functional logistic model.

This paper is organized as follows. In Section 2 we describe a multi-class functional logistic model for data that the covariates are given by orthonormal wavelet expansions. In Section 3 we describe a wavelet-based functionalization of time-dependent observations. Section 4 provide an estimation procedure of multi-class logistic model based on a regularized log-likelihood and Newton-Raphson algorithm. In Section 5 we present model selection criteria to choose the smoothing parameter. A numerical study is given in Section 6. Finally, in Section 7, our proposed method is illustrated in a real data example given by an application to digitized analog signals of "phonemes", where this problem forms the subject of sound recognition in signal analysis.

2. Multi-class logistic model for functional data

Suppose we have n independent observations

$$\{(x_i(t),\, y_i)\,;\, i = 1, \ldots, n\}, \tag{1}$$

where $x_i(t)$ are functions given on $t \in \mathcal{T}$, and $y_i \in \{1, 2, \ldots, K\}$ denote classes to which $x_i(t)$ belong. We assume that the class labels y_i are generated from certain probability distributions $\Pr(Y_i = k|x_i(t))$ which are represented by

$$\log \frac{\Pr(Y_i = k|x_i(t))}{\Pr(Y_i = K|x_i(t))} = \beta_{k,0} + \int_{\mathcal{T}} \beta_k(t)\, x_i(t)\, \mathrm{d}t, \quad k = 1, \ldots, K-1, \tag{2}$$

where $\beta_{k,0}$ and $\beta_k(t)$ are unknown model parameters.

We also assume that covariate functions $x_i(t)$ are given in the form

$$x_i(t) = \sum_{m=1}^{M} \alpha_{i,m}\phi_m(t).$$

Then for $\beta_k(t) = \sum_{m=1}^{M} \beta_{k,m}\phi_m(t)$ expanded by the same bases as that of $x_i(t)$, it follows that the right-hand side of equation (2) may be

$$\beta_{k,0} + (\beta_{k,1}, \ldots, \beta_{k,M})\mathcal{J}(\alpha_{i,1}, \ldots, \alpha_{i,M})^T, \tag{3}$$

where $\mathcal{J}$ denotes the $M \times M$ matrix with (m_1, m_2)th elements given by $\int \phi_{m_1}(t)\phi_{m_2}(t)\, \mathrm{d}t$. Cardot and Sarda (2005) proposed an estimation of the coefficient function for functional generalized models based on a B-splines expansion and penalized likelihood. Araki *et al.* (2004) considerd the use of radial basis function networks for the functional logistic discrimination of sufficiently smooth time course data.

In this paper, we consider the use of wavelets for the bases $\{\phi_m(t)\,;\, m = 1, \ldots, M\}$. The orthonormal property of wavelets, i.e., $\int \phi_{m_1}(t)\phi_{m_2}(t)\, \mathrm{d}t = \delta_{m_1,m_2}$ yields that $\mathcal{J} = I$, and thus equation (2) is equivalent to

$$\log \frac{\Pr(Y_i = k|x_i(t))}{\Pr(Y_i = K|x_i(t))} = \boldsymbol{\beta}_k^T \boldsymbol{\alpha}_i\,, \quad k = 1, \ldots, K-1,$$

where $\boldsymbol{\alpha}_i = (1, \alpha_{i,1}, \ldots, \alpha_{i,M})^T$ and $\boldsymbol{\beta}_k = (\beta_{k,0}, \beta_{k,1}, \ldots, \beta_{k,M})^T$. It follows that the estimation of the model results in the estimation of the vector of coefficients $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_{K-1}^T)^T$.

3. Wavelet-based functionalization of irregulary sampled data

Although we assumed that the covariates of data (1) are given by wavelet expansions, in practice, it is natural that the coefficients $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n$ are unknown, while we instead observe the following

$$[\{(x_{i,l}, t_{i,l})\,;\, l = 1, \ldots, L_i\},\, y_i]. \tag{4}$$

Suppose that $\{x_{i,l}\,;\, l = 1, \ldots, L_i\}$ in data (4) are generated from the model

$$x_{i,l} = \boldsymbol{\alpha}_i^T \boldsymbol{\phi}_{i,l} + \varepsilon_{i,l}, \quad l = 1, \ldots, L_i,$$

where $\boldsymbol{\phi}_{i,l} = (\phi_1(t_{i,l}), \ldots, \phi_M(t_{i,l}))^T$, $\{t_{i,l}\,;\, l = 1, \ldots, L_i\}$ are fixed observational points, $\{\varepsilon_{i,l}\}$ are independently and normally distributed with mean 0 and variance σ_i^2, and $\boldsymbol{\alpha}_i$ are unknown. Note that, in this case, the observational points $\{t_{i,l}\,;\, l = 1, \ldots, L_i\}$ may vary among individuals. Hence there may be problems in constructing the discriminant model directly by using $\{x_{i,l}\,;\, l = 1, \ldots, L_i\}$ as covariate vectors.

It follows that probability model for data (4) may be represented by the density function

$$f(y_i\,;\boldsymbol{\alpha}_i, \boldsymbol{\beta}) = f(y_i \,|\, \boldsymbol{\alpha}_i\,;\boldsymbol{\beta}) \prod_{l=1}^{L_i} f(x_{i,l} \,|\, t_{i,l}\,;\boldsymbol{\alpha}_i, \sigma_i^2), \tag{5}$$

where $f(y_i \,|\, \boldsymbol{\alpha}_i\,;\boldsymbol{\beta}) = \Pr(Y_i = y_i | x_i(t))$ is the model for functional data (1) with given $\boldsymbol{\alpha}_i$, and

$$f(x_{i,l} \,|\, t_{i,l}\,;\boldsymbol{\alpha}_i, \sigma_i^2) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left\{ -\frac{(x_{i,l} - \boldsymbol{\alpha}_i^T \boldsymbol{\phi}_{i,l})^2}{2\sigma_i^2} \right\}, \tag{6}$$

is the model of $\{x_{i,l}\,;\, l = 1, \ldots, L_i\}$ with unknown $\boldsymbol{\alpha}_i$.

By means of the generalized linear models for functional data, James (2002) suggested the estimation of the full model (5) directly from the discrete observation (4) by using the EM algorithm (Dempster *et al.*, 1977), in which the coefficients $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n$ are considered as unobserved latent variables having some prior density. In this method, however, there may be difficulties in deciding how to determine the coefficients $\boldsymbol{\alpha}_{i^*}$ for the future observations.

To avoid this problem, we consider the following 2-step estimation procedure:

step 1. estimate the parameter $\boldsymbol{\alpha}_i$ of the model (6) for $i = 1, \ldots, n$.

step 2. estimate the parameter $\boldsymbol{\beta}$ of model $f(y_i \,|\, \boldsymbol{\alpha}_i \,; \boldsymbol{\beta})$ with $\boldsymbol{\alpha}_i \equiv \widehat{\boldsymbol{\alpha}}_i$ estimated in step 1.

We can then estimate the conditional probabilitis for unsupervised future observations as $\Pr(Y_{i^*} = k | x_{i^*}(t)) = \text{logit}^{-1}(\widehat{\boldsymbol{\beta}}_k^T \widehat{\boldsymbol{\alpha}}_{i^*})$, where we estimate $\boldsymbol{\alpha}_{i^*}$ in the same way as that used for $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n$.

We apply a wavelet-based regression method of Fujii and Konishi (2005) for the estimation in step 1. It might be noticed that the vast majority of wavelet-based regression estimation including Donoho (1995) has been conducted within the setting that given data is of decimal length and has equally spaced observational points. For the case that the observational points are not decimal and irregularly spaced such as that given by data (4), the corresponding basis matrix is no longer orthogonal, and wavelet-based decomposition/reconstruction procedure can not be directly applied. Several different approaches for irregular observational points have been made such as by Hall and Patil (1996) and Fujii and Konishi (2005) among others.

There may be an advantage to use a method of Fujii and Konishi (2005) because one can automatically choose smoothing parameters in estimating each $\boldsymbol{\alpha}_i$ by using model selection criteria (see Fujii and Konishi (2005, Section 3) for further details).

4. Estimation of the discriminant model

In this section, we assume that the coefficients $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n$ are already estimated in step 1 of Section 3, and they are given by $\boldsymbol{\alpha}_i \equiv \widehat{\boldsymbol{\alpha}}_i$. Here, we describe the procedure of estimation in step 2 of Section 3. We consider the maximization of regularized log-likelihood function given by

$$\ell_\lambda(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log f(y_i \,|\, \boldsymbol{\alpha}_i \,; \boldsymbol{\beta}) - \frac{n}{2} \lambda \, \boldsymbol{\beta}^T \mathcal{K} \boldsymbol{\beta}, \tag{7}$$

where the $(M+1)(K-1) \times (M+1)(K-1)$ diagonal matrix $\mathcal{K}$ has elements given by

$$\mathcal{K}_{(m,m)} = \begin{cases} 0 & m \equiv 1 \pmod{M+1}, \\ 1 & \text{otherwise.} \end{cases}$$

For an explicit representation of the log-likelihood function, we define $y_{i,1}, \ldots, y_{i,K-1}$ in place of the response $y_i \in \{1, \ldots, K\}$ by

$$y_{i,k} = \begin{cases} 1 & \text{if} \quad k = y_i, \\ 0 & \text{otherwise,} \end{cases}$$

and let $\mu_{i,k} = \Pr(Y_i = k|x_i(t))$ for $k = 1, \ldots, K-1$. It then follows from the multinomial nature of the distribution that

$$\begin{aligned}\log f(y_i \,|\, \boldsymbol{\alpha}_i \,; \boldsymbol{\beta}) &= \sum_{k=1}^{K-1} y_{i,k} \log \mu_{i,k} + \left(1 - \sum_{k=1}^{K-1} y_{i,k}\right) \log \left(1 - \sum_{k=1}^{K-1} \mu_{i,k}\right) \\ &= \sum_{k=1}^{K-1} y_{i,k} \boldsymbol{\beta}_k^T \boldsymbol{\alpha}_i - \log \left\{1 + \sum_{k=1}^{K-1} \exp(\boldsymbol{\beta}_k^T \boldsymbol{\alpha}_i)\right\}.\end{aligned}$$

Let $\boldsymbol{\mu}$, $\boldsymbol{y}$ and $\boldsymbol{\eta}$ be $n(K-1)$ dimensional vectors whose $\{i + n(k-1)\}$th elements are $\mu_{i,k}$, $y_{i,k}$ and $\eta_{i,k} = \boldsymbol{\beta}_k^T \boldsymbol{\alpha}_i$, respectively. It then follows that

$$\frac{\partial \ell_\lambda(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{\partial \boldsymbol{\eta}^T}{\partial \boldsymbol{\beta}}(\boldsymbol{y} - \boldsymbol{\mu}) - n\lambda \mathcal{K} \boldsymbol{\beta}, \qquad \frac{\partial^2 \ell_\lambda(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = -\frac{\partial \boldsymbol{\eta}^T}{\partial \boldsymbol{\beta}} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\eta}^T} \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\beta}^T} - n\lambda \mathcal{K},$$

where the elements of the above matrices are given respectively by

$$\frac{\partial \eta_{i,k}}{\partial \beta_{l,m}} = \delta_{k,l} \alpha_{i,m}, \quad \frac{\partial \mu_{i,k}}{\partial \eta_{j,l}} = \delta_{i,j} \mu_{i,k} (\delta_{k,l} - \mu_{j,l}).$$

Further, the regularized log-likelihood function (7) can be maximized by using the Newton-Raphson algorithm represented as follows:

$$\boldsymbol{\beta}^{new} = \boldsymbol{\beta}^{old} - \left(\frac{\partial^2 \ell_\lambda(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\right)^{-1} \frac{\partial \ell_\lambda(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \bigg|_{\boldsymbol{\beta} = \boldsymbol{\beta}^{old}}. \tag{8}$$

We start with $\boldsymbol{\beta}^{old} = \mathbf{0}$, and then update $\boldsymbol{\beta}^{old}$ with $\boldsymbol{\beta}^{new}$ calculated by equation (8) repeatedly until a certain convergence criterion is satisfied.

5. Model selection

To find an optimal model we must choose smoothing parameter λ. In this section, we present two different model selection criteria derived from Kullback-Leibler information and Bayes' analysis.

5.1. Generalized information criterion

Generalized information criterion (GIC) had been derived by Konishi and Kitagawa (1996) as a bias corrected estimator of the Kullback-Leibler information (Kullback and Leibler, 1951), which define a distance between true model and the model fitted by using the methods such as penalized log-likelihood estimation.

Hence by using the result given in Konishi and Kitagawa (1996, p. 889), we have the model selection criterion for evaluating the fitted logistic model $f(y_i|\boldsymbol{\alpha}_i\,;\widehat{\boldsymbol{\beta}}_\lambda)$ estimated by maximizing the penalized log-likelihood function (7),

$$\mathrm{GIC} = -2\,\ell_0(\widehat{\boldsymbol{\beta}}_\lambda) + 2\,\mathrm{tr}(R_\lambda^{-1}Q_\lambda),$$

where $\ell_\lambda(\cdot)$ is given by equation (7) and

$$\begin{aligned} R_\lambda &= -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2\{\log f(y_i\,|\,\boldsymbol{\alpha}_i\,;\boldsymbol{\beta}) - (\lambda/2)\boldsymbol{\beta}^T\mathcal{K}\boldsymbol{\beta}\}}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\bigg|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_\lambda} \\ &= \frac{1}{n}\frac{\partial\boldsymbol{\eta}^T}{\partial\boldsymbol{\beta}}\frac{\partial\boldsymbol{\mu}}{\partial\boldsymbol{\eta}^T}\frac{\partial\boldsymbol{\eta}}{\partial\boldsymbol{\beta}^T}\bigg|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_\lambda} + \lambda\mathcal{K}, \end{aligned} \tag{9}$$

$$\begin{aligned} Q_\lambda &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial\{\log f(y_i\,|\,\boldsymbol{\alpha}_i\,;\boldsymbol{\beta}) - (\lambda/2)\boldsymbol{\beta}^T\mathcal{K}\boldsymbol{\beta}\}}{\partial\boldsymbol{\beta}}\frac{\partial\log f(y_i\,|\,\boldsymbol{\alpha}_i\,;\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^T}\bigg|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_\lambda} \\ &= \frac{1}{n}\left\{\frac{\partial\boldsymbol{\eta}^T}{\partial\boldsymbol{\beta}}\mathrm{diag}(\boldsymbol{y}-\widehat{\boldsymbol{\mu}}_\lambda) - \lambda\mathcal{K}\widehat{\boldsymbol{\beta}}_\lambda\mathbf{1}_{n(K-1)}^T\right\}\mathrm{diag}(\boldsymbol{y}-\widehat{\boldsymbol{\mu}}_\lambda)\frac{\partial\boldsymbol{\eta}}{\partial\boldsymbol{\beta}^T}. \end{aligned}$$

5.2. Generalized bayesian information criterion

Konishi *et al.* (2004) extended Schwarz's BIC (Schwarz, 1978) to the evaluation of models fitted by the maximum penalized log-likelihood method or the method of regularization.

Let the prior density for the unknown parameter vector $\boldsymbol{\beta}$ to be a multivariate normal distribution given by

$$\pi(\boldsymbol{\beta}\,|\,\lambda) = (2\pi)^{-M(K-1)/2}(n\lambda)^{M(K-1)/2}|\mathcal{K}|_+^{1/2}\exp\left(-\frac{n\lambda}{2}\boldsymbol{\beta}^T\mathcal{K}\boldsymbol{\beta}\right),$$

where $\mathcal{K}$ is a matrix of rank $M(K-1)$ which appears in equation (7), and $|\mathcal{K}|_+$ denotes the product of $M(K-1)$ non-zero eigenvalues of $\mathcal{K}$. Then by using the result given in Konishi *et al.* (2004, p. 30), we have generalized bayesian information criterion (GBIC) given by

$$\mathrm{GBIC} = -2\,\ell_\lambda(\widehat{\boldsymbol{\beta}}_\lambda) - M(K-1)\log\lambda + \log|R_\lambda| - \log|\mathcal{K}|_+\,,$$

where the matrix R_λ is given by equation (9).

We choose the optimal values of the smoothing parameter λ by minimizing either GIC or GBIC criterion.

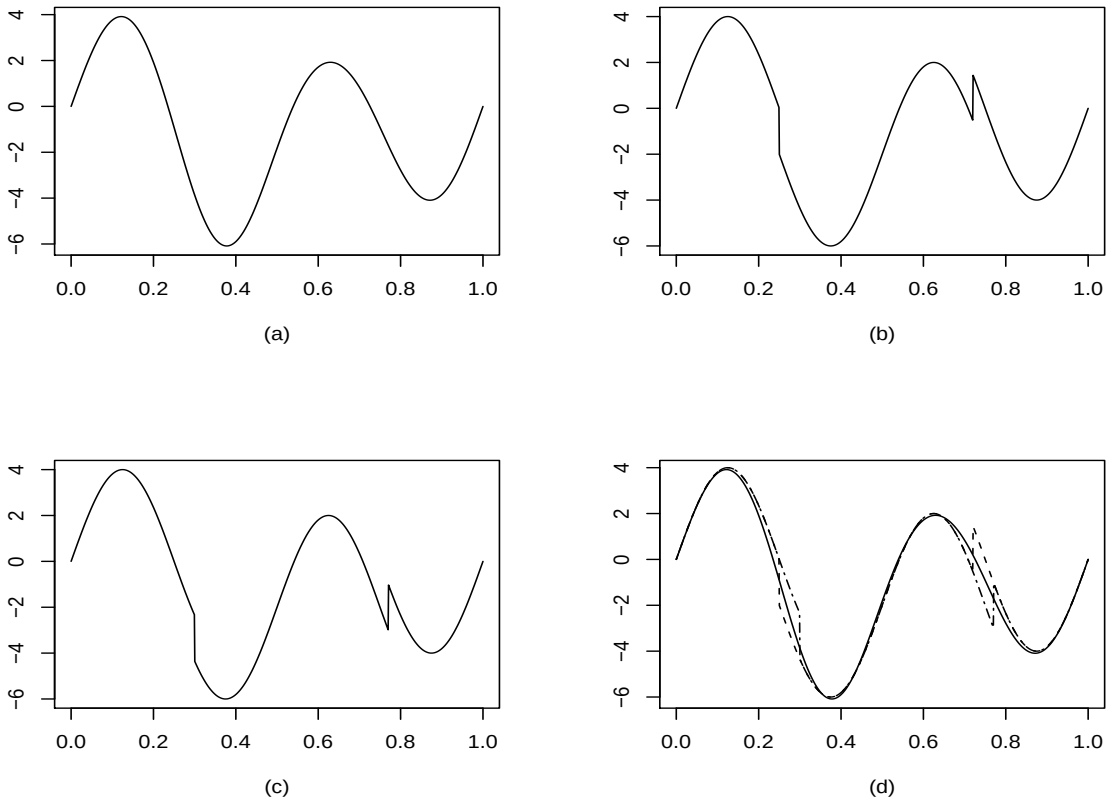

Figure 1: (a), (b) and (c) are the plots of true functions $x(t)$ for class 1, 2 and 3, respectively. (d) displays the true functions for three classes simultaneously.

6. Numerical study

In this section, we describe a multi-class discriminant analyses of simulated data. For the simulations, we consider the following sinusoidal functions as true covariates.

$$x(t) = \begin{cases} 4\sin(4\pi t) - 2\exp\{-(8t-3)^2\} - 2\exp\{-(8t-5)^2\} & \text{for class 1,} \\ 4\sin(4\pi t) - \mathrm{sgn}(t-.25) - \mathrm{sgn}(.72-t) & \text{for class 2,} \\ 4\sin(4\pi t) - \mathrm{sgn}(t-.3) - \mathrm{sgn}(.77-t) & \text{for class 3.} \end{cases}$$

Figure 1 plots the true functions on $t \in [0,1]$. The functions for class 2 and class 3 have jump discontinuities at $\{.25, .72\}$ and $\{.3, .77\}$ respectively, while the function for class 1 has no jump discontinuities, but smoothly approximates the other functions.

For each class, we generated a data set as follows:

$$x_{i,l} = x(t_{i,l}) + \varepsilon_{i,l}, \quad l = 1, \ldots, 100,$$

for $i = 1, \ldots, 200$, where observational points $\{t_{i,l}\,; l = 1, \ldots, 100\}$ are independently, uni-

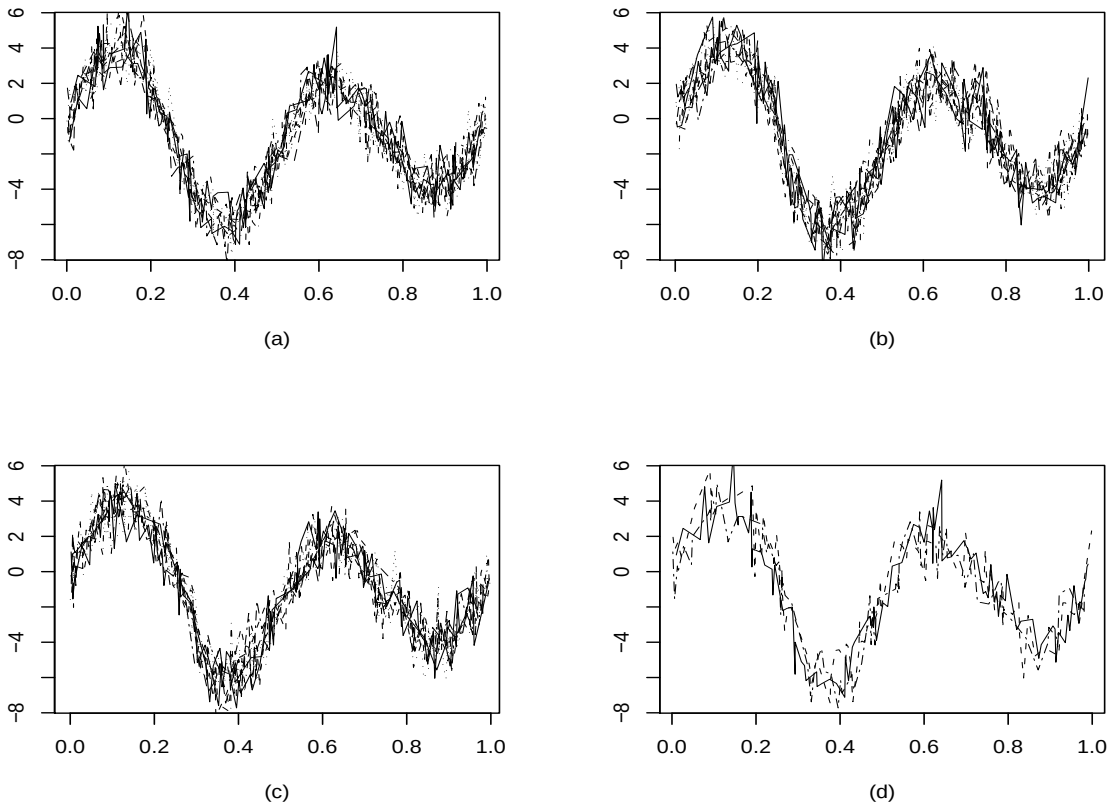

Figure 2: (a), (b) and (c) are the plots of each 10 pieces of noisy data generated for class 1, 2 and 3, respectively. (d) displays each 1 piece of noisy data for the three classes simultaneously.

formly distributed on $[0, 1]$, and noises $\{\varepsilon_{i,l}\,; l = 1, \ldots, 100\}$ are independently, normally distributed with mean 0 and variance 1. Figure 2 plots the generated data including noises. In the functionalization step, we used symmlet-5 as the wavelet bases, and used the criterion GBIC for selecting the optimum number of bases and the other smoothing parameters in the context of regularized wavelet-based regression estimation (see Fujii and Konishi (2005, Section 3.2)). For all classes, the most frequently selected number of bases was $M = 2^5$, so we fixed this parameter in the following analyses.

In each class, we randomly allocated 100 of 200 observations to a training set, and the rest 100 observations to a testing set. We estimated an optimum model according to each of the criteria, GIC and GBIC defined in Section 4, and then assessed the estimated model by calculating error rates for the testing set.

We repeated 100 times the above discriminant analysis for randomly allocated training/testing set. In total, misclassification rates for the testing set are 17.53% and 16.98%

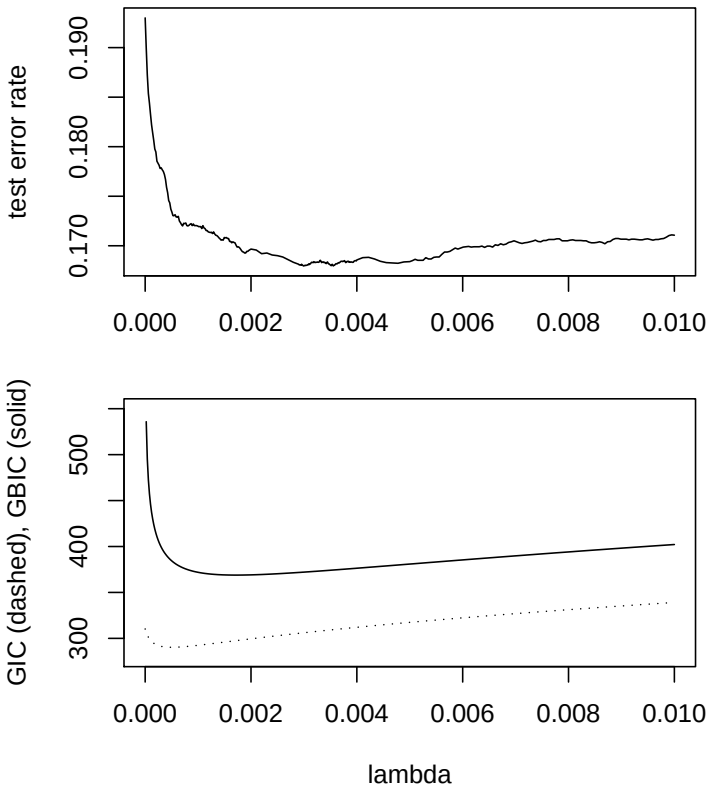

Figure 3: Upper panel: The curve of test error rate, calculated with respect to the smoothing parameter λ. Lower panel: The curves drawn by the model selection criteria (dashed line: GIC, solid line: GBIC).

for GIC and GBIC, respectively. The lower plot of Figure 3 shows the curves drawn by the values of GIC and GBIC, calculated with respect to a set of fixed smoothing parameter λ, while the upper plot shows the values of test error rate. These values are averaged over 100 repetition.

Using Fourier bases and cubic B-splines in the functionalization, we performed the same analyses as that we did for the wavelet bases. It may be noted that Fourier bases are orthonormal, while B-splines are not. Hence we calculated matrix $\mathcal{J}$ of functional linear model (3) for cubic B-splines $\{B_m(t)\,;\, m = 1, \ldots, M\}$. The (m_1, m_2)th elements $\mathcal{J}_{(m_1,m_2)} = \int B_{m_1}(t)B_{m_2}(t)\,\mathrm{d}t$ are given by $\mathcal{J}_{(m_1,m_2)}/\Delta = 214/315,\ 1163/21504,\ 1/42,\ 1/322560$ for (m_1, m_2) such that $|m_1 - m_2| = 0, 1, 2, 3$, respectively, and 0 for the other (m_1, m_2), where Δ denotes a distance of an equidistant knots sequence.

GBIC selected $M = 9$ as the optimum number of Fourier bases, and $M = 13$ for

Table 1: Discriminant results for the three types of bases each with the use of GIC and GBIC.

$n = 300$	$L_i = 100$	$\hat{y} = 1$		$\hat{y} = 2$		$\hat{y} = 3$	
$\varepsilon_{i,l} \sim \mathcal{N}(0,1)$		GIC	GBIC	GIC	GBIC	GIC	GBIC
class 1	symmlet-5	73.56	74.07	15.50	15.71	10.94	10.22
	Fourier	69.24	69.22	22.72	22.75	8.04	8.03
	cubic B-splines	72.42	72.31	18.87	19.00	8.71	8.69
class 2	symmlet-5	13.83	12.65	84.44	85.69	1.73	1.66
	Fourier	18.63	18.76	76.88	76.75	4.49	4.49
	cubic B-splines	12.74	12.79	83.10	83.03	4.16	4.18
class 3	symmlet-5	8.61	8.60	1.97	2.10	89.42	89.30
	Fourier	5.36	5.34	1.52	1.53	93.12	93.13
	cubic B-splines	8.58	8.39	1.11	1.28	90.31	90.33

cubic B-splines. For Fourier bases, test errors are 20.25% and 20.30% according to GIC and GBIC respectively, while for cubic B-splines, the test errors are 18.06% and 18.11%. Table 1 shows average breakdowns of the repeated discriminations with the use of the three types of bases.

The simulation results show efficiency of the proposed discriminant rule based on the wavelet bases. The regularization method successfully works with the use of proposed model selection criteria for the estimation of functional logistic model. It might be also said for this simulation that the criterion GBIC reduces test errors more than that of GIC.

7. Real data example

The "phoneme" data has frequently been analyzed in sound recognition. We use a dataset available at the Stanford University web-site[1], which was illustrated in the paper on penalized discriminant analysis by Hastie *et al.* (1995). The phonemes are transcribed as follows; "sh" as in "she", "dcl" as in "dark", "iy" as the vowel in "she", "aa" as the vowel in "dark", and "ao" as the first vowel in "water".

4509 speech frames are sampled from continuous speech of 50 male speakers. The each speech frame is represented by 512 digitized samples of 32 msec duration at a 16 kHz sampling rate, and it represents one of the above five phonemes. From the each speech frame, a log-periodogram of length 256 on a 0-8 kHz frequency range was computed. Thus

[1]URL: `http://www-stat.stanford.edu/~tibs/ElemStatLearn/`

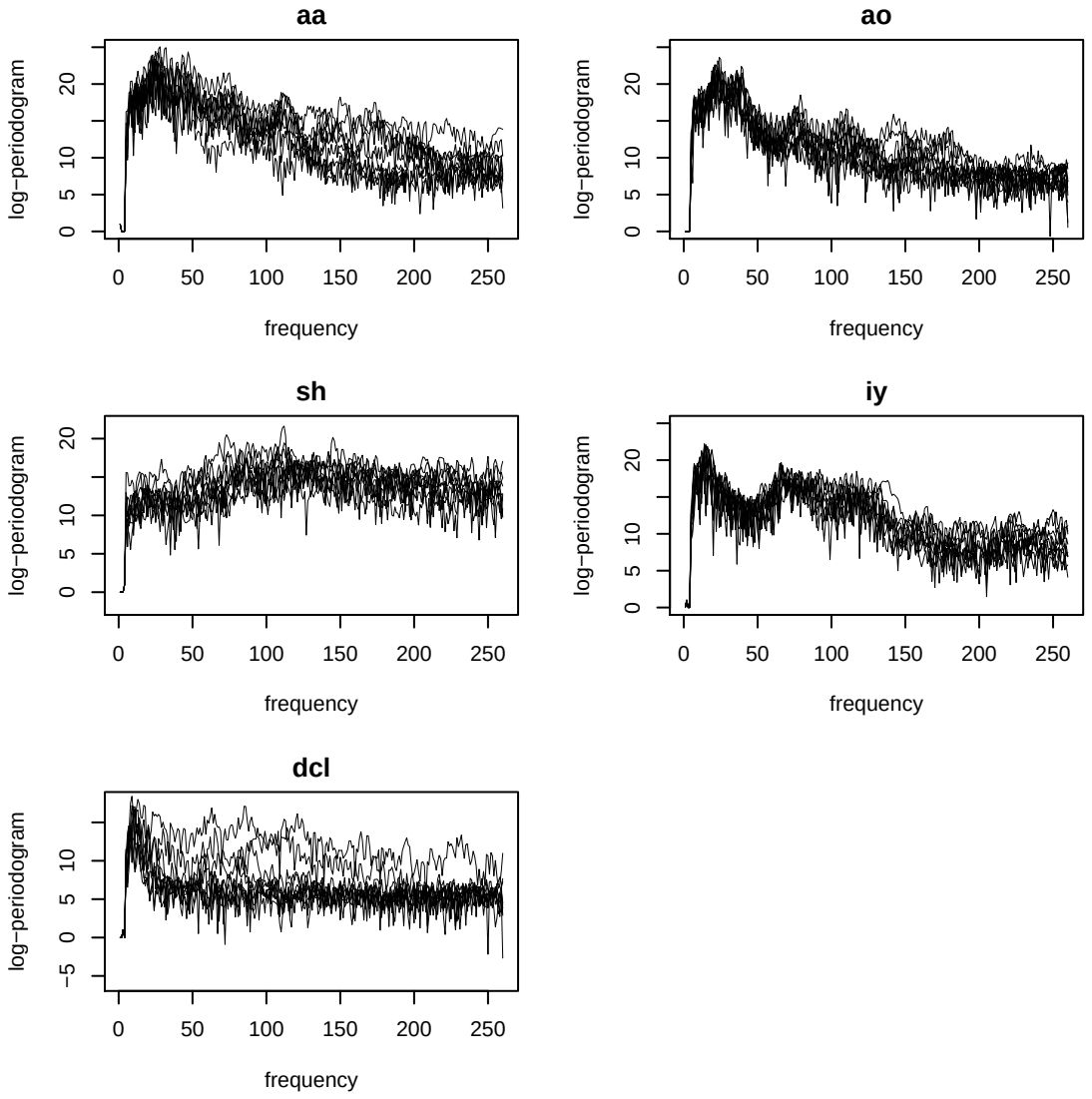

Figure 4: The log-periodograms of five phonemes (10 speech frames for each phoneme).

the dataset consists of 4509 log-periodograms of length 256 with known class (phoneme) memberships. A breakdown of the 4509 log-periodograms into phoneme frequencies is as follows; 695 "aa"s, 1022 "ao"s, 757 "dcl"s, 1163 "iy"s and 872 "sh"s. The dataset is thus represented in the form

$$[\{(x_{i,l}, t_l),\, y_i\}\,;\, l = 1, \ldots, 256\,;\, i = 1, \ldots, 4509], \tag{10}$$

where $x_{i,l}$ are the log-periodograms, t_l are the frequencies and $y_i \in \{1, 2, \ldots, 5\}$ are the class labels ("aa", "ao" "sh", "iy" or "dcl"). Figure 4 shows a sample of 10 log-periodograms of the five phonemes respectively.

In the functionalization step, we used symmlet-10 as the wavelet bases, and used the criterion GBIC for selecting the optimum number of bases and smoothing parameters in the regularized wavelet-based regression estimation. The number of bases $M = 2^6$ is optimal for almost all log-periodograms, while optimum values of the other smoothing parameters differ for each log-periodogram. Thus, we selected the smoothing parameters individually.

To perform a classification of the functionalized data, we randomly allocated 50 individuals from each class to a training set, and the rest individuals to a testing set. Thus we totally used individuals of population $n = 250$ to estimate a discriminant model, and used the rest 4259 individuals as the testing data. We then select an optimum λ by assessing the model for $\hat{\boldsymbol{\beta}}_\lambda$ via the model selection criterion. Smoothing parameter λ is selected by using either GIC or GBIC criterion. GIC selects an optimum smoothing parameter $\lambda = 0.836$, while GBIC selects $\lambda = 0.702$. The test errors are 10.07% and 10.00% for GIC and GBIC, respectively. The corresponding discrimination results are shown in Table 2.

Nextly, to make a comparison with our proposed method based on functionalization, we aimed to conduct the discriminant methods based on multivariate vectors $\boldsymbol{x}_i$ of length 256 given as (10), assuming that $\boldsymbol{x}_i$ in class k are independently, normally distributed with covariance matrix given by

$$\Sigma_k(\epsilon) = \epsilon\hat{\Sigma}_k + (1-\epsilon)\hat{\Sigma}, \qquad k = 1, 2, 3, 4, 5,$$

where $\hat{\Sigma}_k$ is a sample covariance matrix of class k and $\hat{\Sigma}$ is a sample covariance matrix of the whole training data. We then conducted Fisher's linear discriminant analysis (LDA), quadratic discriminant analysis (QDA) and regularized discriminant analysis (RDA) by taking $\epsilon = 0$, $\epsilon = 1$ and $\epsilon \in (0, 1)$, respectively. The above discriminant procedures

for multivariate observations are detailed in Hastie *et al.* (2003) and in references given therein.

However, all of the above discriminant methods for multivariate vectors failed to estimate a discriminant functions with the training set of population $n = 250$, because the variance-covariance matrices of LDA, QDA and RDA became singular.

To conduct the discriminant methods for multivariate vectors successfully, we then increased the population of training set as $n = 500$, by randomly allocating 100 individuals from each class to the training set. The method RDA gives the minimum test error 10.900% when $\epsilon = 0.01$, while LDA gives 10.975%. QDA failed once again in this situation. GIC selected an optimum smoothing parameter $\lambda = 0.7571$, while GBIC selected $\lambda = 0.5834$. The test errors are 9.204% and 9.030% for GIC and GBIC, respectively. The corresponding discrimination results of the methods other than QDA are shown in Table 3. The results show that the proposed functional discriminant procedures are superior in generalization ability to the other procedures.

Table 2: The result for phoneme data ($n = 250$). Totally, the test errors are 10.073% (GIC; $\lambda = 0.8355$) and 10.002% (GBIC; $\lambda = 0.7019$). LDA, QDA and RDA can not be calculated because of singularity.

$n = 250$		$\hat{y} = 1$	$\hat{y} = 2$	$\hat{y} = 3$	$\hat{y} = 4$	$\hat{y} = 5$	test errors (%)
aa (class 1)	GIC	493	152	0	0	0	23.57
	GBIC	491	154	0	0	0	23.88
ao (class 2)	GIC	235	735	0	0	0	24.23
	GBIC	232	738	0	0	0	23.92
sh (class 3)	GIC	0	0	822	0	0	0.00
	GBIC	0	0	822	0	0	0.00
iy (class 4)	GIC	0	2	16	1089	6	2.16
	GBIC	0	2	16	1089	6	2.16
dcl (class 5)	GIC	0	2	6	8	691	2.26
	GBIC	0	1	6	7	693	1.98

REFERENCES

Araki, Y., Konishi, S. and Imoto, S. (2004): Functional discriminant analysis for microarray gene expression data via radial basis function networks. *Proc. COMPSTAT 2004*, 613–620, Physica-Verlag/Springer.

Table 3: The result for phoneme data ($n = 500$). Totally, the test errors are 9.204% (GIC; $\lambda = 0.7571$), 9.030% (GBIC; $\lambda = 0.5834$), 10.975% for LDA and 10.900% for RDA with $\epsilon = 0.01$. QDA can not be calculated because of singularity.

$n = 500$		$\hat{y} = 1$	$\hat{y} = 2$	$\hat{y} = 3$	$\hat{y} = 4$	$\hat{y} = 5$	test errors (%)
aa (class 1)	LDA	392	203	0	0	0	34.12
	RDA	390	205	0	0	0	34.45
	GIC	441	154	0	0	0	25.88
	GBIC	438	157	0	0	0	26.39
ao (class 2)	LDA	181	740	0	0	1	19.74
	RDA	179	742	0	0	1	19.52
	GIC	171	749	0	0	2	18.76
	GBIC	164	756	0	0	2	18.00
sh (class 3)	LDA	1	0	770	0	1	0.26
	RDA	0	0	771	0	1	0.13
	GIC	0	0	772	0	0	0.00
	GBIC	0	0	772	0	0	0.00
iy (class 4)	LDA	3	0	2	1042	16	1.97
	RDA	3	0	2	1042	16	1.97
	GIC	0	2	13	1043	5	1.88
	GBIC	0	1	11	1045	6	1.69
dcl (class 5)	LDA	0	0	4	28	625	4.87
	RDA	0	0	3	27	627	4.57
	GIC	0	3	6	13	635	3.35
	GBIC	0	2	6	13	636	3.20

Cardot, H. and Sarda, P. (2005): Estimation in generalized linear models for functional data via penalized likelihood. *J. Mult. Anal.* **92**, 24–41.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977): Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39**, 1–22.

Donoho, D.L., Johnston, I.M., Kerkyacharian, G. and Picard, D. (1995): Wavelet shrinkage: Asymptopia? *J. Roy. Statist. Soc. Ser. B* **57**, 301–369.

Fujii, T. and Konishi, S. (2005): Nonlinear regression modeling via regularized wavelets and smoothing parameter selection. To appear in *J. Mult. Anal.*

Hall, P. and Patil, P. (1996): On the choice of smoothing parameter, threshold and truncation in nonparametric regression by nonlinear wavelet methods. *J. Roy. Statist. Soc. Ser. B* **58**, 361–377.

Hastie, T., Buja, A. and Tibshirani, R. (1995): Penalized discriminant analysis. *Ann. Statist.* **23**, 73–102.

Hastie, T., Tibshirani, R. and Friedman, J. (2003): *The Elements of Statistical Learning.* Springer, New York, 4th corrected printing.

James, G.M. (2002): Generalized linear models with functional predictors. *J. Roy. Statist. Soc. Ser. B* **64**, 411–432.

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

Konishi, S., Ando, T. and Imoto, S. (2004): Bayesian information criteria and smoothing parameter selection in radial basis function networks. *Biometrika* **91**(1), 27–43.

Ramsay, J.O. and Silverman, B.W. (2005): *Functional Data Analysis.* Springer, New York, 2nd edition.

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

List of MHF Preprint Series, Kyushu University

21st Century COE Program
Development of Dynamic Mathematics with High Functionality

MHF2005-1 Hideki KOSAKI
Matrix trace inequalities related to uncertainty principle

MHF2005-2 Masahisa TABATA
Discrepancy between theory and real computation on the stability of some finite element schemes

MHF2005-3 Yuko ARAKI & Sadanori KONISHI
Functional regression modeling via regularized basis expansions and model selection

MHF2005-4 Yuko ARAKI & Sadanori KONISHI
Functional discriminant analysis via regularized basis expansions

MHF2005-5 Kenji KAJIWARA, Tetsu MASUDA, Masatoshi NOUMI, Yasuhiro OHTA & Yasuhiko YAMADA
Point configurations, Cremona transformations and the elliptic difference Painlevé equations

MHF2005-6 Kenji KAJIWARA, Tetsu MASUDA, Masatoshi NOUMI, Yasuhiro OHTA & Yasuhiko YAMADA
Construction of hypergeometric solutions to the q - Painlevé equations

MHF2005-7 Hiroki MASUDA
Simple estimators for non-linear Markovian trend from sampled data:
I. ergodic cases

MHF2005-8 Hiroki MASUDA & Nakahiro YOSHIDA
Edgeworth expansion for a class of Ornstein-Uhlenbeck-based models

MHF2005-9 Masayuki UCHIDA
Approximate martingale estimating functions under small perturbations of dynamical systems

MHF2005-10 Ryo MATSUZAKI & Masayuki UCHIDA
One-step estimators for diffusion processes with small dispersion parameters from discrete observations

MHF2005-11 Junichi MATSUKUBO, Ryo MATSUZAKI & Masayuki UCHIDA
Estimation for a discretely observed small diffusion process with a linear drift

MHF2005-12 Masayuki UCHIDA & Nakahiro YOSHIDA
AIC for ergodic diffusion processes from discrete observations

MHF2005-13 Hiromichi GOTO & Kenji KAJIWARA
Generating function related to the Okamoto polynomials for the Painlevé IV equation

MHF2005-14 Masato KIMURA & Shin-ichi NAGATA
Precise asymptotic behaviour of the first eigenvalue of Sturm-Liouville problems with large drift

MHF2005-15 Daisuke TAGAMI & Masahisa TABATA
Numerical computations of a melting glass convection in the furnace

MHF2005-16 Raimundas VIDŪNAS
Normalized Leonard pairs and Askey-Wilson relations

MHF2005-17 Raimundas VIDŪNAS
Askey-Wilson relations and Leonard pairs

MHF2005-18 Kenji KAJIWARA & Atsushi MUKAIHIRA
Soliton solutions for the non-autonomous discrete-time Toda lattice equation

MHF2005-19 Yuu HARIYA
Construction of Gibbs measures for 1-dimensional continuum fields

MHF2005-20 Yuu HARIYA
Integration by parts formulae for the Wiener measure restricted to subsets in $\mathbb{R}^d$

MHF2005-21 Yuu HARIYA
A time-change approach to Kotani's extension of Yor's formula

MHF2005-22 Tadahisa FUNAKI, Yuu HARIYA & Mark YOR
Wiener integrals for centered powers of Bessel processes, I

MHF2005-23 Masahisa TABATA & Satoshi KAIZU
Finite element schemes for two-fluids flow problems

MHF2005-24 Ken-ichi MARUNO & Yasuhiro OHTA
Determinant form of dark soliton solutions of the discrete nonlinear Schrödinger equation

MHF2005-25 Alexander V. KITAEV & Raimundas VIDŪNAS
Quadratic transformations of the sixth Painlevé equation

MHF2005-26 Toru FUJII & Sadanori KONISHI
Nonlinear regression modeling via regularized wavelets and smoothing parameter selection

MHF2005-27 Shuichi INOKUCHI, Kazumasa HONDA, Hyen Yeal LEE, Tatsuro SATO, Yoshihiro MIZOGUCHI & Yasuo KAWAHARA
On reversible cellular automata with finite cell array

MHF2005-28 Toru KOMATSU
Cyclic cubic field with explicit Artin symbols

MHF2005-29 Mitsuhiro T. NAKAO, Kouji HASHIMOTO & Kaori NAGATOU
A computational approach to constructive a priori and a posteriori error estimates for finite element approximations of bi-harmonic problems

MHF2005-30 Kaori NAGATOU, Kouji HASHIMOTO & Mitsuhiro T. NAKAO
Numerical verification of stationary solutions for Navier-Stokes problems

MHF2005-31 Hidefumi KAWASAKI
A duality theorem for a three-phase partition problem

MHF2005-32 Hidefumi KAWASAKI
A duality theorem based on triangles separating three convex sets

MHF2005-33 Takeaki FUCHIKAMI & Hidefumi KAWASAKI
An explicit formula of the Shapley value for a cooperative game induced from the conjugate point

MHF2005-34 Hideki MURAKAWA
A regularization of a reaction-diffusion system approximation to the two-phase Stefan problem

MHF2006-1 Masahisa TABATA
Numerical simulation of Rayleigh-Taylor problems by an energy-stable finite element scheme

MHF2006-2 Ken-ichi MARUNO & G R W QUISPEL
Construction of integrals of higher-order mappings

MHF2006-3 Setsuo TANIGUCHI
On the Jacobi field approach to stochastic oscillatory integrals with quadratic phase function

MHF2006-4 Kouji HASHIMOTO, Kaori NAGATOU & Mitsuhiro T. NAKAO
A computational approach to constructive a priori error estimate for finite element approximations of bi-harmonic problems in nonconvex polygonal domains

MHF2006-5 Hidefumi KAWASAKI
A duality theory based on triangular cylinders separating three convex sets in R^n

MHF2006-6 Raimundas VIDŪNAS
Uniform convergence of hypergeometric series

MHF2006-7 Yuji KODAMA & Ken-ichi MARUNO
N-Soliton solutions to the DKP equation and Weyl group actions

MHF2006-8 Toru KOMATSU
Potentially generic polynomial

MHF2006-9 Toru KOMATSU
Generic sextic polynomial related to the subfield problem of a cubic polynomial

MHF2006-10 Shu TEZUKA & Anargyros PAPAGEORGIOU
Exact cubature for a class of functions of maximum effective dimension

MHF2006-11 Shu TEZUKA
On high-discrepancy sequences

MHF2006-12 Raimundas VIDŪNAS
Detecting persistent regimes in the North Atlantic Oscillation time series

MHF2006-13 Toru KOMATSU
Tamely Eisenstein field with prime power discriminant

MHF2006-14 Nalini JOSHI, Kenji KAJIWARA & Marta MAZZOCCO
Generating function associated with the Hankel determinant formula for the solutions of the Painlevé IV equation

MHF2006-15 Raimundas VIDŪNAS
Darboux evaluations of algebraic Gauss hypergeometric functions

MHF2006-16 Masato KIMURA & Isao WAKANO
New mathematical approach to the energy release rate in crack extension

MHF2006-17 Toru KOMATSU
Arithmetic of the splitting field of Alexander polynomial

MHF2006-18 Hiroki MASUDA
Likelihood estimation of stable Lévy processes from discrete data

MHF2006-19 Hiroshi KAWABI & Michael RÖCKNER
Essential self-adjointness of Dirichlet operators on a path space with Gibbs measures via an SPDE approach

MHF2006-20 Masahisa TABATA
Energy stable finite element schemes and their applications to two-fluid flow problems

MHF2006-21 Yuzuru INAHAMA & Hiroshi KAWABI
Asymptotic expansions for the Laplace approximations for Itô functionals of Brownian rough paths

MHF2006-22 Yoshiyuki KAGEI
Resolvent estimates for the linearized compressible Navier-Stokes equation in an infinite layer

MHF2006-23 Yoshiyuki KAGEI
Asymptotic behavior of the semigroup associated with the linearized compressible Navier-Stokes equation in an infinite layer

MHF2006-24 Akihiro MIKODA, Shuichi INOKUCHI, Yoshihiro MIZOGUCHI & Mitsuhiko FUJIO
The number of orbits of box-ball systems

MHF2006-25 Toru FUJII & Sadanori KONISHI
Multi-class Logistic Discrimination via Wavelet-based Functionalization and Model Selection Criteria