Nonlinear logistic discrimination via regularized Gaussian basis expansions

川野, 秀一
九州大学大学院数理学府

小西, 貞則
九州大学大学院数理学研究院

出版情報：Communications in Statistics : Simulation and Computation. 38 (7), pp.1414-1425, 2009-08. Taylor & Francis
バージョン：
権利関係：

MI Preprint Series

Kyushu University
The Grobal COE Program
Math-for-Industry Education & Research Hub

Nonlinear logistic discrimination via regularized Gaussian basis expansions

S. Kawano & S. Konishi

MI 2009-4

(Received January 23, 2009)

Faculty of Mathematics
Kyushu University
Fukuoka, JAPAN

Nonlinear logistic discrimination via regularized Gaussian basis expansions

Shuichi Kawano[1] and Sadanori Konishi[2]

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

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

s-kawano@math.kyushu-u.ac.jp konishi@math.kyushu-u.ac.jp

Abstract: We consider the problem of constructing multi-class classification methods for analyzing data with complex structure. A nonlinear logistic discriminant model is introduced based on Gaussian basis functions constructed by the self-organizing map. In order to select adjusted parameters, we employ model selection criteria derived from information-theoretic and Bayesian approaches. Numerical examples are conducted to investigate the performances of the proposed multi-class discriminant procedure. Our modeling procedure is also applied to protein structure recognition in life science. The results indicate the effectiveness of our strategy in terms of prediction accuracy.

Key Words and Phrases: Logistic regression, Model selection, Multi-class classification, Regularization, Self-organizing map.

1 Introduction

Multi-class classification problems have received an enormous amount of attention in the fields of statistics, artificial intelligence and bioinformatics (see, e.g., McLachlan, 1992; Hastie *et al.*, 2001; Bishop, 2006). One of well-known statistical tools of multi-class classification is based on linear logistic regression models (e.g., Day and Kerridge, 1967; Anderson, 1975; Seber, 1984; Hosmer and Lemeshow, 1989; Zhu and Hastie, 2004). In order to process data generated from complex structures, many researchers have studied nonlinear logistic discriminant models that are obtained by replacing a linear predictor in logistic models with a linear combination of basis functions (see, e.g., Hastie and Thibshirani, 1990; Simonoff, 1996; Loader, 1999). The natural cubic spline and B-spline have been widely used as basis functions in nonlinear logistic models. However, these modeling procedures for high dimensions may suffer from the *curse of dimensionality*.

To overcome this problem, Ando and Konishi (2008) proposed using nonlinear logistic models with Gaussian basis functions for classifying high-dimensional data.

There still remains, however, a problem in constructing Gaussian basis functions. Although Gaussian bases are usually constructed by using the k-means clustering algorithm, this algorithm depends on initial values and consequently yields different nonlinear logistic models corresponding to each set of initial values. To overcome this problem, Kawano and Konishi (2007) proposed spline-based Gaussian basis functions, which have advantages associated with B-spline bases, and demonstrated the efficiency of this nonlinear modeling method in the context of regression problems. However, this construction method is restricted to low-dimensional data (at most two dimensions). This is a major problem since data treated in classification problems generally exist in high-dimensional spaces.

In this article, we employ the self-organizing map (SOM) presented by Kohonen (1997) to construct Gaussian basis functions and propose a multi-class logistic discriminant model with these basis functions along with the technique of regularization. Our proposed models are easily applied to analyze complex or high-dimensional data, and also yield more stable prediction error rates than models based on the k-means clustering algorithm. Crucial issues in the model constructing process are the choices of adjusted parameters, which include the number of basis functions, a regularization parameter and a hyper-parameter involved in Gaussian bases. In order to select these parameters, we use information-theoretic and Bayesian type criteria. Numerical examples are conducted to examine the effectiveness of the proposed multi-class discriminant procedure. We also apply our modeling strategy to analyze protein structure data.

This article is organized as follows. Section 2 describes a nonlinear logistic model with basis expansions for multi-class classification. This section also gives estimation and evaluation procedures for models based on a penalized log-likelihood method. In Section 3, we consider the problem of constructing Gaussian basis functions, and present a nonlinear logistic discriminant model with Gaussian basis functions via the SOM. In Section 4, we demonstrate the performance of our proposed models using some numerical studies and a real data example. Some concluding remarks are given in Section 5.

2 Preliminaries

2.1 Nonlinear logistic modeling using basis expansions

Suppose we have n independent observations $\{(\boldsymbol{x}_\alpha, g_\alpha); \alpha = 1, \cdots, n\}$, where $\boldsymbol{x}_\alpha$ are p -dimensional explanatory variables and $g_\alpha \in \{1, 2, \cdots, L\}$ indicate the class label to which $\boldsymbol{x}_\alpha$ belong. We assume that the conditional probabilities given by $\boldsymbol{x}_\alpha$, which are called posterior probabilities, can be expressed as

$$\log\left\{\frac{\Pr(g_\alpha = k|\boldsymbol{x})}{\Pr(g_\alpha = L|\boldsymbol{x})}\right\} = w_{k0} + \sum_{j=1}^{m} w_{kj}\phi_j(\boldsymbol{x}) = \boldsymbol{w}_k^T\boldsymbol{\phi}(\boldsymbol{x}), \quad k = 1, \cdots, L-1, \tag{1}$$

where $\boldsymbol{w}_k = (w_{k0}, w_{k1}, \cdots, w_{km})^T$ is an unknown parameter vector for class k and $\boldsymbol{\phi}(\boldsymbol{x}) = (1, \phi_1(\boldsymbol{x}), \cdots, \phi_m(\boldsymbol{x}))^T$ is a vector of basis functions. For basis functions $\phi(\boldsymbol{x})$, we shall use Gaussian basis functions with a hyper-parameter given by

$$\phi_j(\boldsymbol{x}; \boldsymbol{\mu}_j, h_j^2, \nu) = \exp\left(-\frac{||\boldsymbol{x} - \boldsymbol{\mu}_j||^2}{2\nu h_j^2}\right), \quad (j = 1, \cdots, m), \tag{2}$$

where $\boldsymbol{\mu}_j$ is a p-dimensional vector that determines the position of the basis function, h_j^2 is the dispersion parameter and ν (> 0) is the hyper-parameter. The hyper-parameter ν plays a key role in adjusting the smoothness of the decision boundary. Ando *et al.* (2008) and Ando and Konishi (2008) have used various numerical examples to demonstrate the effectiveness of nonlinear models using the basis functions that contain the hyper-parameter. The parameters $\boldsymbol{\mu}_j$, h_j^2 are estimated using the procedures described in Section 3, while the hyper-parameter ν is determined by the model evaluation criteria given in Section 2.2.

From Equation (1) the posterior probability can be rewritten as

$$\Pr(g_\alpha = k|\boldsymbol{x}_\alpha) = \frac{\exp\{\boldsymbol{w}_k^T\boldsymbol{\phi}(\boldsymbol{x}_\alpha)\}}{1 + \sum_{j=1}^{L-1}\exp\{\boldsymbol{w}_j^T\boldsymbol{\phi}(\boldsymbol{x}_\alpha)\}}, \quad k = 1, \cdots, L-1, \tag{3}$$

$$\Pr(g_\alpha = L|\boldsymbol{x}_\alpha) = 1 - \sum_{k=1}^{L-1}\Pr(g_\alpha = k|\boldsymbol{x}_\alpha) = \frac{1}{1 + \sum_{j=1}^{L-1}\exp\{\boldsymbol{w}_j^T\boldsymbol{\phi}(\boldsymbol{x}_\alpha)\}}. \tag{4}$$

Since these posterior probabilities depend on the parameter $\boldsymbol{w} = (\boldsymbol{w}_1^T, \cdots, \boldsymbol{w}_{L-1}^T)^T$, we use the notation $\Pr(g_\alpha = k|\boldsymbol{x}) := \pi_k(\boldsymbol{x}; \boldsymbol{w})$.

For simplicity, we introduce an $(L-1)$-dimensional response variable $\boldsymbol{y} = (y_1, \cdots, y_{L-1})^T$, the components of which are either 0 or 1. The k-th element of $\boldsymbol{y}_\alpha$ is set to 1 if the corresponding $\boldsymbol{x}_\alpha$ belongs to the k-th class, i.e.,

$$\boldsymbol{y}_\alpha = (y_1^{(\alpha)}, \cdots, y_{L-1}^{(\alpha)})^T = \begin{cases} (0, \cdots, 0, \overset{(k)}{1}, 0, \cdots, 0)^T & \text{if } \ g_\alpha = k, \quad (k = 1, \cdots, L-1), \\ (0, \cdots, 0)^T & \text{if } \ g_\alpha = L. \end{cases}$$

The response $\boldsymbol{y}_\alpha$ is assumed to be distributed according to a multinomial distribution with the posterior probabilities $\pi_k(\boldsymbol{x}_\alpha; \boldsymbol{w})$ expressed by the following probability function:

$$f(\boldsymbol{y}_\alpha | \boldsymbol{x}_\alpha; \boldsymbol{w}) = \prod_{k=1}^{L-1} \pi_k(\boldsymbol{x}_\alpha; \boldsymbol{w})^{y_k^{(\alpha)}} \{\pi_L(\boldsymbol{x}_\alpha; \boldsymbol{w})\}^{1 - \sum_{l=1}^{L-1} y_l^{(\alpha)}}. \tag{5}$$

From the multinomial distribution we obtain the following log-likelihood function

$$\begin{aligned} \ell(\boldsymbol{w}) &= \sum_{\alpha=1}^{n} \log f(\boldsymbol{y}_\alpha | \boldsymbol{x}_\alpha; \boldsymbol{w}) \\ &= \sum_{\alpha=1}^{n} \left[\sum_{k=1}^{L-1} y_k^{(\alpha)} \log \pi_k(\boldsymbol{x}_\alpha; \boldsymbol{w}) + \left(1 - \sum_{l=1}^{L-1} y_l^{(\alpha)} \right) \log \pi_L(\boldsymbol{x}_\alpha; \boldsymbol{w}) \right]. \end{aligned} \tag{6}$$

The maximum likelihood estimator, which is a widely used estimator, of an unknown parameter $\boldsymbol{w}$ can easily be obtained by maximizing the log-likelihood function (6). However, the maximum likelihood method often yields an unstable parameter estimate, i.e., a parameter estimate tends to infinity. To overcome this problem, we estimate the parameter vector $\boldsymbol{w}$ by maximizing the regularized or penalized log-likelihood function

$$\ell_\lambda(\boldsymbol{w}) = \ell(\boldsymbol{w}) - \frac{n\lambda}{2} \sum_{k=1}^{L-1} \boldsymbol{w}_k^T K \boldsymbol{w}_k, \tag{7}$$

where λ (> 0) is a regularization parameter that reduces the variance of the parameter estimate and K is an $(m+1) \times (m+1)$ positive semi-definite matrix (see, e.g., Imoto and Konishi, 2003; Konishi and Kitagawa, 2008). The maximum penalized likelihood estimator $\hat{\boldsymbol{w}}$ is the solution of $\partial \ell_\lambda(\boldsymbol{w}) / \partial \boldsymbol{w} = \boldsymbol{0}$. This equation is generally nonlinear with respect to the parameter vector $\boldsymbol{w}$. We use the Fisher's scoring method to obtain the solution $\hat{\boldsymbol{w}}$ (see Green and Silverman, 1994 for details).

Given the estimate $\hat{\boldsymbol{w}}$, a future observation $\boldsymbol{x}$ is assigned to class k that has the maximum posterior probability $\pi_k(\boldsymbol{x}; \hat{\boldsymbol{w}})$ among L classes, where

$$\pi_k(\boldsymbol{x}_\alpha; \hat{\boldsymbol{w}}) = \frac{\exp\{\hat{\boldsymbol{w}}_k^T \boldsymbol{\phi}(\boldsymbol{x}_\alpha)\}}{1 + \sum_{j=1}^{L-1} \exp\{\hat{\boldsymbol{w}}_j^T \boldsymbol{\phi}(\boldsymbol{x}_\alpha)\}}, \quad k = 1, \cdots, L-1, \tag{8}$$

$$\pi_L(\boldsymbol{x}_\alpha;\hat{\boldsymbol{w}}) = \frac{1}{1+\sum_{j=1}^{L-1}\exp\{\hat{\boldsymbol{w}}_j^T\boldsymbol{\phi}(\boldsymbol{x}_\alpha)\}}. \tag{9}$$

We then obtain a statistical model in the form

$$f(\boldsymbol{y}_\alpha|\boldsymbol{x}_\alpha;\hat{\boldsymbol{w}}) = \prod_{k=1}^{L-1}\{\pi_k(\boldsymbol{x}_\alpha;\hat{\boldsymbol{w}})\}^{y_k^{(\alpha)}}\{\pi_L(\boldsymbol{x}_\alpha;\hat{\boldsymbol{w}})\}^{1-\sum_{l=1}^{L-1}y_l^{(\alpha)}}. \tag{10}$$

This statistical model depends on the number of basis functions m and the values of the regularization parameter λ and hyper-parameter ν. We use model selection criteria from information-theoretic and Bayesian approaches given in Ando and Konishi (2008) to select the values of these adjusted parameters.

2.2 Model selection criteria

The generalized information criterion proposed by Konishi and Kitagawa (1996) enables us to evaluate statistical models with various types of estimators, including the robust and penalized likelihood estimator. Using the result given in Konishi and Kitagawa (1996, p.876), we obtain a model selection criterion for evaluating the nonlinear logistic model as follows:

$$\mathrm{GIC} = -2\sum_{\alpha=1}^{n}\log f(\boldsymbol{y}_\alpha|\boldsymbol{x}_\alpha;\hat{\boldsymbol{w}}) + 2\mathrm{tr}(QR^{-1}), \tag{11}$$

where Q and R are $(L-1)(m+1)\times(L-1)(m+1)$ matrices given by

$$Q = \frac{1}{n}\{(F-G)\odot E\}^T\{(F-G)\odot E\} - \frac{\lambda}{n}I\hat{\boldsymbol{w}}\mathbf{1}_n^T\{(F-G)\odot E\}, \tag{12}$$

$$R = -\frac{1}{n}(G\odot E)^T(G\odot E) + \frac{1}{n}H + \lambda I \tag{13}$$

with $E=(\Phi,\cdots,\Phi)$, $F=(\boldsymbol{y}_{(1)}\mathbf{1}_{m+1}^T,\cdots,\boldsymbol{y}_{(L-1)}\mathbf{1}_{m+1}^T)$, $G=(\boldsymbol{\pi}_{(1)}\mathbf{1}_{m+1}^T,\cdots,\boldsymbol{\pi}_{(L-1)}\mathbf{1}_{m+1}^T)$, $H=\mathrm{diag}\{\Phi^T\mathrm{diag}\{\boldsymbol{\pi}_{(1)}\}\Phi,\cdots,\Phi^T\mathrm{diag}\{\boldsymbol{\pi}_{(L-1)}\}\Phi\}$, $I=\mathrm{diag}\{K,\cdots,K\}$, $\boldsymbol{y}_{(k)}=(y_k^{(1)},\cdots,y_k^{(n)})^T$, $\Phi=(\boldsymbol{\phi}(\boldsymbol{x}_1),\cdots,\boldsymbol{\phi}(\boldsymbol{x}_n))^T$ and $\boldsymbol{\pi}_{(k)}=(\pi_k(\boldsymbol{x}_1;\hat{\boldsymbol{w}}),\cdots,\pi_k(\boldsymbol{x}_n;\hat{\boldsymbol{w}}))^T$. Here the operator $\odot$ denotes the Hadamard product.

Konishi *et al.* (2004) proposed the generalized Bayesian information criteria by extending the Bayesian information criteria (Schwarz, 1978) to evaluate statistical models estimated by the maximum penalized likelihood method. Using the result given in Konishi

et al. (2004, p.30), we obtain a model evaluation criterion given by

$$\begin{aligned}\text{GBIC} = -2\sum_{\alpha=1}^{n}\log f(\boldsymbol{y}_\alpha|\boldsymbol{x}_\alpha;\hat{\boldsymbol{w}}) + n\lambda\sum_{k=1}^{L-1}\hat{\boldsymbol{w}}_k^T K\hat{\boldsymbol{w}}_k - (L-1)\log|K|_+ \\ + \log|R| - (L-1)(m+1-d)\log\lambda - (L-1)d\,\log\left(\frac{2\pi}{n}\right),\end{aligned} \tag{14}$$

where $|K|_+$ is the product of the positive eigenvalues of K with rank d and R is given in Equation (13).

We select the number of basis functions and the values of the regularization parameter and the hyper-parameter by minimizing either the generalized information criterion (GIC) or the generalized Bayesian information criterion (GBIC).

3 Construction of Gaussian basis functions

3.1 Gaussian basis functions and its problem

In this section, we consider the problem of constructing Gaussian basis functions. The centers $\boldsymbol{\mu}_j$ and width parameters h_j^2 included in Gaussian basis functions given in Equation (2) are generally determined by using the k-means clustering algorithm (Moody and Darken, 1989). This algorithm divides a set of observations $\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n\}$ into m clusters $\{C_1,\cdots,C_m\}$ corresponding to the number of basis functions. The centers $\boldsymbol{\mu}_j$ and the width parameters h_j^2 are then respectively determined by $\hat{\boldsymbol{\mu}}_j = \sum_{\alpha\in C_j}\boldsymbol{x}_\alpha/n_j$ and $\hat{h}_j^2 = \sum_{\alpha\in C_j}||\boldsymbol{x}_\alpha - \hat{\boldsymbol{\mu}}_j||^2/n_j$, where n_j is the number of observations that belongs to the j-th cluster C_j. Replacing $\boldsymbol{\mu}_j$ with $\hat{\boldsymbol{\mu}}_j$ and h_j^2 with $\hat{h}_j^2$, we obtain a set of m basis functions given by

$$\phi_j(\boldsymbol{x};\hat{\boldsymbol{\mu}}_j,\hat{h}_j^2,\nu) = \exp\left(-\frac{||\boldsymbol{x}-\hat{\boldsymbol{\mu}}_j||^2}{2\nu\hat{h}_j^2}\right), \quad j=1,\cdots,m. \tag{15}$$

It should, however, be noted that models based on the basis functions constructed using the k-means clustering algorithm have some drawbacks. These drawbacks are due to the different initial values in the k-means clustering algorithm, which imply that clusters determined using this clustering algorithm are strongly dependent on the initial values. To illustrate this point, data $\{(x_{1\alpha},x_{2\alpha},g_\alpha);\alpha=1,\cdots,300\}$ are generated from normal

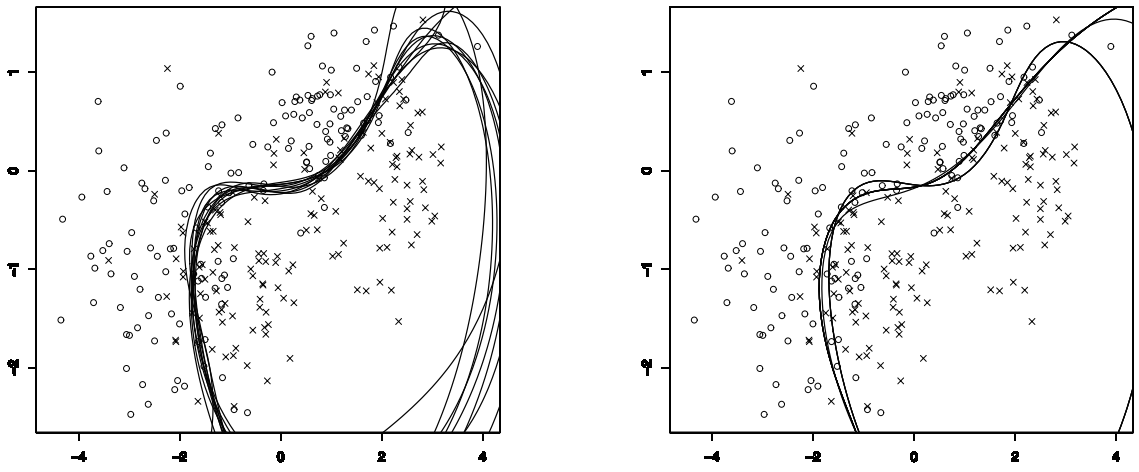

Figure 1: 10 estimated decision boundaries for the same simulated data set. The left panel shows boundaries estimated by models constructed using the k-means clustering algorithm. The right panel shows the result obtained using our modeling method. Class 1 samples are indicated by circles and class 2 samples are denoted by crosses.

mixture distributions as follows:

$$\begin{aligned}
\text{Class 1} \quad & 0.5\, N\left(\begin{pmatrix} 1 \\ 0.5 \end{pmatrix}, \begin{pmatrix} 1 & 0.2 \\ 0.2 & 0.2 \end{pmatrix}\right) + 0.5\, N\left(\begin{pmatrix} -2.5 \\ -1 \end{pmatrix}, \begin{pmatrix} 0.75 & 0 \\ 0 & 0.75 \end{pmatrix}\right), \\
\text{Class 2} \quad & 0.5\, N\left(\begin{pmatrix} -1 \\ -1 \end{pmatrix}, \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}\right) + 0.5\, N\left(\begin{pmatrix} 2 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}\right).
\end{aligned} \tag{16}$$

The left panel in Figure 1 shows 10 estimated decision boundaries for the same data set. We observe that the models provide a different set of decision boundaries, which demonstrates that the k-means clustering algorithm estimates different centers of basis functions for different sets of initial values.

This is undesirable for practical applications since the same estimation model produces different results; different decision boundaries were obtained in spite of using exactly the same data set. To overcome this problem, Kawano and Konishi (2007) proposed new Gaussian bases, which have the advantages associated with B-spline basis functions. They investigated the efficiency of nonlinear regression models based on these new Gaussian basis functions. However, spline-based construction methods cannot be directly applied to high-dimensional data sets since the number of basis functions may exceed the sample

size. In this study, therefore, we employ Gaussian basis functions based on the SOM.

3.2 Gaussian basis functions using the self-organizing map

The SOM is an unsupervised neural network, which visualizes complex high-dimensional data by drawing a low-dimensional map (Kohonen, 1997). When the SOM is used as a clustering method it constructs more stable clusters than the k-means algorithm: the SOM is robust against variation in initial values, whereas the k-means algorithm is strongly dependent on the initial values.

The SOM algorithm is given by the following procedure. First, a set of reference vectors $\{\boldsymbol{p}_j \in \mathbf{R}^p; j = 1, \cdots, m\}$ is prepared. Second, we select the node of the reference vector $\boldsymbol{p}_c$ that minimizes the distance between the reference vector and observations of the explanatory variables $\boldsymbol{x}_\alpha$, i.e.,

$$c = \arg\min_j \{||\boldsymbol{x}_\alpha - \boldsymbol{p}_j||\}. \tag{17}$$

Third, if we have the t-th values of $\boldsymbol{p}_c^{(t)}$, the $(t+1)$-th updated values $\boldsymbol{p}_c^{(t+1)}$ are given by

$$\boldsymbol{p}_j^{(t+1)} = \boldsymbol{p}_j^{(t)} + h_c(t)\left[\boldsymbol{x}_\alpha - \boldsymbol{p}_j^{(t)}\right], \quad (j = 1, \cdots, m), \tag{18}$$

where $h_c(t)$ is a monotone decreasing function of the number of iterations t. A Gaussian neighborhood kernel is generally used as the monotone decreasing function in the following equation:

$$h_c(t) = \alpha(t)\ \exp\left\{-\frac{||\boldsymbol{p}_c - \boldsymbol{p}_j^{(t)}||^2}{2\sigma^2(t)}\right\}, \tag{19}$$

where $\alpha(t)$ (> 0) is the learning rate and $\sigma^2(t)$ determines the width of the function. Both $\alpha(t)$ and $\sigma^2(t)$ are monotone decreasing functions of the numbers of iterations, and they could be selected to be linear. Fourth, we alternate between the second step and the third step for all observations $\{\boldsymbol{x}_\alpha; \alpha = 1, \cdots, n\}$. Finally, we continually repeat the second step to the forth step until the convergence condition is satisfied. The resulting clusters C_j $(j = 1, \cdots, m)$ are given by $\{\boldsymbol{x}_\alpha; j = \arg\min_k\{||\boldsymbol{x}_\alpha - \boldsymbol{p}_k||\}\}$, where $\boldsymbol{p}_k$ are the convergence values of the above procedure. For more details, we refer the reader to Kohonen (1997).

We obtain m Gaussian basis functions given by

$$\phi_j^{(som)}(\boldsymbol{x}; \hat{\boldsymbol{\mu}}_j^{(som)}, \hat{h}_j^{(som)2}, \nu) = \exp\left(-\frac{||\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j^{(som)}||^2}{2\nu \hat{h}_j^{(som)2}}\right), \quad j = 1, \cdots, m, \tag{20}$$

where $\hat{\boldsymbol{\mu}}_j^{(som)}$ and $\hat{h}_j^{(som)2}$ are respectively the estimated centers and width parameters obtained from the clusters $C_j^{(som)}$ $(j = 1, \cdots, m)$ using the SOM. We then introduce a following nonlinear logistic model using these basis functions,

$$\log\left\{\frac{\Pr(g_\alpha = k|\boldsymbol{x}_\alpha)}{\Pr(g_\alpha = L|\boldsymbol{x}_\alpha)}\right\} = w_{k0} + \sum_{j=1}^{m} w_{kj}\phi_j^{(som)}(\boldsymbol{x}_\alpha; \hat{\boldsymbol{\mu}}_j^{(som)}, \hat{h}_j^{(som)2}, \nu), \quad k = 1, \cdots, L-1.$$

Our proposed models are more stable than models whose basis functions are constructed using the k-means algorithm, since the basis functions included in our models are constructed using the SOM. To demonstrate this, we fitted our model to the same data set given in (16). The right panel of Figure 1 shows 10 estimated decision boundaries constructed by our models. The decision boundaries depicted in the right panel of Figure 1 are more stable than those in the left panel, demonstrating that our proposed models are superior to models using the k-means algorithm in the sense that they give a smaller variance in the estimated decision boundaries. In the next section, we use some numerical examples to compare our models with several other models in terms of prediction accuracy and stability.

4 Numerical examples

4.1 Synthetic data

We investigate the performance of our modeling methodology by analyzing synthetic data (Ripley, 1996). This synthetic data consists of two classes with two-dimensional explanatory variables; 250 values of training data and 1000 values of test data were prepared.

Nonlinear logistic models with basis functions constructed using the SOM or the k-means clustering method are fitted to the data set. The number of basis functions was fixed to 25, and the values of the regularization parameter and hyper-parameter in the

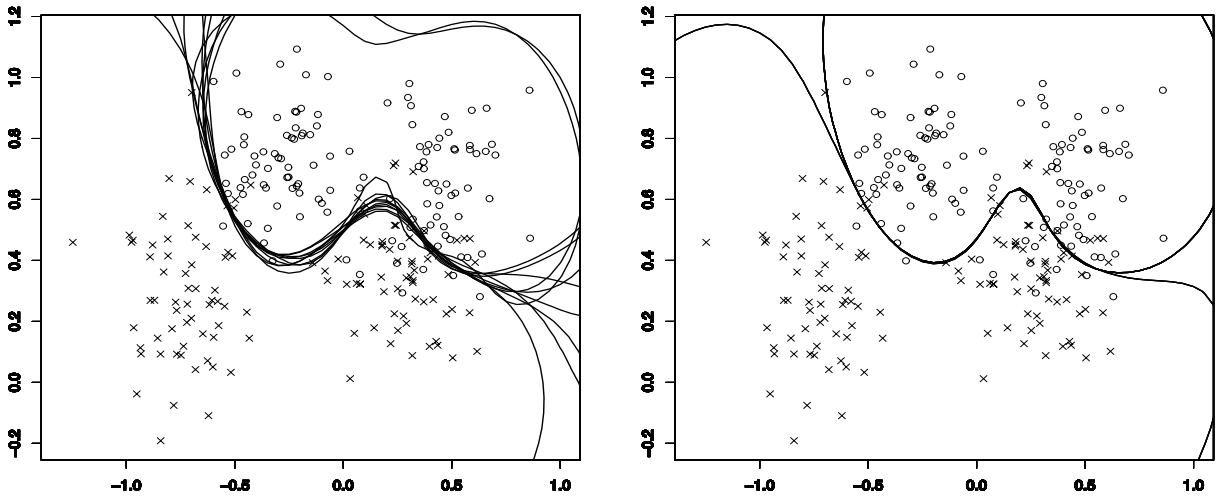

Figure 2: 10 estimated decision boundaries for the same data set. The left panel shows boundaries estimated by models constructed using the k-means clustering algorithm. The right panel shows the results obtained using our modeling method. Class 1 samples are indicated by circles and Class 2 samples are denoted by crosses.

Gaussian basis function were selected using the GIC. We then repeated the above procedure 10 times using exactly the same data set. Figure 2 shows that the models estimated by the SOM are more stable than those constructed using the k-means clustering algorithm. This result is similar to that given in Section 3. We also compared the performances of nonlinear logistic modeling procedures based on Gaussian bases using the SOM (NLMsom) and k-means clustering (NLMk) with that of other procedures. The regularization parameter and the hyper-parameter in the NLMsom and NLMk models were selected by using the GIC or GBIC, while the number of basis functions was fixed as 25 since this numerical study required a considerable amount of computation. Table 1 shows a summary of the prediction errors. To obtain results for generalized additive model (GAM), we used the `mgcv` function from the `mgcv` package in `R` (Wood, 2004). For the NLMsom or NLMk models, the prediction errors listed in Table 1 are the averages of 50 results obtained using the same data set, and the figures in parentheses indicate the standard errors. Table 1 shows that the proposed methods perform well; they yield a relatively small prediction error with a small variance.

Table 1: Prediction errors (%) for synthetic data. Figures in parentheses indicate the standard errors for 50 repetitions. The results for methods 9 to 12 are from Ripley (1994).

	Method	Prediction error
1.	NLMsom with GIC	9.40 (0)
2.	NLMsom with GBIC	9.50 (0)
3.	NLMk with GIC	9.47 (0.38)
4.	NLMk with GBIC	9.51 (0.16)
5.	Linear discriminant analysis	10.8
6.	Quadratic discriminant analysis	10.2
7.	Linear logistic discriminant model	11.4
8.	GAM	9.50
9.	CART	10.1
10.	1 nearest neighbor	15.0
11.	2 nearest neighbor	13.4
12.	3 nearest neighbor	13.0

4.2 Wave form data

In the second experiment, we consider the problem of multi-class classification by analyzing waveform data (Hastie *et al.*, 2001). The waveform data consist of three classes with 21-dimensional predictors, and were generated from the following functions:

$$x_k = \begin{cases} uH_1(k) + (1-u)H_2(k) + \varepsilon_k & \text{if } g = 1 \\ uH_1(k) + (1-u)H_3(k) + \varepsilon_k & \text{if } g = 2 \\ uH_2(k) + (1-u)H_3(k) + \varepsilon_k & \text{if } g = 3 \end{cases} \qquad k = 1, \cdots, 21. \tag{21}$$

where u is uniform on [0,1], ε_k are the standard normal variates and H_i are the shifted triangular waveforms, $H_1(k) = \max\{6-|k-11|, 0\}$, $H_2(k) = H_1(k-4)$, $H_3(k) = H_1(k+4)$. We generated 300 sets of training data with equal prior probability for each class and 500 sets of test data.

We compared the performances of several different modeling procedures. Using the GIC or GBIC as the model selection criterion, we selected the regularization parameter and the hyper-parameter in the NLMsom or NLMk models. As in Section 4.1, we fixed the number of basis functions as 15, since the computational demanding was quite high. A summary of the prediction errors is given in Table 2. We obtained the results for the

Table 2: Prediction errors (%) for waveform data. Figures in parentheses indicate the standard errors for 50 repetitions. The results for methods 9 to 11 are from Hastie *et al.* (2001).

	Method	Prediction error
1.	NLMsom with GIC	15.6 (0.10)
2.	NLMsom with GBIC	15.7 (0.10)
3.	NLMk with GIC	16.0 (0.70)
4.	NLMk with GBIC	16.1 (0.74)
5.	Linear discriminant analysis	21.2
6.	Quadratic discriminant analysis	21.6
7.	Linear logistic discriminant model	21.2
8.	GAM	22.0
9.	Classification tree	28.9
10.	FDA (MARS (degree 1))	19.1
11.	FDA (MARS (degree 2))	21.5

generalized additive model (GAM) using the `mgcv` function from the `mgcv` package in `R`. The prediction errors for the NLMsom and NLMk models were obtained by averaging the results obtained by 50 iterations from the same data set, and the standard errors are given in parentheses. For this data set, our proposed models using the GIC or GBIC give lower prediction errors than other models and provide more stable prediction error rates than the models based on the k-means clustering algorithm.

4.3 Recognition of protein structure data

We applied the proposed multi-class discriminant procedure to a protein structure data set that was analyzed in Ding and Dubchak (2001). This data set consists of four structural classes: all-α, all-β, α/β, $\alpha+\beta$. For each class, the percentage compositions of the 20 amino acids form a part of the predictors. The remainder of the predictors is defined by the structural or physicochemical properties extracted from the primary protein sequence, allowing us to generate 125-dimensional predictors. See Dubchak *et al.* (1995, 1999) for details regarding this method of generating the predictors. The data set can be obtained from the website (*http://ranger.uta.edu/∼chqding/protein*).

Table 3: The number of training data sets and test data sets for each class.

Class	Training data	Test data
all-α	55	61
all-β	109	117
α/β	115	145
$\alpha+\beta$	34	62
Total	313	385

Table 4: Prediction errors (%) for protein data. The results for SVM and KNN are from Shi and Suganthan (2003).

	Method	Prediction error
1.	NLMsom with GIC	23.3
2.	NLMsom with GBIC	22.8
3.	LDA	NA
4.	QDA	NA
5.	LLDA	23.9
6.	SVM	23.1
7.	KNN	28.9

The number of training sets and test sets for each class in this study are listed in Table 3. Our modeling method was applied to the data set with the help of regularization. The number of basis functions and the values of the regularization parameter and hyperparameter were selected using the GIC or GBIC. The values of the adjusted parameters for the data set are $m = 28$, $\lambda = 10^{-8.0}$ and $\nu = 11.5$ for the GIC, while they are $m = 30$, $\lambda = 10^{-4.4}$ and $\nu = 5.0$ for the GBIC. We compared the performance of our procedure with that of linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), linear logistic discriminant analysis (LLDA), support vector machine (SVM) and K-nearest neighbor classifier (KNN). Table 4 summarizes the prediction errors obtained using these methods. Discriminant functions could not be constructed from LDA and QDA because of their singular variance-covariance matrices. Table 4 shows that the nonlinear logistic discriminant models based on the SOM provide relatively lower prediction errors than other methods.

5 Concluding remarks

In this article, we introduced a nonlinear logistic model based on Gaussian basis functions constructed by using the SOM in the framework of multi-class classification problem. In order to choose the values of adjusted parameters, we employ model selection criteria from the information-theoretic and Bayesian viewpoints. Some numerical examples and a real data analysis demonstrated that our modeling strategies yield smaller prediction error rates than several previously developed models. Due to the stability and the predictive performance of the estimated models, our multi-class logistic discrimination procedure has the potential to be useful in a variety of practical applications.

References

Anderson, J.A. (1975). Quadratic logistic discrimination. *Biometrika*, **62**, 149–154.

Ando, T. and Konishi, S. (2008). Nonlinear logistic discrimination via regularized radial basis functions for classifying high-dimensional data. *Annals of the Institute of Statistical Mathematics* (to appear).

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

Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York.

Day, N.E. and Kerridge, D.F. (1967). A general maximum likelihood discriminant. *Biometrics*, **23**, 313–324.

Ding, C.H. and Dubchak, I. (2001). Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics*, **17**, 349–358.

Dubchak, I., Muchnik, I., Holbrook, S.R. and Kim, S.H. (1995). Prediction of protein folding class using global description of amino acid sequence. *Proceedings of the National Academy of Sciences*, **92**, 8700–8704.

Dubchak, I., Muchnik, I., Mayor, C., Dralyuk, I. and Kim, S.H. (1999). Recognition of a protein fold in the context of the structural classification of proteins (SCOP)

classification. *Proteins*, **35**, 401–407.

Green, P.J. and Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models*. Chapman & Hall, London.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman & Hall, London.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer, New York.

Hosmer, D.W. and Lemeshow, S. (1989). *Applied Logistic Regression*. Wiley, New York.

Imoto, S. and Konishi, S. (2003). Selection of smoothing parameters in B-spline nonparametric regression models using information criteria. *Annals of the Institute of Statistical Mathematics*, **55**, 671–687.

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

Kohonen, T. (1997). *Self-Organizing Maps*. Springer, New York.

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*. Springer, New York.

Loader, C. (1999). *Local Regression and Likelihood*. Springer, New York.

McLachlan, G.J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, New York.

Moody, J. and Darken, C.J. (1989). Fast learning in networks of locally-tuned processing units. *Neural Computation*, **1**, 281–294.

Ripley, B.D. (1994). Neural networks and related methods for classification. *Journal of the Royal Statistical Society Series B*, **56**, 409–456.

Ripley, B.D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press, UK.

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

Seber, G.A.F. (1984). *Multivariate Observations*. Wiley, New York.

Shi, S.Y.M. and Suganthan, P.N. (2003). Feature analysis and classification of protein secondary structure data. *Lecture Notes in Computer Science*, **2714**, 1151–1158.

Simonoff, J.S. (1996). *Smoothing Methods in Statistics*. Springer, New York.

Wood, S.N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, **99**, 673–686.

Zhu, J. and Hastie, T. (2004). Classification of gene microarrays by penalized logistic regression. *Biostatistics*, **5**, 427–443.

List of MI Preprint Series, Kyushu University

The Grobal COE Program
Math-for-Industry Education & Research Hub

MI

MI2008-1 Takahiro ITO, Shuichi INOKUCHI & Yoshihiro MIZOGUCHI
Abstract collision systems simulated by cellular automata

MI2008-2 Eiji ONODERA
The intial value problem for a third-order dispersive flow into compact almost Hermitian manifolds

MI2008-3 Hiroaki KIDO
On isosceles sets in the 4-dimensional Euclidean space

MI2008-4 Hirofumi NOTSU
Numerical computations of cavity flow problems by a pressure stabilized characteristic-curve nite element scheme

MI2008-5 Yoshiyasu OZEKI
Torsion points of abelian varieties with values in n nite extensions over a p-adic eld

MI2008-6 Yoshiyuki TOMIYAMA
Lifting Galois representations over arbitrary number elds

MI2008-7 Takehiro HIROTSU & Setsuo TANIGUCHI
The random walk model revisited

MI2008-8 Silvia GANDY, Masaaki KANNO, Hirokazu ANAI & Kazuhiro YOKOYAMA
Optimizing a particular real root of a polynomial by a special cylindrical algebraic decomposition

MI2008-9 Kazufumi KIMOTO, Sho MATSUMOTO & Masato WAKAYAMA
Alpha-determinant cyclic modules and Jacobi polynomials

MI2008-10 Sangyeol LEE & Hiroki MASUDA
Jarque-Bera Normality Test for the Driving Lévy Process of a Discretely Observed Univariate SDE

MI2008-11 Hiroyuki CHIHARA & Eiji ONODERA
A third order dispersive flow for closed curves into almost Hermitian manifolds

MI2008-12 Takehiko KINOSHITA, Kouji HASHIMOTO and Mitsuhiro T. NAKAO
On the L^2 a priori error estimates to the nite element solution of elliptic problems with singular adjoint operator

MI2008-13 Jacques FARAUT and Masato WAKAYAMA
Hermitian symmetric spaces of tube type and multivariate Meixner-Pollaczek polynomials

MI2008-14 Takashi NAKAMURA
Riemann zeta-values, Euler polynomials and the best constant of Sobolev inequality

MI2008-15 Takashi NAKAMURA
Some topics related to Hurwitz-Lerch zeta functions

MI2009-1 Yasuhide FUKUMOTO
Global time evolution of viscous vortex rings

MI2009-2 Hidetoshi MATSUI & Sadanori KONISHI
Regularized functional regression modeling for functional response and predictors

MI2009-3 Hidetoshi MATSUI & Sadanori KONISHI
Variable selection for functional regression model via the L_1 regularization

MI2009-4 Shuichi KAWANO & Sadanori KONISHI
Nonlinear logistic discrimination via regularized Gaussian basis expansions