Ann Inst Stat Math (2009) 61:331–353
DOI 10.1007/s10463-007-0143-3

Nonlinear logistic discrimination via regularized radial basis functions for classifying high-dimensional data

Tomohiro Ando · Sadanori Konishi

Received: 8 August 2005 / Revised: 16 April 2007 / Published online: 10 August 2007

Abstract A flexible nonparametric method is proposed for classifying high-dimensional data with a complex structure. The proposed method can be regarded as an extended version of linear logistic discriminant procedures, in which the linear predictor is replaced by a radial-basis-expansion predictor. Radial basis functions with a hyperparameter are used to take the information on covariates and class labels into account; this was nearly impossible within the previously proposed hybrid learning framework. The penalized maximum likelihood estimation procedure is employed to obtain stable parameter estimates. A crucial issue in the model-construction process is the choice of a suitable model from candidates. This issue is examined from information-theoretic and Bayesian viewpoints and we employed Ando et al. (*Japanese Journal of Applied Statistics*, *31*, 123–139, 2002)'s model evaluation criteria. The proposed method is available not only for the high-dimensional data but also for the variable selection problem. Real data analysis and Monte Carlo experiments show that our proposed method performs well in classifying future observations in practical situations. The simulation results also show that the use of the hyperparameter in the basis functions improves the prediction performance.

Keywords Bayes approach · Information criteria · Maximum penalized likelihood method · Radial basis functions

T. Ando (✉)
Graduate School of Business Administration, Keio University, 2-1-1 Hiyoshi-Honcho, Kohoku-ku, Yokohama-shi, Kanagawa 223-8523, Japan
e-mail: andoh@kbs.keio.ac.jp

S. Konishi
Faculty of Mathematics, Kyushu University, 6-10-1 Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan
e-mail: konishi@math.kyushu-u.ac.jp

1 Introduction

The methods of multiclass classification provide one of the most important tools in various fields of research, including finance, economics, engineering, artificial intelligence, and bioinformatics. One well-known statistical method of multiclass classification is based on linear logistic discriminant models (Seber 1984; Hosmer and Lemeshow 1989), which assume that the log-odds ratios of the posterior probabilities can be expressed as linear combinations of the p-dimensional feature variables $\boldsymbol{x} = (x_1, \ldots, x_p)^T$:

$$\log\left\{\frac{\Pr(g = k|\boldsymbol{x})}{\Pr(g = G|\boldsymbol{x})}\right\} = w_{k0} + \sum_{j=1}^{p} w_{kj}x_j, \quad k = 1, \ldots, G-1. \tag{1}$$

Here G is the number of groups, the categorical variable $g \in \{1, \ldots, G\}$ is an indicator of the class label, and $\Pr(g = k|\boldsymbol{x})$ is the posterior probability of $g = k$ given the feature variables $\boldsymbol{x}$. When the unknown parameters $\{w_{kj};\ j = 0, \ldots, p, k = 1, \ldots, G-1\}$ are estimated by the maximum likelihood method, a future observation is generally classified into one of several groups that gives the maximum posterior probability.

Although linear logistic discriminant models have become a standard tool for multiclass classification, this method has some disadvantages. Firstly, linear decision boundaries are often too crude for complex data, and therefore nonlinear decision boundaries would be more attractive (Hastie et al. 1994). Secondly, a large number of predictors relative to the sample size leads to unstable maximum likelihood parameter estimates. In addition, the existence of multicollinearity may result in infinite maximum likelihood parameter estimates and, consequently, incorrect classification results.

To overcome these problems, we can replace the linear predictor in (1) with a linear combination of radial basis functions (Bishop 1995; Ripley 1996; Webb 1999). Unfortunately, a problem still remains in the construction of the radial basis functions. The previously proposed hybrid learning methods (Broomhead and Lowe 1988; Moody and Darken 1989; Ranganath and Arun 1997) construct radial basis functions in a completely unsupervised way and do not take class label information into account. Ando et al. (2002) therefore used radial basis functions with a hyperparameter (Ando et al. 2001, 2005; Konishi et al. 2004). This method can easily be applied to high-dimensional data and also the variable selection problem. Moreover, a clear improvement is obtained by the use of the hyperparameter in the radial basis functions.

The unknown parameters are estimated by the penalized maximum likelihood method, or the regularization method (Green and Silverman 1994; Eilers and Marx 1996), since the maximum likelihood method does not yield satisfactory results for fitting our model to high-dimensional data with a complex structure. The essential points in the model-building process are the determination of the number of basis functions and of the values of the smoothing parameter and hyperparameter. This problem can be investigated from an information-theoretic (Akaike 1973, 1974) and also a Bayesian (Schwarz 1978) point of view. Ando et al. (2002) derived tailor-made versions of the generalized information criterion (GIC; Konishi and Kitagawa 1996) and the Bayesian information criterion (BIC; Konishi et al. 2004) for evaluating the

goodness of the radial basis function network classification model, estimated by the penalized maximum likelihood method. The estimated models are evaluated by using these criteria.

The contributions of the paper are as follows: firstly, since the computational difficulties caused by high dimensionality were not focused by Ando et al. (2002), the paper focuses on the high-dimensional data. Secondly, when each feature takes only 0, 1 values, the direct application of k-means algorithm is not suitable in the basis function construction step, because it works well for continuous data. To solve this problem, we have introduced a new idea in Sect. 5.3. Thirdly, the variable selection problem was not focused by Ando et al. (2002). We therefore considered the variable selection problem.

This article is organized as follows. In Sect. 2, we extend linear logistic discriminant models to nonlinear models by replacing the linear predictor with a linear combination of radial basis functions. Section 3 describes the model estimation procedure and illustrates some characteristics of our modeling procedure. In Sect. 4, we describe two types of model evaluation criteria. Section 5 conduct real data analysis and Monte Carlo simulations performed to investigate the performance of our proposed procedure. The numerical results indicate that the proposed method performs well in practical situations even when the dimensionality of the feature variables is large. Conclusions are given in Sect. 6.

2 Radial basis functions for logistic regression models

One of the assumptions made in logistic discrimination is that the log-odds of the posterior probabilities can be expressed as a linear combination of the p-dimensional feature variables. However, when high-dimensional data with a complex structure are analyzed, a suitable decision boundary that separates the data into several different groups will often be nonlinear. We have therefore extended the class of linear logistic discriminant models to nonlinear models by replacing the linear predictor with a linear combination of radial basis functions:

$$\log\left\{\frac{\Pr(g=k|\boldsymbol{x})}{\Pr(g=G|\boldsymbol{x})}\right\} = w_{k0} + \sum_{j=1}^{m} w_{kj}\phi_j(\boldsymbol{x}), \tag{2}$$

where $\{\phi_j(\boldsymbol{x}); j = 1, \ldots, m\}$ are a set of radial basis functions and $\{w_{kj}; j = 0, \ldots, m, k = 1, \ldots, G-1\}$ are a set of unknown parameters to be estimated. For a perspective on radial basis functions, see Girosi et al. (1995), Bishop (1995), Ripley (1996), Webb (1999), and the references given therein.

For radial basis functions $\phi_j(\boldsymbol{x})$ in (2), Ando et al. (2002) used a Gaussian radial basis with a hyperparameter (Ando et al. 2001):

$$\phi_j(\boldsymbol{x}) = \exp\left(-\frac{||\boldsymbol{x}-\boldsymbol{\mu}_j||^2}{2\nu\sigma_j^2}\right), \quad j = 1, 2, \ldots, m, \tag{3}$$

where $\boldsymbol{\mu}_j$ is a p-dimensional vector determining the location of the basis function, σ_j^2 is the scale parameter, and ν is the hyperparameter. This basis function avoids several problems that occur in the previously proposed methods (see for example Ando et al. 2005); these problems are described in Sect. 3.3. As will be shown in Sect. 3.4, the hyperparameter ν plays an essential role in controlling the smoothness of decision boundaries.

It may easily be seen that log-posterior-odds models of the form (2) can be rewritten in terms of the following posterior probabilities:

$$\begin{aligned} \Pr(g = k|\boldsymbol{x}) &= \frac{\exp\left\{\boldsymbol{w}_k^T \boldsymbol{\phi}(\boldsymbol{x})\right\}}{1 + \sum_{j=1}^{G-1} \exp\left\{\boldsymbol{w}_j^T \boldsymbol{\phi}(\boldsymbol{x})\right\}}, \quad k = 1, \ldots, G-1, \\ \Pr(g = G|\boldsymbol{x}) &= \frac{1}{1 + \sum_{k=1}^{G-1} \exp\left\{\boldsymbol{w}_j^T \boldsymbol{\phi}(\boldsymbol{x})\right\}}, \end{aligned} \tag{4}$$

where $\boldsymbol{w}_k = (w_{k0}, \ldots, w_{km})^T$ is an $(m+1)$-dimensional parameter vector and $\boldsymbol{\phi}(\boldsymbol{x}) = (1, \phi_1(\boldsymbol{x}), \ldots, \phi_m(\boldsymbol{x}))^T$ is an $(m+1)$-dimensional vector of basis functions. These posterior probabilities $\Pr(g = k|\boldsymbol{x})$ depend on a set of parameters $\boldsymbol{w} = (\boldsymbol{w}_1^T, \ldots, \boldsymbol{w}_{G-1}^T)^T$, and so we denote these posterior probabilities as $\Pr(g = k|\boldsymbol{x}) := \pi_k(\boldsymbol{x}; \boldsymbol{w})$.

We now define the G-dimensional vector $\boldsymbol{y} = (y_1, \ldots, y_G)^T$ that indicates group membership. The kth element of $\boldsymbol{y}$ is set to be one or zero according to whether $\boldsymbol{x}$ belongs or does not belong to the kth group as follows:

$$\boldsymbol{y} = (y_1, \ldots, y_G)^T = (0, \ldots, \overset{(k-1)}{0}, \overset{(k)}{1}, \overset{(k+1)}{0}, \ldots, 0)^T \quad \text{if } g = k.$$

This implies that $\boldsymbol{y}$ is the kth unit column vector if $g = k$.

Assuming that the random variable $\boldsymbol{y}$ is distributed according to a multinomial distribution with probabilities $\pi_k(\boldsymbol{x}; \boldsymbol{w})(k = 1, \ldots, G)$, our model (2) can be expressed in the following probability density form:

$$f(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{w}) = \prod_{k=1}^{G} \pi_k(\boldsymbol{x}; \boldsymbol{w})^{y_k}, \tag{5}$$

where $\pi_k(\boldsymbol{x}; \boldsymbol{w})$ are the posterior probabilities given in (4).

The problems are how to construct the Gaussian radial basis functions (3) and how to estimate the unknown parameter $\boldsymbol{w}$ in the model, which will be discussed in the next section.

3 Estimation of model parameters

Suppose that we have a set of n independent observations $\{(\boldsymbol{x}_\alpha, g_\alpha); \alpha = 1, \ldots, n\}$, where the $\boldsymbol{x}_\alpha$ are the vectors of p feature variables and g_α are the class labels. Our model estimation procedure consists of two stages. In the first stage, a set of Gaussian radial

basis functions $\{\phi_j(\boldsymbol{x});\ j = 1, \ldots, m\}$ are constructed or, equivalently, the centers $\boldsymbol{\mu}_j$ and the scale parameters σ_j^2 in the Gaussian radial basis (3) are determined. In the second stage, the unknown parameter vector $\boldsymbol{w}$ is estimated by the penalized maximum likelihood method.

3.1 Construction of the Gaussian radial basis

Ando et al. (2002) determined the centers $\boldsymbol{\mu}_j$ and the scale parameters σ_j^2 in the Gaussian radial basis by using the k-means clustering algorithm (MacQueen 1967). This algorithm divides a set of observations $\{\boldsymbol{x}_\alpha;\ \alpha = 1, \ldots, n\}$ into m clusters $A_1, \ldots, A_m$ that correspond to the number of basis functions. The centers and the scale parameters are then determined by $\boldsymbol{\mu}_j = \sum_{\alpha \in A_j} \boldsymbol{x}_\alpha / n_j$ and $\sigma_j^2 = \sum_{\alpha \in A_j} ||\boldsymbol{x}_\alpha - \boldsymbol{c}_j||^2 / n_j$, respectively, where n_j is the number of observations which belong to the jth cluster A_j. Using an appropriate value of the hyperparameter ν, we then obtain a set of m Gaussian radial basis functions. Various basis construction procedures have been proposed, which will be described in Sect. 3.3.

As will be shown in Sect. 3.4, the hyperparameter ν in the Gaussian radial basis function plays an important role in determining the smoothness of the decision boundaries. We can optimize the value of the hyperparameter by using model selection criteria, which will be discussed in Sect. 4.

3.2 Penalized maximum likelihood estimation

The vector of unknown parameters $\boldsymbol{w}$ is estimated by maximizing the penalized log-likelihood function

$$\ell_\lambda(\boldsymbol{w}) = \sum_{\alpha=1}^{n} \log f(\boldsymbol{y}_\alpha | \boldsymbol{x}_\alpha; \boldsymbol{w}) - \frac{n\lambda}{2} \boldsymbol{w}^T \boldsymbol{w}, \tag{6}$$

where $\boldsymbol{y}_\alpha = (y_1^{(\alpha)}, \ldots, y_G^{(\alpha)})^T$ indicates the class label of the αth observation, and λ is the smoothing parameter. For details of the penalized maximum likelihood method, we refer to Green and Silverman (1994), Eilers and Marx (1996), and references given therein.

The penalized maximum likelihood estimates $\hat{\boldsymbol{w}}$ are given by the solution of $\partial \ell_\lambda(\boldsymbol{w}) / \partial \boldsymbol{w} = \boldsymbol{0}$, which is obtained by employing a Newton–Raphson algorithm. Using the first and second derivatives of $\ell_\lambda(\boldsymbol{w})$, given by

$$\frac{\partial \ell_\lambda(\boldsymbol{w})}{\partial \boldsymbol{w}_k} = \sum_{\alpha=1}^{n} \left\{ y_k^{(\alpha)} - \pi_k(\boldsymbol{x}_\alpha; \boldsymbol{w}) \right\} \boldsymbol{\phi}(\boldsymbol{x}_\alpha) - n\lambda \boldsymbol{w}_k, \quad k = 1, \ldots, G-1,$$

$$\frac{\partial \ell_\lambda(\boldsymbol{w})}{\partial \boldsymbol{w}_m \partial \boldsymbol{w}_l^T} = \begin{cases} \displaystyle\sum_{\alpha=1}^{n} \pi_m(\boldsymbol{x}_\alpha; \boldsymbol{w})(1 - \pi_m(\boldsymbol{x}_\alpha; \boldsymbol{w})) \boldsymbol{\phi}(\boldsymbol{x}_\alpha) \boldsymbol{\phi}(\boldsymbol{x}_\alpha)^T - n\lambda I_{m+1}, & (l = m), \\ \displaystyle\sum_{\alpha=1}^{n} \pi_m(\boldsymbol{x}_\alpha; \boldsymbol{w}) \pi_l(\boldsymbol{x}_\alpha; \boldsymbol{w}) \boldsymbol{\phi}(\boldsymbol{x}_\alpha) \boldsymbol{\phi}(\boldsymbol{x}_\alpha)^T, & (l \neq m), \end{cases}$$

respectively, we optimize the parameter vector $\boldsymbol{w}$ by use of the following iterative system:

$$\boldsymbol{w}^{\text{new}} = \boldsymbol{w}^{\text{old}} - \left\{ \frac{\partial^2 \ell_\lambda(\boldsymbol{w}^{\text{old}})}{\partial \boldsymbol{w} \partial \boldsymbol{w}^T} \right\}^{-1} \frac{\partial \ell_\lambda(\boldsymbol{w}^{\text{old}})}{\partial \boldsymbol{w}},$$

where I_{m+1} is an $(m+1) \times (m+1)$ identity matrix. The parameter vector $\boldsymbol{w}$ is updated until a suitable convergence criterion is satisfied.

3.3 Some remarks and previous studies

Generally, a Gaussian radial basis function is given by

$$\phi_j(\boldsymbol{x}) = \exp\left(-\frac{||\boldsymbol{x} - \boldsymbol{\mu}_j||^2}{2\sigma_j^2} \right), \quad j = 1, 2, \ldots, m. \tag{7}$$

The radial basis functions $\phi_j(\boldsymbol{x})$ overlap with each other to capture the information from the input data. Since the amount of overlap of the basis functions is controlled by the width parameters, the values of those width parameters play an essential role in determining the smoothness of the decision boundary.

Unfortunately, the previously proposed hybrid learning methods determine the width parameters by using a completely unsupervised approach (Bishop 1995; Broomhead and Lowe 1988; Karayiannis and Mi 1997; Moody and Darken 1989; Sato 1996; Ranganath and Arun 1997). Such heuristic approaches do not always give sufficiently good prediction results (Ando et al. 2001, 2005; Konishi et al. 2004). To construct a more flexible and data-adaptive learning procedure, we have introduced the Gaussian basis (3). In Sect. 5.2, using Monte Carlo simulations, we compare the performance of our basis construction approach with that of other methods and show the effectiveness of our approach.

In low-dimensional (one- or two-dimensional) cases, one approach for determining the centers is to use a uniform grid (Bishop 1995; Nabney 2002). Unfortunately, in high-dimensional cases, it is nearly impossible to use this approach, since the number of basis functions m may become much larger than the sample size n, which we call overparameterization. Girosi et al. (1995) used the data points $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ for the centers in their numerical studies. However, this approach also causes an overparameterization problem.

Another approach to constructing a Gaussian radial basis is to use fully supervised learning that simultaneously optimizes the Gaussian radial basis $\phi_j(\boldsymbol{x})$ and the parameter vector $\boldsymbol{w}$ by maximizing the penalized log-likelihood function (6). Xu et al. (1995) and Xu (1998) implemented the maximum likelihood method by using the well-known EM algorithm. However, there are a number of disadvantages. Convergence to a global minimum cannot be guaranteed, since the problem is nonlinear with respect to the centers and widths of the basis functions. In fact, Moody and Darken (1989) reported that fully supervised learning does not guarantee that the basis functions will be localized well in numerical simulations. In addition, an overparameterization

problem frequently arises when we regard the centers and the scale parameters as unknown parameters to be optimized. Consider the situation, for example, when the number of groups, the sample size, and the dimension of $\boldsymbol{x}$ are $G = 3$, $n = 100$, and $p = 10$, respectively. In this case, even if we use only a set of $m = 5$ basis functions, an overparameterization problem occurs. Furthermore, the computational time for the fully supervised learning method is much larger than that for hybrid learning methods. We therefore employed the hybrid learning approach.

3.4 Some characteristics of the proposed model

The purpose of this subsection is to illustrate some characteristics of the proposed model by means of a simulation study. We show that (a) the smoothness of the decision boundary is mainly controlled by ν, and (b) the smoothing parameter λ has the effect of reducing the variances of the parameter estimates $\hat{\boldsymbol{w}}$ or, equivalently, it controls the stability of the decision boundary.

A set of simulated data $\{(x_{1\alpha}, x_{2\alpha}, g_\alpha), \alpha = 1, \ldots, 100\}$ were generated from equal mixtures of normal distributions with centers $(0.3, -0.7)$ and $(0.3, 0.3)$ in class 1 and $(0.7, 0.2)$ and $(0.7, 0.3)$ in class 2, with a common covariance matrix $\Sigma = 0.03 I_2$, where I_2 is a two-dimensional identity matrix. Figure 1 shows the true decision boundary obtained from the Bayes rule. As shown in Fig. 1, the Bayes decision boundary $\{\boldsymbol{x};\ P(g = 1|\boldsymbol{x}) = P(g = 2|\boldsymbol{x}) = 0.5\}$ represents a nonlinear structure. It is clear that the linear logistic discriminant model (1) cannot capture the true structure well.

We investigate first the effect of the smoothing parameter. The proposed model was estimated with various values of the smoothing parameter λ. In this experiment, the values of the smoothing parameter were specified as $\log_{10}(\lambda) = -1, -3, -5$, and -7, respectively. We set $m = 20$ and $\nu = 10$. Figure 2 shows the estimated decision boundaries $\{\boldsymbol{x}; \pi_1(\boldsymbol{x}; \hat{\boldsymbol{w}}) = \pi_2(\boldsymbol{x}; \hat{\boldsymbol{w}}) = 0.5\}$ obtained from 50 Monte Carlo simulations.

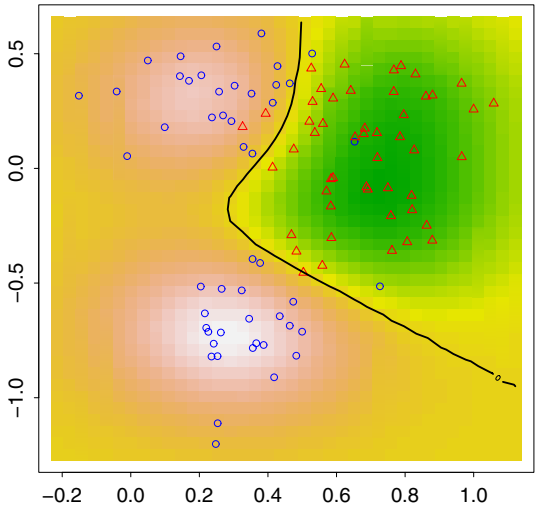


Fig. 1 The Bayes boundary (*solid line*). Samples are marked by *open circles* ($g_\alpha = 1$) and *open triangles* ($g_\alpha = 2$). As the posterior probability $P(g = 2|\boldsymbol{x})$ becomes larger, the color becomes green

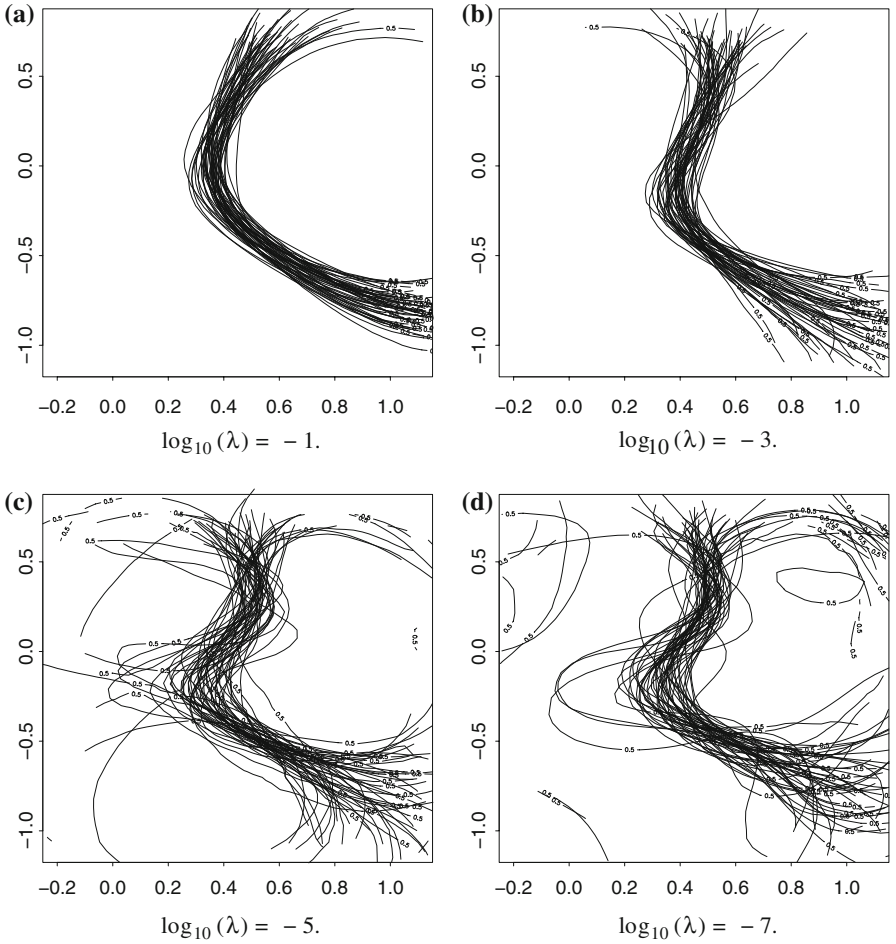


Fig. 2 Comparison of the variances of the estimated decision boundaries obtained through 50 Monte Carlo simulations

It can be seen from Fig. 2 that the stability of our model is closely related to the value of the smoothing parameter; as the value of the smoothing parameter becomes smaller, the variance of the estimated decision boundary becomes large. The variance of the decision boundary can be reduced by using a relatively large smoothing parameter. However, too large a smoothing parameter leads to a linear decision boundary, which cannot capture the nonlinear structure well.

Boxplots of the training errors and prediction errors obtained from 50 Monte Carlo simulations are also shown in Fig. 3. As the smoothing parameter becomes smaller, the training error becomes small. Note that we cannot use the training error as a measure of the prediction ability of the estimated model, since we can make the training error small by using a more complicated model. In fact, the smallest value of the

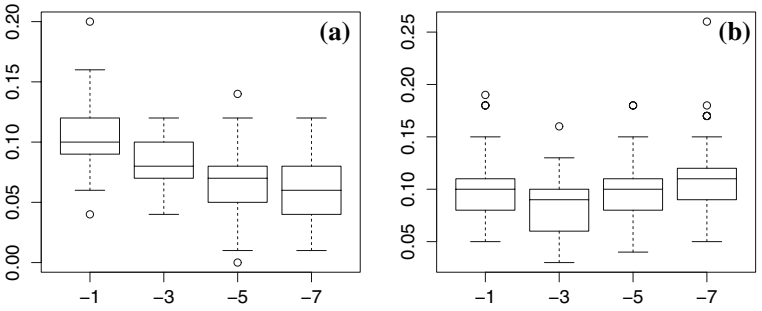


Fig. 3 Boxplots of **a** the training errors and **b** the prediction errors obtained from various values of the smoothing parameter $\log_{10}(\lambda)$

smoothing parameter $\log_{10}(\lambda) = -7$ gives the smallest median value of the training error, whereas it does not minimize the median value of the prediction error. On the other hand, an appropriate choice of $\log_{10}(\lambda) = -3$ gives the smallest median value of the prediction error.

We next illustrate the effect of the hyperparameter ν in the Gaussian radial basis function (3). Using the penalized maximum likelihood method, we fitted the proposed model (5) with $\log_{10}(\nu) = 0$, 1, and 2, respectively. In this simulation, we fixed the number of basis functions and the value of the smoothing parameter at $m = 20$ and $\log_{10}(\lambda) = -3$. Figure 4 compares the Bayes decision boundary and the estimated decision boundaries. The estimated decision boundaries in Fig. 4a and c are obviously undersmoothed and oversmoothed, respectively. We can see from Fig. 4b that an appropriate choice of ν gives a good approximation to the system underlying the data.

These simulation studies indicate that the crucial issue in the model building process is the choice of λ and ν. Additionally, the number of basis functions m should be optimized. In the next section, we present two types of model selection criterion, derived from information-theoretic and Bayesian viewpoints.

4 Model selection

Perhaps the most standard approach to selecting the adjusted parameters λ, ν, and m would be the minimization of a cross-validated misclassification rate. Unfortunately, owing to the use of a nonlinear optimization algorithm, the computational cost becomes larger as the sample size n increases. To overcome this problem, we have constructed two analytical model evaluation criteria; they are closely related to the misclassification rate tested on future observations (unseen data). Hereafter, the estimator $\hat{\boldsymbol{w}}$ is the maximizer of the penalized log-likelihood function (6).

4.1 An information-theoretic approach

Akaike (1973, 1974) proposed an information criterion, AIC, as an estimator of the Kullback–Leibler information (Kullback and Leibler 1951) from a predictive point of

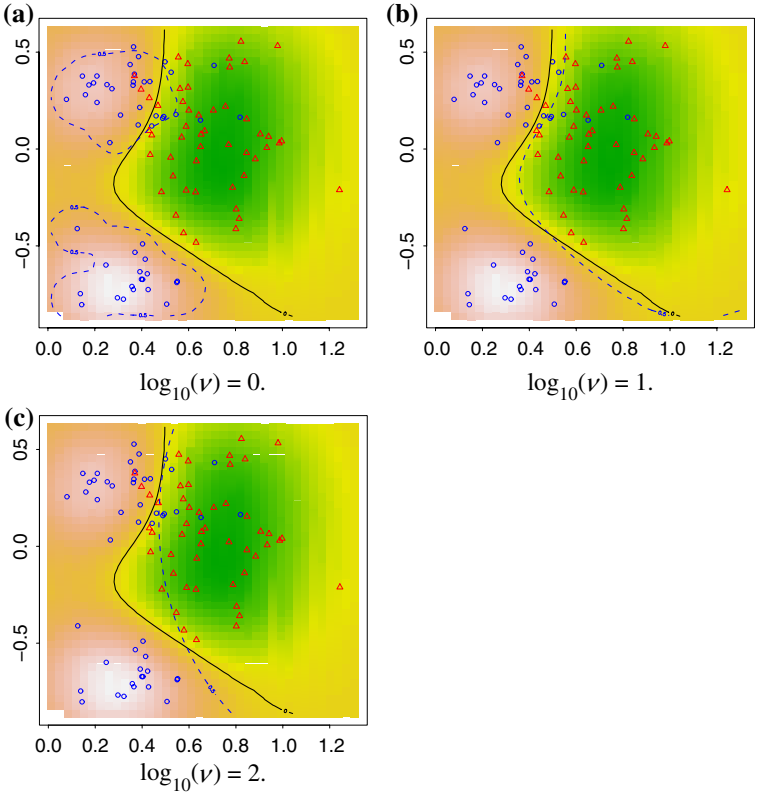


Fig. 4 The effect of hyperparameter ν. The *dashed lines* and *solid lines* represent the estimated decision boundaries and Bayes decision boundary, respectively

view. However, AIC theoretically covers only the models estimated by the maximum likelihood method. If the models were constructed by the penalized maximum likelihood method, a problem might arise in the theoretical justification for the automatic use of AIC. Ando et al. (2002) therefore presented an information criterion for evaluating the proposed model, estimated by use of the penalized maximum likelihood method within the framework of nonlinear discriminant models.

An information criterion is generally constructed by correcting the bias of the log-likelihood in the estimation of the expected log-likelihood

$$\ell^*(\hat{\boldsymbol{w}}) = \sum_{\alpha=1}^{n} \sum_{k=1}^{G} \int z_k \ \log \ \pi_k(\boldsymbol{x}_\alpha; \hat{\boldsymbol{w}}) \mathrm{d}H(\boldsymbol{z}_\alpha | \boldsymbol{x}_\alpha),$$

where the expectation value is taken over the true distribution $H(\boldsymbol{z}|\boldsymbol{x})$. If the true distribution is replaced by the empirical distribution function, it follows from (5) that the log-likelihood of the model is

$$\ell(\hat{\boldsymbol{w}}) = \sum_{\alpha=1}^{n} \log f(\boldsymbol{y}_\alpha|\boldsymbol{x}_\alpha; \hat{\boldsymbol{w}}) = \sum_{\alpha=1}^{n}\sum_{k=1}^{G} y_k^{(\alpha)} \log \pi_k(\boldsymbol{x}_\alpha; \hat{\boldsymbol{w}}). \tag{8}$$

The bias of the log-likelihood in estimating the expected log-likelihood is then given by

$$b(H) = \sum_{\alpha=1}^{n} \int \left\{\ell(\hat{\boldsymbol{w}}) - \ell^*(\hat{\boldsymbol{w}})\right\} \mathrm{d}H(\boldsymbol{y}_\alpha|\boldsymbol{x}_\alpha).$$

If the bias can be estimated by appropriate procedures and $\hat{b}$ is obtained, the bias-corrected log-likelihood is given by $\ell(\hat{\boldsymbol{w}}) - \hat{b}$, which is usually used in the form $\mathrm{IC} = -2\{\ell(\hat{\boldsymbol{w}}) - \hat{b}\}$.

We use Theorem 2.1 in Konishi and Kitagawa (1996, p. 876), which states that the bias of the log-likelihood (8) in estimating the expected log-likelihood is asymptotically given by

$$b(H) = \mathrm{tr}\left\{R(H)^{-1}Q(H)\right\} + o(1).$$

Here $R(H)$ and $Q(H)$ are $(G-1)(m+1)$-dimensional matrices, given respectively by

$$R(H) = -\int \left.\frac{\partial^2\{\log f(z|\boldsymbol{x}; \boldsymbol{w}) - \lambda \boldsymbol{w}^T\boldsymbol{w}/2\}}{\partial \boldsymbol{w}\partial \boldsymbol{w}^T}\right|_{\boldsymbol{w}=\boldsymbol{T}(H)} \mathrm{d}H(z|\boldsymbol{x}),$$

$$Q(H) = \int \left.\frac{\partial\{\log f(z|\boldsymbol{x}; \boldsymbol{w}) - \lambda \boldsymbol{w}^T\boldsymbol{w}/2\}}{\partial \boldsymbol{w}} \frac{\partial \log f(z|\boldsymbol{x}; \boldsymbol{w})}{\partial \boldsymbol{w}^T}\right|_{\boldsymbol{w}=\boldsymbol{T}(H)} \mathrm{d}H(z|\boldsymbol{x}),$$

and $\boldsymbol{T}(H)$ is the statistical functional defined by

$$\int \frac{\partial}{\partial \boldsymbol{w}}\left(\log f(z|\boldsymbol{x}; \boldsymbol{w}) - \frac{\lambda}{2}\boldsymbol{w}^T\boldsymbol{w}\right)\Big|_{\boldsymbol{w}=\boldsymbol{T}(H)} \mathrm{d}H(z|\boldsymbol{x}) = \boldsymbol{0}.$$

Note that the estimator $\hat{\boldsymbol{w}}$ can be obtained in this equation by replacing H by the empirical distribution $\hat{H}$, that is, $\hat{\boldsymbol{w}} = \boldsymbol{T}(\hat{H})$.

Replacing the unknown distribution H by the empirical distribution $\hat{H}$, we obtain a tailor-made version of the generalized information criterion (Konishi and Kitagawa 1996) for evaluating the proposed model $f(\boldsymbol{y}|\boldsymbol{x}; \hat{\boldsymbol{w}})$ estimated by the penalized maximum likelihood method, as follows:

$$\mathrm{GIC} = -2\sum_{\alpha=1}^{n}\sum_{k=1}^{G} y_k^{(\alpha)} \log \pi_k(\boldsymbol{x}_\alpha; \hat{\boldsymbol{w}}) + 2\mathrm{tr}\left\{R(\hat{H})^{-1}Q(\hat{H})\right\}, \tag{9}$$

where $R(\hat{H})$ and $Q(\hat{H})$ are given respectively by

$$R(\hat{H}) = -\frac{1}{n}(C \otimes A)^T (C \otimes A) + \frac{1}{n} D + \lambda I_{(G-1)(m+1)},$$
$$Q(\hat{H}) = \frac{1}{n}((B - C) \otimes A)^T ((B - C) \otimes A) - \frac{\lambda}{n} \hat{\boldsymbol{w}} \mathbf{1}_n^T ((B - C) \otimes A);$$

$A = \overbrace{(\Phi, \ldots, \Phi)}^{G-1}$, $B = (\boldsymbol{y}_{(1)} \mathbf{1}_{m+1}^T, \ldots, \boldsymbol{y}_{(G-1)} \mathbf{1}_{m+1}^T)$, $C = (\boldsymbol{\pi}_{(1)} \mathbf{1}_{m+1}^T, \ldots, \boldsymbol{\pi}_{(G-1)} \times \mathbf{1}_{m+1}^T)$, $D = \mathrm{diag}\{\Phi^T \mathrm{diag}\{\boldsymbol{\pi}_{(1)}\}\Phi, \ldots, \Phi^T \mathrm{diag}\{\boldsymbol{\pi}_{(G-1)}\}\Phi\}$, $\Phi = (\boldsymbol{\phi}(\boldsymbol{x}_1), \ldots, \boldsymbol{\phi}(\boldsymbol{x}_n))^T$, $\boldsymbol{y}_{(k)} = (y_k^{(1)}, \ldots, y_k^{(n)})^T$, and $\boldsymbol{\pi}_{(k)} = (\pi_k(\boldsymbol{x}_1; \hat{\boldsymbol{w}}), \ldots, \pi_k(\boldsymbol{x}_n; \hat{\boldsymbol{w}}))^T$. Here the operator $\otimes$ means the elementwise product (suppose that the arbitrary matrices $A_{ij} = (a_{ij})$, $B_{ij} = (b_{ij})$ are given; then $A_{ij} \otimes B_{ij} = (a_{ij} \times b_{ij})$).

We choose the optimum values of the smoothing parameter λ, the hyperparameter ν, and the number of basis functions m which minimize the value of the information criterion GIC in (9).

Ando et al. (2001) derived tailor-made versions of GIC for the evaluation of the radial basis function network Gaussian regression models. GIC for evaluating the radial basis function network and B-spline generalized linear models were presented by Ando et al. (2005) and Imoto and Konishi (2003), respectively. Fujii and Konishi (2006) and Nonaka and Konishi (2005) obtained GIC for the evaluation of wavelets and local likelihood regression models estimated by the method of regularization, respectively.

4.2 Bayesian approach

Schwarz (1978) proposed a Bayesian information criterion, BIC. As in the case of AIC, Schwarz's BIC covers only models estimated by the maximum likelihood method (Konishi et al. 2004). Thus, the problem of constructing a Bayesian information criterion for evaluating models estimated by the penalized maximum likelihood method still remains.

Suppose we are interested in selecting a model from a set of candidate models $M_1, \ldots, M_r$ for a given set of n observations $D_n = \{(\boldsymbol{x}_\alpha, g_\alpha); \alpha = 1, \ldots, n\}$. In the proposed model (5), the differences of each model M_k are characterized by the combination of the number of basis functions m, the values of the smoothing parameter λ, and the hyperparameter ν.

The Bayes approach to selecting a model is to choose the model with the largest posterior probability from among a set of candidate models:

$$P(M_j|D_n) \propto P(M_j) \int \prod_{\alpha=1}^{n} f(\boldsymbol{y}_\alpha|\boldsymbol{x}_\alpha; \boldsymbol{w}) \pi(\boldsymbol{w}) \mathrm{d}\boldsymbol{w}, \quad j = 1, \ldots, r, \tag{10}$$

where $\pi(\boldsymbol{w})$ and $P(M_j)$ are the prior distribution of $\boldsymbol{w}$ and the prior probability for model M_j, respectively. A crucial problem in constructing a criterion based on the

posterior probability of the model is the computation of the high-dimensional integral (10).

Under certain regularity conditions, Konishi et al. (2004) used the Laplace approximation (Tierney and Kadane 1986; Tierney et al. 1989; Kass et al. 1990) to compute this high-dimensional integral and obtained

$$\int \prod_{\alpha=1}^{n} f(\boldsymbol{y}_\alpha|\boldsymbol{x}_\alpha;\boldsymbol{w})\pi(\boldsymbol{w})\mathrm{d}\boldsymbol{w} = \frac{(2\pi)^{q/2}}{n^{q/2}|U(\hat{\boldsymbol{w}})|^{1/2}} \exp\left\{n \cdot u(\hat{\boldsymbol{w}}, D_n)\right\}\left\{1 + O_p(n^{-1})\right\},$$

where q is the dimension of $\boldsymbol{w}$, $\log \pi(\boldsymbol{w}) = O(n)$,

$$u(\boldsymbol{w}, D_n) = \frac{1}{n}\log\left\{\prod_{\alpha=1}^{n} f(\boldsymbol{y}_\alpha|\boldsymbol{x}_\alpha;\boldsymbol{w})\pi(\boldsymbol{w})\right\} \quad \text{and} \quad U(\boldsymbol{w}) = -\frac{\partial^2 u(\boldsymbol{w}, D_n)}{\partial \boldsymbol{w}\partial \boldsymbol{w}^T},$$

and $\hat{\boldsymbol{w}}$ is the mode of $u(\boldsymbol{w}, D_n)$. Taking the logarithm of the resulting formula, Konishi et al. (2004) extended Schwarz's BIC to cover the evaluation of models estimated by the penalized maximum likelihood method.

Concerning the penalized maximum likelihood method, we implicitly specify the prior distributions $\pi(\boldsymbol{w})$ of the parameters of each model to be a $(G-1)(m+1)$-variate normal distribution $\pi(\boldsymbol{w}) = (n\lambda)^{-p/2}(2\pi)^{p/2}\exp\{-n\lambda \boldsymbol{w}^T\boldsymbol{w}/2\}$. Substituting the prior distribution $\pi(\boldsymbol{w})$ into $u(\boldsymbol{w}, D_n)$ and taking the first derivative of $u(\boldsymbol{w}, D_n)$, we find that the estimator $\hat{\boldsymbol{w}}$ is given by the solution of the following equation:

$$\begin{aligned}\frac{\partial u(\boldsymbol{w}, D_n)}{\partial \boldsymbol{w}} &= \frac{\partial}{\partial \boldsymbol{w}}\left\{\sum_{\alpha=1}^{n} f(\boldsymbol{y}_\alpha|\boldsymbol{x}_\alpha;\boldsymbol{w}) - \frac{n\lambda}{2}\boldsymbol{w}'\boldsymbol{w} + \frac{p}{2}\log\left(\frac{n\lambda}{2\pi}\right)\right\} \\ &= \frac{\partial \ell_\lambda(\boldsymbol{w})}{\partial \boldsymbol{w}} = \boldsymbol{0}. \end{aligned} \tag{11}$$

This implies that $\hat{\boldsymbol{w}}$ is the maximizer of the penalized log-likelihood function.

Assuming equal prior probabilities for a model within a set of candidate models and some regularity conditions for the Laplace approximation, Ando et al. (2002) obtained a Bayesian information criterion (Konishi et al. 2004) that evaluates the proposed model estimated by the penalized maximum likelihood method:

$$\begin{aligned}\mathrm{BIC} = &-2\sum_{\alpha=1}^{n}\sum_{k=1}^{G} y_k^{(\alpha)} \log \pi_k(\boldsymbol{x}_\alpha;\hat{\boldsymbol{w}}) + n\lambda \hat{\boldsymbol{w}}^T\hat{\boldsymbol{w}} \\ &+ \log|R(\hat{H})| - (G-1)(m+1)\log\lambda, \end{aligned} \tag{12}$$

where $R(\hat{H})$ is the $(G-1)(m+1)$-dimensional matrix given in (9).

The adjusted parameters λ, ν, and m are determined from the minimizer of BIC in (12).

BIC for evaluating the radial basis function network generalized linear models was presented by Konishi et al. (2004). Ando et al. (2004) obtained BIC for the evaluation

of adaptive learning machines. BIC for the evaluation of wavelets regression models estimated by the method of regularization was proposed by Fujii and Konishi (2006).

5 Numerical results

In this section, we analyze various datasets by applying of the proposed nonlinear discriminant procedure. In the model selection process, we considered a three-dimensional grid search with regard to the adjusted parameters λ, ν, and m. Since the ranges of the grids generally depend on the data structure, it is difficult to identify suitable ranges. Generally, if the data structure is complicated, the optimal number of basis functions m may be large and the optimal values of λ and ν may be small. If the data structure is simple, the optimal value of m may be small and those of λ and ν may be large.

5.1 Analysis of benchmark datasets

We have investigated the performance of our proposed method by analyzing waveform data (Breiman et al. 1984) and vowel recognition data (Hastie et al. 1994; Ripley 1994). The waveform data consisted of three classes with 21 feature variables, and were generated from the following probability system:

$$x_k = \begin{cases} uH_1(k) + (1-u)H_2(k) + \varepsilon_k \ (g=1) \\ uH_1(k) + (1-u)H_3(k) + \varepsilon_k \ (g=2) \\ uH_2(k) + (1-u)H_3(k) + \varepsilon_k \ (g=3) \end{cases} \quad k = 1, \ldots, 21, \tag{13}$$

where u is uniform on $(0, 1)$, ε_k are standard normal variables, and $H_i(k)$ are shifted triangular waveforms; $H_1(k) = \max(6 - |k - 11|, 0)$, $H_2(k) = H_1(k - 4)$, and $H_3(k) = H_1(k + 4)$. We generated 300 values of training data with equal prior probability for each class by using the probability system (13). In the same way, 500 values of test data were also generated to compute the prediction error.

The vowel recognition data consisted of 11 classes with 10 feature variables. This data contained 528 training observations from eight speakers (four male and four female) and 462 test observations from seven speakers (four male and three female).

The proposed model (5) was fitted to both datasets by maximizing the penalized log-likelihood function (6). Figure 5 shows the relationship between the prediction errors and the values of the model selection criteria GIC given in (9) and BIC given in (12), obtained through analysis of the waveform data. As the values of our proposed model selection criteria become smaller, the prediction error also becomes smaller. This indicates that the proposed model selection criteria are useful for evaluating the prediction accuracy of the estimated model.

The values of some adjusted parameters $\boldsymbol{\theta} = (m, \log_{10}(\lambda), \nu)$ were chosen as minimizers of GIC or BIC. The number of basis functions ranged from 10 to 30. The candidates for the smoothing parameter and the hyperparameter for the waveform data were chosen on a geometrical grid with 50 knots between $\log_{10}(\lambda) = -2.0$ and $\log_{10}(\lambda) = -6.0$, and on a geometrical grid with 50 knots between $\log_{10}(\nu) = 0$ and $\log_{10}(\nu) = 1$, respectively. For the vowel recognition data, the candidates for the

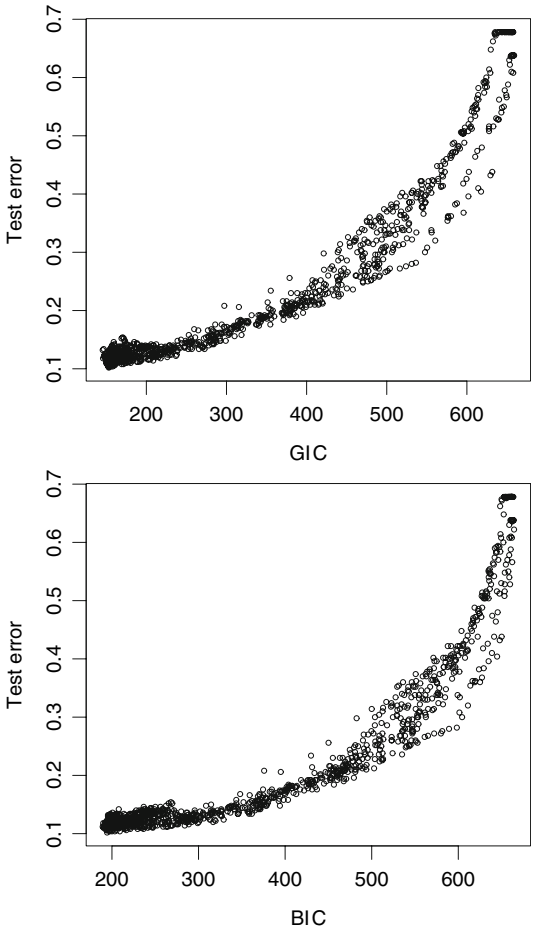


Fig. 5 The relationships between the prediction error and the values of model selection criteria

smoothing parameter and the hyperparameter were chosen on a geometrical grid with 50 knots between $\log_{10}(\lambda) = -3.0$ and $\log_{10}(\lambda) = -7.5$ and on a geometrical grid with 50 knots between $\log_{10}(\nu) = 0$ and $\log_{10}(\nu) = 1.75$, respectively.

The average values of the adjusted parameters for 10 runs of the waveform data, selected by use of GIC and BIC, were $\boldsymbol{\theta} = (15.0, -5.27, 1.71)$ and $\boldsymbol{\theta} = (13.2, -4.54, 2.46)$, respectively. In the case of the vowel recognition data, the use of GIC and BIC selected the adjusted parameters such that $\boldsymbol{\theta} = (20, -6.40, 3.16)$ and $\boldsymbol{\theta} = (10, -3.55, 1.50)$, respectively. Thus the procedure using BIC tends to choose fewer basis functions and larger values of λ than does the procedure based on GIC.

Table 1 summarizes the prediction errors. The prediction errors for the waveform data are average values over ten runs. When the results were compared with those of the previously proposed methods, our method performed very well; it gives the best prediction error.

Table 1 Comparison of prediction errors (%). The results, except for those obtained from our modeling strategy, are due to Hastie et al. (1994, 2001), Ando (2003), and Ando et al. (2004)

Method	Waveform	Vowel
Proposed model (5) with GIC	14.5	35.0
Proposed model (5) with BIC	14.2	35.9
Linear discriminant analysis	19.1	56
Quadratic discriminant analysis	20.5	67
Classification tree	28.9	56
Flexible discriminant analysis (MARS degree = 1)	19.1	45
Flexible discriminant analysis (MARS degree = 2)	21.5	42
Single-layer perceptron	–	67
Multilayer perceptron (88 hidden units)	–	49
Gaussian node network (528 hidden units)	–	45
Nearest-neighbor	–	44
Ando et al. (2004)'s adaptive learning machines	15.6	41
Ando (2003)'s kernel flexible discriminant analysis	15.3	40

5.2 Comparison of basis function construction methods

By means of an analysis of synthetic data (Ripley 1994), we have investigated the performance of the Gaussian radial functions with a hyperparameter (BF_v) given in (3) with that of ordinary Gaussian radial basis functions (7). The synthetic data consisted of two-dimensional feature variables and a binary class distribution.

A set of ordinary Gaussian radial basis functions was constructed by using the following five unsupervised procedures; these were originally developed in research on radial-basis-function networks. Moody and Darken (1989) used a k-means clustering algorithm to position the centers $\boldsymbol{\mu}_k$ while the scale parameters σ_k were determined by a "P-nearest-neighbor" heuristic, which uses the averaged Euclidean distance of the P nearest neighbors from each basis function. We used three such procedures, based on the first-nearest neighbor (MD_1), the second-nearest neighbor (MD_2), and the third-nearest neighbor (MD_3). Ranganath and Arun (1997) also used a k-means clustering algorithm to position the centers $\boldsymbol{\mu}_k$ and calculated the scale parameters by using the distance from the cluster center to the center nearest to another cluster (RA). For the scale parameters σ_k^2, Karayiannis and Mi (1997) suggested the use of

$$s_k^2 = \sqrt{\frac{1}{|A_k|} \sum_{\boldsymbol{x}_\alpha \in A_k} ||\boldsymbol{x}_\alpha - \boldsymbol{c}_k||}, \quad k = 1, \ldots, m,$$

where $|A_k|$ denotes the cardinality of A_k (KM). Broomhead and Lowe (1988) determined the centers $\boldsymbol{\mu}_k$ by randomly selecting from $\{\boldsymbol{x}_\alpha; \alpha = 1, \ldots, n\}$. The scale parameters were determined by using the maximum distance between the selected centers (BL). Note that these methods construct the basis functions in a completely

Table 2 Comparison of prediction errors obtained from various basis construction methods. The values of the adjusted parameters were chosen by use of GIC. The results, except for RA and KM, are due to Ando et al. (2001)

Method	Prediction error (%)	m	$\log_{10}(\lambda)$	ν
BF_ν	**9.6**	25	−6.45	7.1
MD_1	10.0	26	−5.33	–
MD_2	10.0	27	−5.33	–
MD_3	10.4	23	−5.33	–
RA	10.2	25	−6.27	–
KM	11.3	29	−5.78	–
BL	10.0	19	−7.00	–

unsupervised way and do not take the class label information into account, whereas our proposed method does. After basis functions were constructed using these methods, the unknown parameter vector $\boldsymbol{w}$ was estimated by maximizing the penalized log-likelihood function (6). The values of the adjusted parameters were chosen as the minimizer of GIC in (9).

Table 2 summarizes the prediction errors and the selected adjusted parameters. The candidate values of m, $\log_{10}(\lambda)$, and ν were set to be $\{10, 11, \ldots, 30\}$, a geometrical grid with 50 knots between $\log_{10}(\lambda) = -4$ and $\log_{10}(\lambda) = -7$, and a geometrical grid with 100 knots between $\log_{10}(\nu) = 0$ and $\log_{10}(\nu) = 1$, respectively.

It may be seen that our proposed method (BF_ν) is superior to the other methods, in the sense that it gives the smallest value of the prediction error. Similar results were also obtained from the use of BIC (12) instead of GIC. The scale parameters play an essential role in determining the smoothness of the decision boundaries. Nevertheless, the previously proposed methods determine the scale parameters heuristically and do not always give sufficient results. Therefore the use of the hyperparameter ν in the Gaussian radial basis functions helps us to improve the prediction accuracy of the classification.

5.3 Character recognition

We have applied our proposed method to the optical recognition of handwritten digits (Alpaydin and Kaynak 1998). Figure 6 shows a set of examples. In the analysis, 32×32 bitmaps were divided into nonoverlapping blocks of 4×4, and the number of pixels was counted in each block. As shown in Fig. 7, this handling generates an 8×8 feature matrix, where each element is an integer.

We constructed a proposed model (5) using 3,823 values of training data and evaluated the prediction performance by using 1,797 values of test data. The model was estimated by maximizing the penalized likelihood function (6). We then chose the adjusted parameters by minimizing GIC (9) and BIC (12). The candidate values of m were in the range from 30 to 100. The candidates for the smoothing parameter were chosen on a geometrical grid with 100 knots between $\log_{10}(\lambda) = -5$ and $\log_{10}(\lambda) = -7$. The candidates for the hyperparameter were chosen on a geometrical grid with 100 knots between $\log_{10}(\nu) = 0$ and $\log_{10}(\nu) = 0.8$.

As a result, a model with $\boldsymbol{\theta} = (61, -5.55, 1.84)$ was selected by use of GIC, and the corresponding training error and prediction error were 1.77 and 4.61%, respectively.

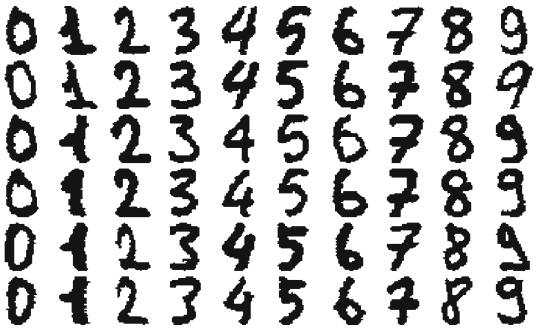
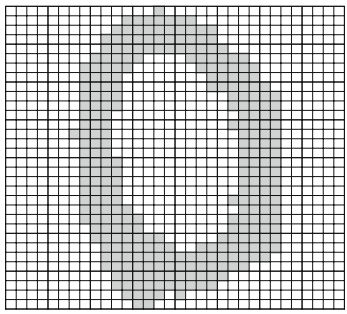


Fig. 6 Examples of optical recognition of handwritten digits data



Preprocessing data

(b)

0	0	6	13	10	0	0	0
0	2	14	5	10	12	0	0
0	4	11	0	1	12	7	0
0	5	8	0	0	9	8	0
0	4	12	0	0	8	8	0
0	3	15	2	0	11	8	0
0	0	13	15	10	15	5	0
0	0	5	13	9	1	0	0

A transformed data.

Fig. 7 An example of dimension reduction procedure. **a** Preprocessing data. **b** A transformed data

The use of BIC selected a model with $\boldsymbol{\theta} = (35, -5.10, 3.16)$, and the corresponding prediction error was 5.73%.

Table 3 summarizes the classification results for the test data using the model selected by use of GIC. The (i, j)th element indicates the number of data; the true number is i and the estimated number is j. So the trace of the matrix indicates the number of correctly classified data. Table 3 indicates that the accuracy rates for the characters 7, 8, and 9 are inferior to those for the other characters. Figure 8 shows some examples that are classified incorrectly. Under the characters, the true label and the classified label are shown. One reason for misclassification could be that it would be difficult even for humans to recognize these characters.

Table 3 Prediction error matrix obtained by the proposed method using GIC

Character	0	1	2	3	4	5	6	7	8	9
0	176	0	0	0	0	2	0	0	0	0
1	0	176	1	0	0	1	0	0	1	3
2	0	2	172	0	0	0	0	2	1	0
3	0	0	2	169	0	2	0	3	1	6
4	0	0	0	0	177	0	0	1	3	0
5	0	0	0	0	1	179	1	0	0	1
6	1	2	0	0	1	0	175	0	2	0
7	0	0	0	0	1	3	0	168	2	5
8	0	8	0	0	0	2	1	1	156	6
9	0	1	0	2	5	3	0	0	3	166

Fig. 8 A set of examples, classified incorrectly. Under the each character, the true label information is provided. The estimated labels are also shown in parentheses

5.4 Cancer classification using microarray data

Target-specific treatment for distinct cancer types has been recognized as an important element in improving medical therapy. The use of molecular information from microarray gene expression data to predict cancer types accurately has received a large amount of attention in recent years (Golub et al. 1999; Alizadeh et al. 2000; Dudoit et al. 2002).

The lymphoma dataset that we used (Alizadeh et al. 2000) consists of gene expression levels obtained from cDNA experiments involving two molecularly distinct diffuse large B-cell lymphomas: germinal-center B-like (GCB) lymphoma and activated B-like (AB) lymphoma. The dataset contains 4,682 gene expression profiles from $n = 42$ subjects. After deleting those gene expression profiles which had any missing information, we used $p = 2,041$ gene expression profiles.

To measure the prediction accuracy, we used a full leave-one-out cross-validation procedure. A proposed model was constructed by using $n - 1$ observations, and its performance was evaluated by using the remaining observation, which was not used for model estimation. The basis functions were reconstructed in each cross-validation run. The unknown parameters were then estimated by maximizing the penalized log-likelihood function (6). The optimum values of the adjusted parameters were chosen as minimizers of BIC in (12). It is clear that the identification of marker genes is equivalent to choosing the best set of genes out of all possible gene combinations. Even if the number of genes p is as small as 100, this is a time-consuming task. We therefore used the forward stepwise selection method to reduce the search space. The candidate

Table 4 List of the most important genes for distinguishing between germinal-center B-like and activated B-like lymphomas

Frequency	Gene description
1.00	Unknown; clone = 1268870
1.00	Unknown; clone = 825199
1.00	BCL-6; clone = 712395
1.00	Unknown; clone = 2020
1.00	JAW1 = lymphoid-restricted membrane protein; clone = 815539
0.98	Unknown; clone = 1333557
0.98	Unknown UG Hs.136345 ESTs; clone = 746300
0.98	Protein; tyrosine phosphatase, non-receptor type 4 = MEG1; clone = 1283105
0.98	Cyclin D2/KIAK0002 = 3' end of KIAK0002 cDNA; clone = 366412
0.98	T-cell protein-tyrosine phosphatase = protein tyrosine phosphatase, non-receptor type 2; clone = 1370148
0.98	Unknown UG Hs.192047 EST; clone = 1353659
0.98	Deoxycytidylate deaminase; clone = 489681
0.98	RPD3 L1 = homologue of yeast RPD3 transcription factor; clone = 548736
0.95	transcription factor ERF-1; clone = 594372
0.95	Unknown UG Hs.28355 ESTs; clone = 703735
0.95	Deoxycytidylate deaminase; clone = 1185959
0.93	RPD3L1 = homologue of yeast RPD3 transcription factor; clone = 814080
0.86	MCL1 = myeloid cell differentiation protein; clone = 50437
0.86	DNA (cytosine-5-)-methyltransferase; clone = 45941
0.83	Unknown; UG Hs.180562 EST; clone = 1334488

values of m, $\log_{10}(\lambda)$, and ν were set to be $\{10, 11, \ldots, 15\}$, $\{10^{-2}, 10^{-3}, \ldots, 10^{-5}\}$, and $\{1, 5, 10, 20\}$, respectively.

The accuracy rate of the full leave-one-out cross-validation study was 97.6% (one observation, with label GCB, was classified into the AB class incorrectly). Table 4 shows a set of genes which are frequently included in models designed using a full leave-one-out cross-validation procedure. Table 4 includes many important genes. For example, the genes BCL-6, MCL-1, JAW1, and RPD3 provide useful information for research on diffuse large B-cell lymphomas (Alizadeh et al. 2000; Shaffer et al. 2002; Troyanskaya et al. 2002; Zhou et al. 2001). Therefore our method identified the important genes for making a diagnostic decision. There are also many unknown genes in Table 4. We hope that these genes will also be important for cancer classification and prognosis in clinical practice.

6 Conclusions

In this article, we have described an extension of linear logistic discriminant models to nonlinear models by replacing the linear predictor with a radial-basis-expansion

predictor. We have utilized Gaussian radial basis functions with a hyperparameter to construct basis functions taking the class label information into account, whereas previously proposed methods could not do this. We have proposed a nonlinear modeling technique, in which a set of basis functions with a hyperparameter is constructed, the unknown parameters are estimated by the penalized maximum likelihood method, and then the estimated model is evaluated to select a suitable one from competing models. Model selection criteria play an essential role in constructing models. We have employed two model selection criteria, from information-theoretic and Bayesian viewpoints, that enable us to evaluate models estimated by the penalized maximum likelihood method.

As demonstrated by various numerical examples, the proposed modeling strategy performs very well even when the dimension of the feature variables is very large. Monte Carlo experiments also showed that the Gaussian radial basis functions with a hyperparameter improved the prediction accuracy, compared with previously proposed methods. We would recommend nonlinear multiclass classification by implementing our proposed method.

Acknowledgments The authors are grateful to the UCI Repository of Machine Learning Databases (Blake and Merz 1998) for providing data on the optical recognition of handwritten digits. The authors would like to acknowledge three anonymous referees for careful reviews and constructive comments that have substantially improved the article.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B.N. Petrov, F Csaki (Eds.), *2nd International Symposium on Information Theory* pp. 267–281. Budapest: Akademiai Kiado.

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

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature, 403*, 503–511.

Alpaydin, E., Kaynak, C. (1998) Cascading classifiers. *Kybernetika, 34*, 369–374.

Ando, T. (2003). Kernel flexible discriminant analysis for classifying high-dimensional data with nonlinear structure and its applications (in Japanese). *Proceedings of the Institute of Statistical Mathematics, 51*, 389–406

Ando, T., Imoto, S., Konishi, S. (2001). Estimating nonlinear regression models based on radial basis function networks (in Japanese). *Japanese Journal of Applied Statistics, 30*, 19–35.

Ando, T., Simauchi, J., Konishi, S. (2002). Nonlinear pattern recognition using radial basis function networks and its application (in Japanese). *Japanese Journal of Applied Statistics, 31*, 123–139.

Ando, T., Imoto, S., Konishi, S. (2004). Adaptive learning machines for nonlinear classification and Bayesian information criteria. *Bulletin of Informatics and Cybernetics, 36*, 147–162.

Ando, T., Imoto, S., Konishi, S. (2005). Nonlinear regression modeling via regularized radial basis function networks. *Journal of Statistical Planning and Inference* (to appear).

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford: Oxford University Press.

Blake, C. L., Merz, C. J. (1998). UCI repository of machine learning databases. http://www.ics.uci.edu/∼mlearn/MLRepository.html, University of California, Department of Information and Computer Sciences.

Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth.

Broomhead, D. S., Lowe, D. (1988). Multivariable functional interpolation and adaptive networks. *Complex Systems, 2*, 321–335.

Dudoit, S., Fridlyand, J., Speed, T. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association, 97*, 77–87.

Eilers, P. H. C., Marx, B. D. (1996). Flexible smoothing with *B*-splines and penalties (with discussion). *Statistical Science, 11*, 89–121.

Fujii, T., Konishi, S. (2006). Nonlinear regression modeling via regularized wavelets and smoothing parameter selection. *Journal of Multivariate Analysis, 97*, 2023–2033.

Girosi, F., Jones, M., Poggio, T. (1995). Regularization theory and neural architectures. *Neural Computation, 7*, 219–269.

Green, P. J., Silverman, B. W. (1994). *Nonparametric regression and generalized linear models*. London: Chapman & Hall.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science, 286*, 531–537.

Hastie, T., Tibshirani, R., Buja, A. (1994). Flexible discriminant analysis by optimal scoring. *Journal of the American Statistical Association, 89*, 1255–1270.

Hastie, T., Tibshirani, R., Friedman, J. (2001). *The elements of statistical learning*. New York: Springer.

Hosmer, D. W., Lemeshow, S. (1989). *Applied logistic regression*, New York: Wiley.

Imoto, S., 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.

Karayiannis, N. B., Mi, G. W. (1997). Growing radial basis neural networks: merging supervised and unsupervised learning with network growth techniques. *IEEE Transactions on Neural Networks, 8*, 1492–1506.

Kass, R. E., Tierney, L., Kadane, J. B. (1990). The validity of posterior expansions based on Laplace's method. In S. Geisser, J. S. Hodges, Press, S. J. Zellner, A. (Eds.), *Essays in honor of George Barnard*, pp. 473–488, Amsterdam: North-Holland.

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

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

Kullback, S., Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics, 22*, 79–86.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In L. M. LeCam, Neyman, J. (Eds.), *Proceeding of the fifth Berkeley symposium on mathematics, statistics, and probability* p. 281. Berkeley: University of California Press.

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

Nabney, I. T. (2002). *NETLAB algorithms for pattern recognition*. UK: Springer.

Nonaka, Y., Konishi, S. (2005). Nonlinear regression modeling using regularized local likelihood method. *Annals of the Institute of Statistical Mathematics, 57*, 617–635.

Ranganath, S., Arun, K. (1997). Face recognition using transform features and neural networks. *Pattern Recognition, 30*, 1615–1622.

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, UK: Cambridge University Press.

Sato, T. (1996). On artificial neural networks as a statistical model (in Japanese). *Proceedings of the Institute of Statistical Mathematics, 44*, 85–98.

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

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

Shaffer, A. L., Rosenwald, A., Staudt, L. M. (2002). Lymphoid malignancies: the dark side of B-cell differentiation. *Nature Reviews Immunology, 2*, 920–933.

Tierney, L., Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association, 81*, 82–86.

Tierney, L., Kass, R. E., Kadane, J. B. (1989). Fully exponential Laplace approximations to expectations and variances of nonpositive functions. *Journal of the American Statistical Association, 84*, 710–716.

Troyanskaya, O. G., Garber, M. E., Brown, P. O., Botstein, D., Altman, R. B. (2002). Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics, 18*, 1454–1461.

Webb, A. (1999). *Statistical pattern recognition*. London: Arnold.
Xu, L. (1998). RBF nets, mixture experts, and Bayesian ying-yang learning. *Neurocomputing, 19*, 223–257.
Xu, L., Jordan, M. I., Hinton, G. E. (1995). An alternative model for mixtures of experts. In J. D. Cowan, et al. (Eds.), *Advances in Neural Information Processing Systems 7*. pp. 633–640. Cambridge, MA: MIT Press.
Zhou, P., Levy, N. B., Xie, H., Qian, L., Lee, C. Y., Gascoyne, R. D., et al. (2001). MCL1 transgenic mice exhibit a high incidence of B-cell lymphoma manifested as a spectrum of histologic subtypes. *Blood, 97*, 3902–3909.