Biometrika (2004), **91,** 1, *pp.* 27–43

Printed in Great Britain

Bayesian information criteria and smoothing parameter selection in radial basis function networks

BY SADANORI KONISHI, TOMOHIRO ANDO

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

konishi@math.kyushu-u.ac.jp ando@math.kyushu-u.ac.jp

AND SEIYA IMOTO

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

imoto@ims.u-tokyo.ac.jp

SUMMARY

By extending Schwarz's (1978) basic idea we derive a Bayesian information criterion which enables us to evaluate models estimated by the maximum penalised likelihood method or the method of regularisation. The proposed criterion is applied to the choice of smoothing parameters and the number of basis functions in radial basis function network models. Monte Carlo experiments were conducted to examine the performance of the nonlinear modelling strategy of estimating the weight parameters by regularisation and then determining the adjusted parameters by the Bayesian information criterion. The simulation results show that our modelling procedure performs well in various situations.

Some key words: Bayes approach; Maximum penalised likelihood; Model selection; Neural network; Nonlinear regression.

1. INTRODUCTION

Recent years have seen the development of various types of nonlinear model such as neural networks, kernel methods and splines. Nonlinear models are generally characterised by including a large number of parameters. Since maximum likelihood methods yield unstable parameter estimates, the adopted model is usually estimated by the maximum penalised likelihood method (Good & Gaskins, 1971; Green & Silverman, 1994) or the method of regularisation or the Bayes approach and so on.

Crucial issues with nonlinear modelling are the choice of a smoothing parameter, the number of basis functions in splines and the number of hidden units in neural networks. Choosing these parameters in the modelling process can be viewed as a model selection and evaluation problem. Schwarz (1978) proposed the Bayesian information criterion, BIC. However, theoretically, the BIC covers only models estimated by the maximum likelihood method. It still remains to construct a criterion for evaluating nonlinear models estimated by the maximum penalised likelihood method.

This paper has two aims. First, the BIC is extended to cover the evaluation of models estimated by the maximum penalised likelihood method. Secondly, the criterion is used

to construct radial basis function network nonlinear regression models. In §2, by extending Schwarz's basic ideas, we present various types of Bayesian information criterion. Section 3 describes nonlinear regression modelling based on radial basis function networks. In §4 we investigate the performance of the nonlinear modelling techniques, using Monte Carlo simulations. Section 5 provides discussion, including possibilities for future work.

2. BAYESIAN APPROACH TO MODEL SELECTION

2·1. *Bayesian information criteria*

Suppose we are interested in selecting a model from a set of candidate models $M_1, \ldots, M_r$ for a given observation vector y of dimension n. It is assumed that model M_k is characterised by the probability density $f_k(y|\theta_k)$, where $\theta_k \in \Theta_k \subset R^{p_k}$ is a p_k-dimensional vector of unknown parameters. Let $\pi_k(\theta_k|\lambda_k)$ be the prior density for parameter vector θ_k under model M_k, where λ_k is a hyperparameter. The posterior probability of the model M_k for a particular dataset y is then given by

$$\mathrm{pr}(M_k|y) = \mathrm{pr}(M_k) \int f_k(y|\theta_k)\pi_k(\theta_k|\lambda_k)d\theta_k \Big/ \sum_{\alpha=1}^{r} \mathrm{pr}(M_\alpha) \int f_\alpha(y|\theta_\alpha)\pi_\alpha(\theta_\alpha|\lambda_\alpha)d\theta_\alpha,$$

where $\mathrm{pr}(M_k)$ is the prior probability for model M_k.

The Bayes approach for selecting a model is to choose the model with the largest posterior probability among a set of candidate models for given values of λ_k. This is equivalent to choosing the model that maximises

$$\mathrm{pr}(M_k) \int f_k(y|\theta_k)\pi_k(\theta_k|\lambda_k)d\theta_k := \mathrm{pr}(M_k) f_k(y|\lambda_k). \tag{1}$$

The quantity $f_k(y|\lambda_k)$ obtained by integrating over the parameter space Θ_k is the marginal probability of the data y under model M_k, and it can be rewritten as

$$f_k(y|\lambda_k) = \int \exp\{nq_k(\theta_k|y, \lambda_k)\}d\theta_k, \tag{2}$$

where

$$q_k(\theta_k|y, \lambda_k) = n^{-1}\{\log f_k(y|\theta_k) + \log \pi_k(\theta_k|\lambda_k)\}. \tag{3}$$

We first consider the case where $\log \pi_k(\theta_k|\lambda_k) = O(n)$. Let $\hat{\theta}_k$ be the mode of $q_k(\theta_k|y, \lambda_k)$. Then, using the Laplace method for integrals in the Bayesian framework developed by Tierney & Kadane (1986), Tierney et al. (1989) and Kass et al. (1990), we have under certain regularity conditions the Laplace approximation to the marginal distribution (2) in the form

$$\begin{aligned} f_k(y|\lambda_k) &= \int \exp\{nq_k(\theta_k|y, \lambda_k)\}d\theta_k \\ &= \frac{(2\pi)^{p_k/2}}{n^{p_k/2}|Q_{\lambda_k}(\hat{\theta}_k)|^{\frac{1}{2}}} \exp\{nq_k(\hat{\theta}_k|y, \lambda_k)\}\{1 + O_p(n^{-1})\}, \end{aligned} \tag{4}$$

where

$$Q_{\lambda_k}(\hat{\theta}_k) = -\frac{\partial^2\{q_k(\theta_k|y, \lambda_k)\}}{\partial\theta_k\partial\theta_k^T}\bigg|_{\theta_k=\hat{\theta}_k}.$$

Substituting the Laplace approximation in equation (1) and taking the logarithm of the resulting formula, we have

$$-2\log\{\mathrm{pr}(M_k)f_k(y|\lambda_k)\} = -2\log f_k(y|\hat{\theta}_k) - 2\log\pi_k(\hat{\theta}_k|\lambda_k) + p_k\log n + \log|Q_{\lambda_k}(\hat{\theta}_k)| - 2\log\mathrm{pr}(M_k) - p_k\log 2\pi + O_p(n^{-1}). \quad (5)$$

Choosing the model with the largest posterior probability among a set of candidate models for given values of λ_k is equivalent to choosing the model that minimises the criterion (5).

We next consider the case where $\log\pi_k(\theta_k|\lambda_k) = O(1)$. Then the mode $\hat{\theta}_k$ of $q_k(\theta_k|y, \lambda_k)$ in (3) can be expanded as

$$\hat{\theta}_k = \hat{\theta}_k^{(\mathrm{ML})} + \frac{1}{n}J_k^{-1}(\hat{\theta}_k^{(\mathrm{ML})})\frac{\partial}{\partial\theta_k}\log\pi_k(\theta_k|\lambda_k)\Big|_{\theta_k=\hat{\theta}_k^{(\mathrm{ML})}} + O_p(n^{-2}), \quad (6)$$

where $\hat{\theta}_k^{(\mathrm{ML})}$ is the maximum likelihood estimate of θ_k in the model $f_k(y|\theta_k)$ and

$$J_k(\hat{\theta}_k^{(\mathrm{ML})}) = -\frac{1}{n}\frac{\partial^2\log f_k(y|\theta_k)}{\partial\theta_k\partial\theta_k^{\mathrm{T}}}\Big|_{\theta_k=\hat{\theta}_k^{(\mathrm{ML})}}.$$

Substituting the stochastic expansion (6) in equation (5) yields

$$-2\log\{\mathrm{pr}(M_k)f_k(y|\lambda_k)\} = -2\log f_k(y|\hat{\theta}_k^{(\mathrm{ML})}) - 2\log\pi_k(\hat{\theta}_k^{(\mathrm{ML})}|\lambda_k) + p_k\log n + \log|J_k(\hat{\theta}_k^{(\mathrm{ML})})| - 2\log\mathrm{pr}(M_k) - p_k\log 2\pi + O_p(n^{-1}).$$

Ignoring the term of order $O(1)$ and higher-order terms in this equation, we have Schwarz's (1978) Bayesian information criterion,

$$\mathrm{BIC} = -2\log f_k(y|\hat{\theta}_k^{(\mathrm{ML})}) + p_k\log n. \quad (7)$$

Suppose that the prior probabilities, $\mathrm{pr}(M_k)$, are all equal, and that the prior density $\pi_k(\theta_k|\lambda_k)$ is sufficiently flat in the neighbourhood of $\hat{\theta}_k$. These conditions lead to the modification of equation (7) to the following:

$$\mathrm{BIC}_{\mathrm{I}} = -2\log f_k(y|\hat{\theta}_k^{(\mathrm{ML})}) + p_k\log n + \log|J_k(\hat{\theta}_k^{(\mathrm{ML})})| - p_k\log 2\pi.$$

This variant, based on the inclusion of the term $\log|J_k(\hat{\theta}_k^{(\mathrm{ML})})|$, is regarded as an improved version of BIC.

The use of Laplace's method for integrals has been extensively investigated as a useful tool for approximating Bayesian predictive distributions, Bayes factors and Bayesian model selection criteria (Davison, 1986; Clarke & Barron, 1994; Kass & Wasserman, 1995; Kass & Raftery, 1995; O'Hagan, 1995; Neath & Cavanaugh, 1997; Pauler, 1998; Lanterman, 2001).

2·2. *Smoothing parameter selection*

We extend the BIC so that it can be applied to the evaluation of models estimated by the method of maximum penalised likelihood. In this subsection we drop the notational dependence on the model M_k and consider the Bayesian approach with equal prior probabilities. It is assumed that the logarithm of a prior density is of order $O(n)$, that is $\log\pi(\theta|\lambda) = O(n)$, where θ is p-dimensional.

Suppose that the model is constructed by maximising the penalised loglikelihood function

$$\ell_\lambda(\theta) = \log f(y|\theta) - \frac{n\lambda}{2}\theta^{\mathrm{T}}D\theta, \quad (8)$$

where D is a $p \times p$ known matrix of rank $p-d$ and λ is a smoothing parameter. The penalty term corresponds to a singular multivariate normal prior density, $\pi(\theta|\lambda)$, where

$$\pi(\theta|\lambda) = (2\pi)^{-(p-d)/2}(n\lambda)^{(p-d)/2}|D|_+^{\frac{1}{2}} \exp\left(-\frac{n\lambda}{2}\theta^{\mathrm{T}} D\theta\right), \tag{9}$$

in which $|D|_+$ is the product of the $(p-d)$ nonzero eigenvalues of D. For the Bayesian justification of the maximum penalised likelihood approach, we refer to Silverman (1985) and Wahba (1990, Ch. 1).

It follows from equation (4) that minus twice the log marginal likelihood, $-2\log\{f(y|\lambda)\}$, can be approximated by

$$-2\log\{f(y|\lambda)\} \doteqdot -2\log f(y|\hat{\theta}) - 2\log\pi(\hat{\theta}|\lambda) + p\log n + \log|Q_\lambda(\hat{\theta})| - p\log 2\pi. \tag{10}$$

By substituting the prior density (9) into this equation, we have

$$\begin{aligned}\mathrm{BIC_P}(\lambda) = &-2\log f(y|\hat{\theta}) + n\lambda\hat{\theta}^{\mathrm{T}} D\hat{\theta} + d\log n + \log|Q_\lambda(\hat{\theta})| - \log|D|_+ \\ &- d\log 2\pi - (p-d)\log\lambda,\end{aligned} \tag{11}$$

where $Q_\lambda(\theta) = -n^{-1}\partial^2\log f(y|\theta)/\partial\theta\,\partial\theta^{\mathrm{T}} + \lambda D$ and the estimator $\hat{\theta}$ is a solution of the equation

$$\frac{\partial q(\theta|y,\lambda)}{\partial\theta} = \frac{\partial}{\partial\theta}\left\{\frac{1}{n}\log f(y|\theta) - \frac{\lambda}{2}\theta^{\mathrm{T}} D\theta + \frac{p-d}{2n}\log\left(\frac{n\lambda}{2\pi}\right) + \frac{1}{2n}\log|D|_+\right\} = 0. \tag{12}$$

This implies that $\hat{\theta}$ is the maximiser of the penalised loglikelihood function (8).

We choose the smoothing parameter λ to minimise $\mathrm{BIC_P}(\lambda)$. In the context of model selection, we minimise $\mathrm{BIC_P}(\lambda)$ over λ for each model, and then choose a model for which $\mathrm{BIC_P}(\lambda)$ is minimised over a set of competing models, which might consist of various types of nonparametric regression model estimated by the maximum penalised likelihood method.

3. RADIAL BASIS FUNCTION NETWORK REGRESSION MODELLING

3·1. *Preamble*

Section 3·2 presents a radial basis function network regression model and derives a Bayesian information criterion in the context of generalised linear models (Nelder & Wedderburn, 1972; McCullagh & Nelder, 1989, Ch. 2). The resulting formulae are applied in §3·3 to special cases involving Gaussian, logistic and Poisson nonlinear regression models based on radial basis function networks. For background about radial basis function networks, we refer to Broomhead & Lowe (1988), Moody & Darken (1989), Poggio & Girosi (1990), Bishop (1995, p. 164), Ripley (1996, p. 131), Webb (1999, p. 140), Ando et al. (2001) and references given therein.

3·2. *Radial basis function network generalised linear models*

Suppose that we have n independent observations y_α corresponding to q-dimensional design points x_α, for $\alpha = 1, \ldots, n$. In generalised linear models y_α are assumed to be drawn from the exponential family of distributions with densities

$$f(y_\alpha|x_\alpha; \xi_\alpha, \psi) = \exp[\{y_\alpha\xi_\alpha - u(\xi_\alpha)\}/\psi + v(y_\alpha, \psi)], \tag{13}$$

where $u(.)$ and $v(.,.)$ are functions specific to each distribution, and ψ is an unknown scale parameter. The conditional expectation $E(Y_\alpha|x_\alpha) = \mu_\alpha = u'(\xi_\alpha)$ is related to the predictor $\eta_\alpha = h(\mu_\alpha)$, where $h(.)$ is a link function. We model the predictor η_α by a radial basis function network of the form

$$\eta_\alpha = \sum_{j=1}^{p} w_j \phi_j(x_\alpha) + w_0 \quad (\alpha = 1, 2, \ldots, n), \tag{14}$$

where $\{\phi_j(x); j = 1, \ldots, p\}$ is a set of radial basis functions. We shall use Gaussian basis functions, such that

$$\phi_j(x) = \phi_j(x; \mu_j, \sigma_j) = \exp\left(-\frac{\|x - \mu_j\|^2}{2v\sigma_j^2} \right) \quad (j = 1, 2, \ldots, p),$$

where μ_j is the q-dimensional vector determining the location of the basis function, σ_j^2 determines the width, v is a hyperparameter and $\|.\|$ is the Euclidian norm. The hyperparameter v adjusts the amount of overlapping among the basis functions so that the network can capture the structure in the data over the region of the input space (Ando et al., 2001).

If we write

$$\eta_\alpha = \sum_{j=1}^{p} w_j \phi_j(x_\alpha) + w_0 = w^{\mathrm{T}} b(x_\alpha),$$

where $w = (w_0, w_1, \ldots, w_p)^{\mathrm{T}}$ and $b(x_\alpha) = (1, \phi_1(x_\alpha), \ldots, \phi_p(x_\alpha))^{\mathrm{T}}$, the parameter ξ_α in (13) can be expressed as $\xi_\alpha = u'^{-1}[h^{-1}\{w^{\mathrm{T}} b(x_\alpha)\}]$. Then it follows from (13) and (14) that the data are summarised by a model from a class of probability densities of the form

$$f(y_\alpha|x_\alpha; w, \psi) = \exp([y_\alpha r\{w^{\mathrm{T}} b(x_\alpha)\} - s\{w^{\mathrm{T}} b(x_\alpha)\}]/\psi + v(y_\alpha, \psi)), \tag{15}$$

where $r(.) = u'^{-1} \circ h^{-1}(.)$ and $s(.) = u \circ u'^{-1} \circ h^{-1}(.)$.

In the model-fitting, the radial basis function network model is generally determined by a two-stage procedure. In the first stage we construct the basis functions by analysing the data on the explanatory variables, using a k-means clustering algorithm. This algorithm divides the input dataset $\{x_1, \ldots, x_n\}$ into p clusters $C_1, \ldots, C_p$ corresponding to the number of basis functions. Then the centres and widths are determined by

$$\hat{\mu}_j = \frac{1}{n_j} \sum_{\alpha \in C_j} x_\alpha, \quad \hat{\sigma}_j^2 = \frac{1}{n_j} \sum_{\alpha \in C_j} \|x_\alpha - \hat{\mu}_j\|^2,$$

where n_j is the number of observations which belong to the jth cluster C_j.

In the second stage, we estimate the weight parameters w by maximising the penalised loglikelihood function

$$l_\lambda(w, \psi) = \sum_{\alpha=1}^{n} ([y_\alpha r\{w^{\mathrm{T}} b(x_\alpha)\} - s\{w^{\mathrm{T}} b(x_\alpha)\}]/\psi + v(y_\alpha, \psi)) - \frac{n\lambda}{2} w^{\mathrm{T}} D w, \tag{16}$$

where D is a $(p+1) \times (p+1)$ positive semidefinite matrix and λ is a regularisation parameter. If a hyperparameter v is used, the smoothness of the fitted model is mainly determined by v, and the regularisation parameter λ has the effect of reducing the variances of the network parameter estimates. For the penalty term, we use the second-order penalty given by

$$\sum_{j=2}^{p} (\Delta^2 w_j)^2 = w^{\mathrm{T}} D_2^{\mathrm{T}} D_2 w,$$

where Δ is the difference operator defined by $\Delta w_j = w_j - w_{j-1}$ and D_2 is a $(p-1) \times (p+1)$ matrix representation of the difference operator Δ^2. The use of difference penalties has been investigated by Whittaker (1923), Tanabe & Tanaka (1983), Green & Yandell (1985) and O'Sullivan et al. (1986).

The maximum penalised likelihood estimator $\hat{w}$ is a solution of the equation $\partial l_\lambda(w, \psi)/\partial w = 0$. This equation is generally nonlinear in w, so we use Fisher's scoring algorithm (Nelder & Wedderburn, 1972; Green & Silverman, 1994). For fixed values of ψ, v, λ and the number of basis functions p, the Fisher scoring iteration may be expressed as

$$w^{\text{new}} = (B^{\mathrm{T}} W B + n\lambda D_2^{\mathrm{T}} D_2)^{-1} B^{\mathrm{T}} W \zeta,$$

where $B = (b(x_1), \ldots, b(x_n))^{\mathrm{T}}$, W is an $n \times n$ diagonal matrix with ith diagonal element $w_{ii} = \{\psi u''(\xi_i) h'(\mu_i)^2\}^{-1}$ and ζ an n-dimensional vector with $\zeta_i = (y_i - \mu_i) h'(\mu_i) + w^{\mathrm{T}} b(x_i)$. If $h(.)$ is a canonical link function such as $h^{-1}(.) = u'(.)$, $W = \mathrm{diag}\,(w_{11}, \ldots, w_{nn})$ and $\zeta = (\zeta_1, \ldots, \zeta_n)^{\mathrm{T}}$ simplify to

$$w_{ii} = u''\{w^{\mathrm{T}} b(x_i)\}/\psi, \quad \zeta_i = (y_i - \mu_i)/u''\{w^{\mathrm{T}} b(x_i)\} + w^{\mathrm{T}} b(x_i).$$

After $\hat{w}$ is obtained, the estimate of the scale parameter $\hat{\psi}$ is obtained as a solution of $\partial l_\lambda(\hat{w}, \psi)/\partial \psi = 0$. Replacing w and ψ in (15) by their sample estimates $\hat{w}$ and $\hat{\psi}$ yields the statistical model $f(y_\alpha | x_\alpha; \hat{w}, \hat{\psi})$, which depends on the values of λ, v and p. We use $\mathrm{BIC_P}$ to choose appropriate values for these parameters. The result is summarised in the following theorem.

THEOREM 1. *Let $f(y_\alpha | x_\alpha; w, \psi)$ be a radial basis function network generalised linear model given by* (15), *and let $f(y_\alpha | x_\alpha; \hat{w}, \hat{\psi})$ be the statistical model estimated by maximising the penalised loglikelihood function* (16). *Then a Bayesian information criterion for evaluating $f(y_\alpha | x_\alpha; \hat{w}, \hat{\psi})$ is*

$$\begin{aligned} \mathrm{BIC_P^{(G)}} = & -2 \sum_{\alpha=1}^{n} ([y_\alpha r\{(\hat{w}^{\mathrm{T}} b(x_\alpha)\} - s\{\hat{w}^{\mathrm{T}} b(x_\alpha)\}]/\hat{\psi} + v(y_\alpha, \hat{\psi})) + n\lambda \hat{w}^{\mathrm{T}} D_2^{\mathrm{T}} D_2 \hat{w} \\ & - 3 \log(2\pi/n) + \log|Q_\lambda(\hat{w}, \hat{\psi})| - \log|D_2^{\mathrm{T}} D_2|_+ - (p-1) \log \lambda, \end{aligned}$$

where

$$Q_\lambda(\hat{w}, \hat{\psi}) = \frac{1}{n\hat{\psi}} \begin{pmatrix} B^{\mathrm{T}} \Gamma B + n\hat{\psi}\lambda D_2^{\mathrm{T}} D_2 & B^{\mathrm{T}} e/\hat{\psi} \\ e^{\mathrm{T}} B/\hat{\psi} & -\hat{\psi} \sum_{\alpha=1}^{n} q_\alpha \end{pmatrix}.$$

Here Γ is an $n \times n$ diagonal matrix, e is an n-dimensional vector and $\sum_\alpha q_\alpha$ is the second derivative of $l_\lambda(w, \psi)/n$ with respect to ψ, with

$$\begin{aligned} \Gamma_{\alpha\alpha} &= \frac{(y_\alpha - \hat{\mu}_\alpha)\{u'''(\hat{\xi}_\alpha) h'(\hat{\mu}_\alpha) + u''(\hat{\xi}_\alpha)^2 h''(\hat{\mu}_\alpha)\}}{\{u''(\hat{\xi}_\alpha) h'(\hat{\mu}_\alpha)\}^3} + \frac{1}{u''(\hat{\xi}_\alpha) h'(\hat{\mu}_\alpha)^2}, \\ e_\alpha &= (y_\alpha - \hat{\mu}_\alpha)/\{u''(\hat{\xi}_\alpha) h'(\hat{\mu}_\alpha)\}, \\ q_\alpha &= 2[y_\alpha r\{\hat{w}^{\mathrm{T}} b(x_\alpha)\} - s\{\hat{w}^{\mathrm{T}} b(x_\alpha)\}]/\hat{\psi}^3 + \partial^2 v(y_\alpha, \psi)/\partial \psi^2|_{\psi=\hat{\psi}}, \end{aligned}$$

for $\alpha = 1, \ldots, n$.

Canonical link functions relate the parameter ξ_α in the exponential family (13) directly to the predictor $\eta_\alpha = \sum_{j=1}^{p} w_j \phi_j(x_\alpha) + w_0$ in (14), and lead to

$$f(y_\alpha|x_\alpha; \hat{w}, \hat{\psi}) = \exp\left([\, y_\alpha \hat{w}^{\mathrm{T}} b(x_\alpha) - u\{\hat{w}^{\mathrm{T}} b(x_\alpha)\}\,]/\hat{\psi} + v(y_\alpha, \hat{\psi})\right). \tag{17}$$

Then we have the following theorem.

THEOREM 2. *Let $h(.)$ be the canonical link function so that $h(.) = u'^{-1}(.)$. Then a Bayesian information criterion for evaluating the statistical model $f(y_\alpha|x_\alpha; \hat{w}, \hat{\psi})$ given by equation* (17) *is*

$$\begin{aligned} \mathrm{BIC}_{\mathrm{P}}^{(\mathrm{C})} = &-2 \sum_{\alpha=1}^{n} \left([\, y_\alpha \hat{w}^{\mathrm{T}} b(x_\alpha) - u\{\hat{w}^{\mathrm{T}} b(x_\alpha)\}\,]/\hat{\psi} + v(y_\alpha, \hat{\psi})\right) + n\lambda \hat{w}^{\mathrm{T}} D_2^{\mathrm{T}} D_2 \hat{w} \\ &- 3 \log(2\pi/n) + \log|Q_\lambda^{(\mathrm{C})}(\hat{w}, \hat{\psi})| - \log|D_2^{\mathrm{T}} D_2|_+ - (p-1)\log\lambda, \end{aligned}$$

where $Q_\lambda^{(\mathrm{C})}(\hat{w}, \hat{\psi})$ can be obtained by replacing Γ, q_α and e in $Q_\lambda(\hat{w}, \hat{\psi})$ by, respectively,

$$\begin{aligned} \Gamma^{(\mathrm{C})} &= \mathrm{diag}\,[u''\{\hat{w}^{\mathrm{T}} b(x_1)\}, \ldots, u''\{\hat{w}^{\mathrm{T}} b(x_n)\}\,], \\ q_\alpha^{(\mathrm{C})} &= 2[\, y_\alpha \hat{w}^{\mathrm{T}} b(x_\alpha) - u\{\hat{w}^{\mathrm{T}} b(x_\alpha)\}\,]/\hat{\psi}^3 + \partial^2 v(y_\alpha, \psi)/\partial\psi^2|_{\psi=\hat{\psi}}, \\ e^{(\mathrm{C})} &= (y_1 - \hat{\mu}_1, \ldots, y_n - \hat{\mu}_n)'. \end{aligned}$$

3·3. *Some special cases*

Example 1: *Radial basis function network Gaussian regression model.* In this case $y_\alpha = m(x_\alpha) + \varepsilon_\alpha (\alpha = 1, \ldots, n)$, where $m(.)$ is an unknown smooth function and the errors ε_α are independently and normally distributed with mean zero and variance σ^2.

We represent $m(.)$ in terms of a radial basis function network as

$$m(x_\alpha) = \sum_{j=1}^{p} w_j \phi_j(x_\alpha) + w_0 \quad (\alpha = 1, 2, \ldots, n),$$

where $\{\phi_j(x); j = 1, \ldots, p\}$ is a prescribed set of Gaussian basis functions. Then the general approach of §3·2 leads to

$$f_{\mathrm{N}}(y_\alpha|x_\alpha; \hat{w}, \hat{\sigma}^2) = \frac{1}{\sqrt{(2\pi\hat{\sigma}^2)}} \exp\left[-\frac{1}{2\hat{\sigma}^2} \{y_\alpha - \hat{w}^{\mathrm{T}} b(x_\alpha)\}^2 \right] \quad (\alpha = 1, \ldots, n), \tag{18}$$

where

$$\hat{w} = (B^{\mathrm{T}} B + n\beta D_2^{\mathrm{T}} D_2)^{-1} B^{\mathrm{T}} y, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{\alpha=1}^{n} \{y_\alpha - \hat{w}^{\mathrm{T}} b(x_\alpha)\}^2$$

with $\beta = \lambda\sigma^2$.

If in the Bayesian framework we specify a Gaussian prior density $\pi(w|\lambda)$ as in (9), the posterior distribution for w is normal with mean vector $\hat{w}$ and covariance matrix $(B'B/\sigma^2 + n\lambda D)^{-1}$; see Silverman (1985) and Spiegelhalter et al. (2002).

Taking $u(\hat{\xi}_\alpha) = \hat{\xi}_\alpha^2/2$, $\psi = \hat{\sigma}^2$, $v(y_\alpha, \psi) = -y_\alpha^2/(2\hat{\sigma}^2) - \log\{\hat{\sigma}\sqrt{(2\pi)}\}$ and $h(\hat{\mu}_\alpha) = \hat{\mu}_\alpha$ in Theorem 2, we obtain the following Bayesian information criterion for evaluating the statistical model $f_N(y_\alpha|x_\alpha; \hat{w}, \hat{\sigma}^2)$ in (18):

$$\begin{aligned} \mathrm{BIC}_{\mathrm{P}}^{(\mathrm{N})}(\beta, v, p) = &(n + p - 1)\log\hat{\sigma}^2 + n\beta \hat{w}^{\mathrm{T}} D_2^{\mathrm{T}} D_2 \hat{w}/\hat{\sigma}^2 + n + (n-3)\log(2\pi) + 3\log n \\ &+ \log|Q_\beta^{(\mathrm{G})}(\hat{w}, \hat{\sigma}^2)| - \log|D_2^{\mathrm{T}} D_2|_+ - (p-1)\log\beta, \end{aligned} \tag{19}$$

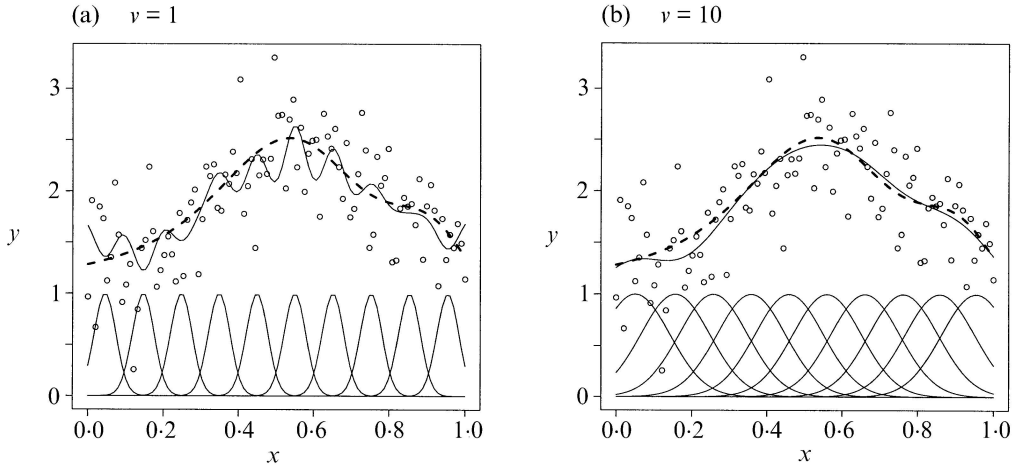

Fig. 1: Example 1. Comparison of the true curve, dashed lines, and the smoothed curve, solid lines, for radial basis functions with $v = 1, 10$.

where

$$Q_\beta^{(G)}(\hat{w}, \hat{\sigma}^2) = \frac{1}{n\hat{\sigma}^2}\begin{pmatrix} B^{\mathrm{T}}B + n\beta D_2^{\mathrm{T}}D_2 & B^{\mathrm{T}}e/\hat{\sigma}^2 \\ e^{\mathrm{T}}B/\hat{\sigma}^2 & n/(2\hat{\sigma}^2) \end{pmatrix}.$$

For illustration, data $\{(y_\alpha, x_\alpha); \alpha = 1, \ldots, 100\}$ were generated from the true model

$$y_\alpha = 0{\cdot}2 \sin(3\pi x_\alpha^2) + \exp\{-(x_\alpha - 0{\cdot}5)^2\} + \exp\{-16(x_\alpha - 0{\cdot}6)^2\} + \varepsilon_\alpha$$

with Gaussian noise $N(0, 0{\cdot}16)$, where the design points are uniformly distributed in $[0, 1]$. Figures 1 (a) and (b) give the fitted curves and the radial basis functions with $v = 1$ and $v = 10$. The fitted curve in Fig. 1 (a) is obviously undersmoothed, while the one in Fig. 1 (b) gives a good representation of the underlying function over the region $[0, 1]$.

Example 2: *Radial basis function network logistic regression model.* Let $y_1, \ldots, y_n$ be independent binary random variables with

$$\mathrm{pr}(Y = 1|x_\alpha) = \pi(x_\alpha), \quad \mathrm{pr}(Y = 0|x_\alpha) = 1 - \pi(x_\alpha),$$

where x_α are q-dimensional explanatory variables. We model $\pi(x_\alpha)$ by

$$\log\left\{\frac{\pi(x_\alpha)}{1 - \pi(x_\alpha)}\right\} = \sum_{j=1}^{p} w_j \phi_j(x_\alpha) + w_0.$$

Estimating the $(p + 1)$-dimensional parameter vector w by maximum penalised likelihood gives the model

$$f_L(y_\alpha|x_\alpha; \hat{w}) = \hat{\pi}(x_\alpha)^{y_\alpha}\{1 - \hat{\pi}(x_\alpha)\}^{1-y_\alpha} \quad (\alpha = 1, \ldots, n), \tag{20}$$

where $\hat{\pi}(x_\alpha) = \exp\{\hat{w}^{\mathrm{T}}b(x_\alpha)\}/[1 + \exp\{\hat{w}^{\mathrm{T}}b(x_\alpha)\}]$ is the estimated conditional probability.

By taking

$$u(\hat{\xi}_\alpha) = \log\{1 + \exp(\hat{\xi}_\alpha)\}, \quad v(y_\alpha, \psi) = 0, \quad h(\hat{\mu}_\alpha) = \log\frac{\hat{\mu}_\alpha}{1 - \hat{\mu}_\alpha}, \quad \psi = 1,$$

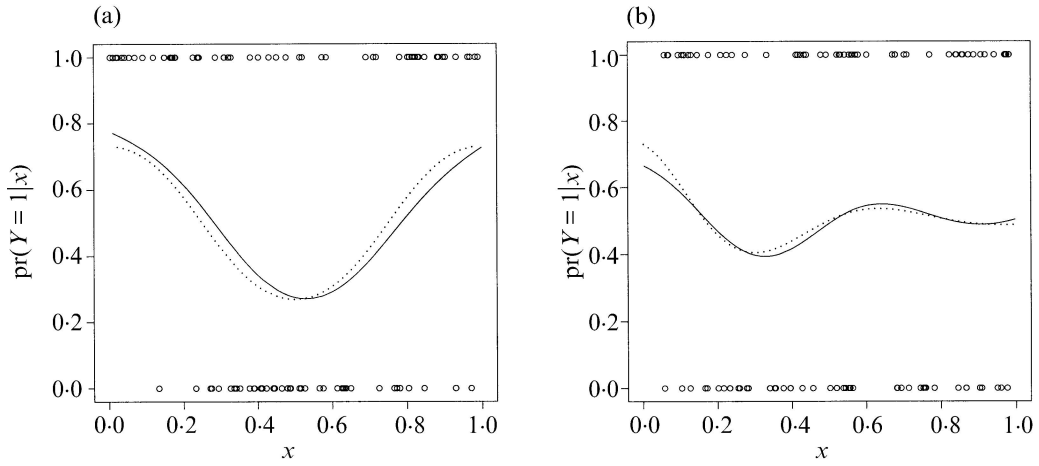

Fig. 2: Example 2. Fitting radial basis function network logistic regression models for the true models
(a) $\mathrm{pr}(Y=1|x)=1/[1+\exp\{-\cos(2\pi x)\}]$,
(b) $\mathrm{pr}(Y=1|x)=1/[1+\exp\{-\exp(-3x)\cos(3\pi x)\}]$.
The broken and solid curves represent the true and estimated conditional probability models, respectively.

in Theorem 2, we obtain

$$\mathrm{BIC}_{\mathrm{P}}^{(\mathrm{L})}(\lambda, v, p) = 2\sum_{\alpha=1}^{n}(\log[1+\exp\{\hat{w}^{\mathrm{T}}b(x_\alpha)\}]-y_\alpha\hat{w}^{\mathrm{T}}b(x_\alpha))+n\lambda\hat{w}^{\mathrm{T}}D_2^{\mathrm{T}}D_2\hat{w}$$
$$-2\log(2\pi/n)+\log|Q_\lambda^{(\mathrm{L})}(\hat{w})|-\log|D_2^{\mathrm{T}}D_2|_+-(p-1)\log\lambda, \qquad (21)$$

where

$$Q_\lambda^{(\mathrm{L})}(\hat{w}) = B^{\mathrm{T}}\Gamma^{(\mathrm{L})}B/n+\lambda D_2^{\mathrm{T}}D_2$$

with $\Gamma_{\alpha\alpha}^{(\mathrm{L})}=\exp\{\hat{w}^{\mathrm{T}}b(x_\alpha)\}/[1+\exp\{\hat{w}^{\mathrm{T}}b(x_\alpha)\}]^2$ as the αth diagonal element of $\Gamma^{(\mathrm{L})}$.

For illustration, binary observations $y_1,\ldots,y_{100}$ were generated from the true models

$$\mathrm{pr}(Y=1|x)=\frac{1}{1+\exp\{-\cos(2\pi x)\}},\quad \mathrm{pr}(Y=1|x)=\frac{1}{1+\exp\{-\exp(-3x)\cos(3\pi x)\}},$$

where the design points are uniformly distributed in $[0,1]$. We fitted the radial basis function network logistic model (20) to these data. Figure 2 shows the true and estimated conditional probability functions; the circles indicate the data.

Example 3: *Radial basis function network Poisson regression model.* Suppose that we have n independent observations y_α, each from a Poisson distribution with conditional expectation $E(Y_\alpha|x_\alpha)=\gamma(x_\alpha)$, where x_α consists of q covariates. It is assumed that the conditional expectation is of the form

$$\log\{\gamma(x_\alpha)\}=\sum_{j=1}^{p}w_j\phi_j(x_\alpha)+w_0=w^{\mathrm{T}}b(x_\alpha)\quad(\alpha=1,2,\ldots,n).$$

We estimate the unknown parameter vector w by maximising the penalised loglikelihood function, and then have the model

$$f_{\mathrm{P}}(y_\alpha|x_\alpha; \hat{w}) = \exp\{-\hat{\gamma}(x_\alpha)\}\hat{\gamma}(x_\alpha)^{y_\alpha}/y_\alpha!, \tag{22}$$

where $\hat{\gamma}(x_\alpha)$ is the estimated conditional expectation.

By taking $u(\hat{\xi}_\alpha) = \exp(\hat{\xi}_\alpha)$, $\psi = 1$, $v(y_\alpha, \psi) = -\log(y_\alpha!)$ and $h(\hat{\mu}_\alpha) = \log(\hat{\mu}_\alpha)$ in Theorem 2, we obtain the model-evaluation criterion

$$\begin{aligned}\mathrm{BIC}_{\mathrm{P}}^{(\mathrm{P})}(\lambda, v, p) = 2\sum_{\alpha=1}^{n}[\exp\{\hat{w}^{\mathrm{T}}b(x_\alpha)\} - y_\alpha\hat{w}^{\mathrm{T}}b(x_\alpha) + \log(y_\alpha!)] + n\lambda\hat{w}^{\mathrm{T}}D_2^{\mathrm{T}}D_2\hat{w} \\ - 2\log(2\pi/n) + \log|Q_\lambda^{(\mathrm{P})}(\hat{w})| - \log|D_2^{\mathrm{T}}D_2|_+ - (p-1)\log\lambda,\end{aligned}$$

where $Q_\lambda^{(\mathrm{P})}(\hat{w}) = B^{\mathrm{T}}\Gamma^{(\mathrm{P})}B/n + \lambda D_2^{\mathrm{T}}D_2$ with $\Gamma_{\alpha\alpha}^{(\mathrm{P})} = \exp\{\hat{w}^{\mathrm{T}}b(x_\alpha)\}$ as the αth diagonal element of $\Gamma^{(\mathrm{P})}$. The adjusted parameters λ, v and p are determined as the minimisers of $\mathrm{BIC}_{\mathrm{P}}^{(\mathrm{P})}(\lambda, v, p)$.

When we set $\lambda = 0$ in (16), the solution becomes the ordinary maximum likelihood estimators. Often the maximum likelihood method yields unstable estimates of weight parameters and so leads to large errors in predicting future observations. Regularisation is essential in Example 2, since some of the maximum likelihood estimates of the weight parameters are often infinite. In our experiments the frequency of convergence was only 395 times out of 1000 repeated Monte Carlo trials for a combination of $(n, p) = (50, 15)$ and 804 times for $(n, p) = (100, 15)$. The frequency of non-convergence increases as the number of basis functions increases.

4. NUMERICAL COMPARISONS

Monte Carlo experiments were conducted to compare the effectiveness of the criteria $\mathrm{BIC}_{\mathrm{P}}^{(\mathrm{N})}$ in (19) and $\mathrm{BIC}_{\mathrm{P}}^{(\mathrm{L})}$ in (21) with AIC-type criteria.

Akaike's information criterion (1973) was derived as an estimator of the Kullback & Leibler (1951) information from the predictive point of view and is given by

$$-2\ell(\hat{\theta}_{\mathrm{ML}}) + 2\,(\text{the number of parameters}),$$

where $\ell(\hat{\theta}_{\mathrm{ML}})$ is the loglikelihood of a model estimated by the maximum likelihood method.

The number of parameters is a measure of the complexity of the model. However, in nonlinear models, especially models estimated by regularisation, the number of parameters is not a suitable measure of model complexity, since the complexity may depend on both the regularisation term and the observed data. The concept of number of parameters was extended to the effective number of parameters by Hastie & Tibshirani (1990, Ch. 3), Moody (1992), Spiegelhalter et al. (2002) and others.

In Example 1, the fitted value $\hat{y}$ is expressed as $\hat{y} = S_\beta y$ for given β, where S_β is the smoother matrix given by $S_\beta = B(B^{\mathrm{T}}B + n\beta D_2^{\mathrm{T}}D_2)^{-1}B^{\mathrm{T}}$. Hastie & Tibshirani (1990) used the trace of the smoother matrix as an approximation to the effective number of parameters. By replacing the number of parameters in AIC and BIC in (7) by $\mathrm{tr}\, S_\beta$, we formally obtain information criteria for the radial basis function network Gaussian regression model (18) in the form

$$\mathrm{AIC}_{\mathrm{M}} = n\log(2\pi\hat{\sigma}^2) + n + 2\,\mathrm{tr}\, S_\beta, \tag{23}$$

$$\mathrm{BIC}_{\mathrm{M}} = n\log(2\pi\hat{\sigma}^2) + n + (\mathrm{tr}\, S_\beta)\log n, \tag{24}$$

where $\hat{\sigma}^2 = \|y - S_\beta y\|^2/n$.

Hurvich et al. (1998) gave an improved version of AIC for choosing a smoothing parameter in various types of nonparametric regression model; see also Sugiura (1978) and Hurvich & Tsai (1989) for Gaussian linear regression and autoregressive time series models. The version for the radial basis function network Gaussian regression model is formally given by

$$\mathrm{AIC_C} = n \log(2\pi\hat{\sigma}^2) + n + 2n(\mathrm{tr}\, S_\beta + 2)/(n - \mathrm{tr}\, S_\beta - 1). \tag{25}$$

An advantage of the criteria $\mathrm{AIC_M}$, $\mathrm{BIC_M}$ and $\mathrm{AIC_C}$ is that they can be applied in an automatic way to each practical situation where there is a smoother matrix. There is however no theoretical justification, since AIC and BIC are criteria for evaluating models estimated by the maximum likelihood method.

Recently, Spiegelhalter et al. (2002) proposed a measure for the effective number of parameters in a model, as the difference between the posterior mean of the deviance and the deviance at the posterior means of the parameters of interest, using an information theoretic argument. They provided a nice review of various types of information theoretic criterion (Akaike, 1973; Takeuchi, 1976; Murata et al., 1994) and the effective number of parameters (Efron, 1986; Wahba, 1990; Hastie & Tibshirani, 1990; Moody, 1992) based on the comparison of their Bayesian complexity measure. Murata et al. (1994) proposed the so-called network information criterion, NIC, for evaluating neural-network models estimated by regularisation. A general theory for constructing information-theoretic criteria based on the Kullback–Leibler (1951) information was given by Konishi & Kitagawa (1996) and Konishi (1999).

We compare the criteria $\mathrm{BIC_P^{(N)}}$ and $\mathrm{BIC_P^{(L)}}$ with the AIC-type criteria $\mathrm{AIC_M}$, $\mathrm{AIC_C}$, NIC and also the BIC-type criterion $\mathrm{BIC_M}$, using Monte Carlo simulations.

Example 4: *Gaussian nonlinear regression models*. In the simulation study, data $\{(x_{1\alpha}, x_{2\alpha}, y_\alpha); \alpha = 1, \ldots, n\}$ were generated from the true model $y_\alpha = m(x_{1\alpha}, x_{2\alpha}) + \varepsilon_\alpha$, where

(a) $m(x_{1\alpha}, x_{2\alpha}) = 0{\cdot}5 + 0{\cdot}001x_{1\alpha} + 1{\cdot}09x_{2\alpha}^2 + 1{\cdot}575x_{1\alpha}x_{2\alpha}$,

(b) $m(x_{1\alpha}, x_{2\alpha}) = \sin(x_{1\alpha} + x_{2\alpha}) \cos(x_{2\alpha})$

and the design points are uniformly distributed in $[-1, 1] \times [-1, 1]$. The errors ε_α are assumed to be independently distributed according to the normal distributions with means 0 and the standard deviations are taken as $\sigma = 0{\cdot}2R_w$ with R_w being the range of $m(.)$ over the input space.

We fitted the model described in Example 1. The adjusted parameters were determined by the Bayesian information criteria $\mathrm{BIC_P^{(N)}}$ in (19), $\mathrm{BIC_M}$ in (24), the AIC-type criteria $\mathrm{AIC_M}$ in (23), $\mathrm{AIC_C}$ in (25) and NIC (Murata et al., 1994).

Tables 1 and 2 compare the average squared errors $\mathrm{ASE} = \sum_{\alpha=1}^{n} \{m(x_\alpha) - \hat{y}_\alpha\}^2/n$ between the true and estimated functions, and the means and standard deviations of the adjusted parameters ν, λ and the number of basis functions. The values in parentheses indicate standard deviations for the means. The simulation results were obtained by averaging over 100 Monte Carlo trials.

Example 5: *Non-Gaussian regression model*. We generated 100 binary observations according to models

(a) $\mathrm{pr}(Y = 1|x) = 1/[1 + \exp\{-0{\cdot}3 \exp(x_1) \cos(\pi x_2) + 0{\cdot}3\}]$,

(b) $\mathrm{pr}(Y = 1|x) = 1/[1 + \exp\{-0{\cdot}3 \exp(x_1 + x_2) + 0{\cdot}8\}]$,

Table 1: *Example* 4. *Comparison of the average squared errors*, ASE, *for true function*

$$m(x) = 0{\cdot}5 + 0{\cdot}001x_1 + 1{\cdot}09x_2^2 + 1{\cdot}575x_1x_2,$$

based on various criteria and using 100 *simulated datasets. Figures in parentheses give estimated standard deviations*

		$\mathrm{BIC}_{\mathrm{P}}^{(\mathrm{N})}$	$\mathrm{BIC}_{\mathrm{M}}$	$\mathrm{AIC}_{\mathrm{C}}$	$\mathrm{AIC}_{\mathrm{M}}$	NIC
$n = 100$	ASE	0·06155	0·06227	0·06283	0·07909	0·08712
		(0·0196)	(0·0208)	(0·0210)	(0·0307)	(0·0352)
	p	16·32	18·59	20·50	24·34	24·84
		(1·548)	(3·028)	(4·382)	(5·793)	(5·515)
	$\log(\lambda)$	−2·124	−2·039	−2·846	−3·282	−3·659
		(0·658)	(0·791)	(0·941)	(0·866)	(0·730)
	v	18·15	20·18	18·40	15·84	14·68
		(5·52)	(4·66)	(5·67)	(6·48)	(6·91)
$n = 200$	ASE	0·02579	0·02691	0·02851	0·02902	0·02858
		(0·0071)	(0·0086)	(0·0087)	(0·0088)	(0·0085)
	p	16·08	18·82	22·26	24·08	24·12
		(1·383)	(4·250)	(5·609)	(6·505)	(6·573)
	$\log(\lambda)$	−2·293	−2·389	−3·250	−3·386	−3·702
		(0·694)	(0·898)	(0·971)	(0·977)	(0·906)
	v	18·64	19·96	18·49	18·49	18·00
		(4·87)	(4·92)	(5·45)	(5·51)	(5·34)

Table 2: *Example* 4. *Comparison of the average squared errors*, ASE, *for true function*

$$m(x) = \sin(x_1 + x_2)\cos(x_2),$$

based on various criteria and using 100 *simulated datasets. Figures in parentheses give estimated standard deviations*

		$\mathrm{BIC}_{\mathrm{P}}^{(\mathrm{N})}$	$\mathrm{BIC}_{\mathrm{M}}$	$\mathrm{AIC}_{\mathrm{C}}$	$\mathrm{AIC}_{\mathrm{M}}$	NIC
$n = 100$	ASE	0·01502	0·01526	0·01568	0·01616	0·01774
		(0·0067)	(0·0065)	(0·0068)	(0·0077)	(0·0086)
	p	16·84	16·76	18·45	18·70	19·20
		(0·687)	(0·843)	(1·365)	(1·462)	(1·505)
	$\log(\lambda)$	−0·688	−0·246	−1·147	−1·178	−2·739
		(0·611)	(0·493)	(1·003)	(1·114)	(1·753)
	v	16·98	18·96	17·34	16·27	15·32
		(6·33)	(5·02)	(6·03)	(6·30)	(6·69)
$n = 200$	ASE	0·00776	0·00798	0·00836	0·00860	0·00872
		(0·0035)	(0·0036)	(0·0038)	(0·0039)	(0·0042)
	p	16·92	16·84	18·65	18·82	19·24
		(0·680)	(0·823)	(1·258)	(1·332)	(1·405)
	$\log(\lambda)$	−0·708	−0·335	−1·176	−1·186	−2·977
		(0·599)	(0·466)	(1·011)	(1·028)	(1·844)
	v	20·20	20·40	18·69	17·22	17·56
		(4·30)	(3·87)	(4·54)	(5·29)	(5·08)

Table 3: *Example* 5. *Comparison of the average squared errors*, ASE, *for true model*

$$\mathrm{pr}(Y=1|x)=1/[1+\exp\{-0{\cdot}3\exp(x_1)\cos(\pi x_2)+0{\cdot}3\}],$$

based on various criteria and using 100 *simulated datasets. Figures in parentheses give estimated standard deviations*

		$\mathrm{BIC}_{\mathrm{P}}^{(\mathrm{L})}$	$\mathrm{BIC}_{\mathrm{M}}$	$\mathrm{AIC}_{\mathrm{C}}$	$\mathrm{AIC}_{\mathrm{M}}$	NIC
$n=100$	ASE	0·01203	0·01199	0·01446	0·01604	0·01790
		(0·0076)	(0·0075)	(0·0090)	(0·0104)	(0·0117)
	p	10·24	9·48	10·65	11·14	10·75
		(2·140)	(1·769)	(2·306)	(2·495)	(2·367)
	$\log(\lambda)$	−0·347	−0·145	−1·101	−1·507	−1·546
		(0·631)	(0·345)	(1·747)	(2·110)	(2·183)
	v	21·40	25·92	20·88	19·62	21·00
		(9·57)	(6·72)	(9·25)	(9·18)	(9·31)
$n=200$	ASE	0·00786	0·00788	0·00864	0·00886	0·00950
		(0·0036)	(0·0033)	(0·0045)	(0·0046)	(0·0048)
	p	11·00	9·41	11·47	11·58	11·51
		(2·420)	(1·892)	(2·492)	(2·458)	(2·439)
	$\log(\lambda)$	−0·433	−0·193	−0·806	−0·913	−0·766
		(0·571)	(0·284)	(1·039)	(1·123)	(1·384)
	v	24·04	26·12	19·88	20·28	21·96
		(8·70)	(7·11)	(9·01)	(9·06)	(8·84)

Table 4: *Example* 5. *Comparison of the average squared errors*, ASE, *for true model*

$$\mathrm{pr}(Y=1|x)=1/[1+\exp\{-0{\cdot}3\exp(x_1+x_2)+0{\cdot}8\}],$$

based on various criteria and using 100 *simulated datasets. Figures in parentheses give estimated standard deviations*

		$\mathrm{BIC}_{\mathrm{P}}^{(\mathrm{L})}$	$\mathrm{BIC}_{\mathrm{M}}$	$\mathrm{AIC}_{\mathrm{C}}$	$\mathrm{AIC}_{\mathrm{M}}$	NIC
$n=100$	ASE	0·01329	0·01375	0·01485	0·01638	0·01664
		(0·0080)	(0·0076)	(0·0087)	(0·093)	(0·0116)
	p	10·18	10·08	10·85	11·08	11·44
		(2·104)	(2·012)	(2·219)	(2·235)	(2·435)
	$\log(\lambda)$	−1·355	−1·234	−2·820	−2·532	−2·943
		(0·074)	(0·065)	(1·261)	(1·355)	(1·844)
	v	22·44	23·20	22·84	18·42	19·66
		(8·63)	(8·31)	(9·45)	(9·09)	(9·61)
$n=200$	ASE	0·00459	0·00470	0·00583	0·00584	0·00639
		(0·0026)	(0·0027)	(0·0039)	(0·0043)	(0·0048)
	p	10·62	9·98	11·08	11·66	12·10
		(2·026)	(1·952)	(2·174)	(2·186)	(2·383)
	$\log(\lambda)$	−1·386	−1·265	−2·873	−2·519	−2·886
		(0·450)	(0·335)	(0·785)	(0·799)	(1·022)
	v	23·04	23·32	23·36	18·54	19·16
		(9·04)	(8·44)	(9·05)	(9·15)	(9·34)

where the design points x_α are uniformly distributed in $[-1, 1] \times [-1, 1]$. We constructed the radial basis function network logistic regression model (20), using the criterion $\mathrm{BIC}_{\mathrm{P}}^{(\mathrm{L})}$ in (21), $\mathrm{BIC}_{\mathrm{M}}$ in (24), the AIC-type criteria $\mathrm{AIC}_{\mathrm{M}}$ in (23), $\mathrm{AIC}_{\mathrm{C}}$ in (25) and NIC. The smoother matrix is given by $S_\lambda = B(B^{\mathrm{T}}WB + n\lambda D_2^{\mathrm{T}} D_2)^{-1} B^{\mathrm{T}} W$, where W is an $n \times n$ diagonal matrix with ith diagonal element $w_{ii} = \hat{\pi}(x_\alpha)\{1 - \hat{\pi}(x_\alpha)\}$.

Tables 3 and 4 compare the average squared errors between the true and estimated conditional probabilities, the means and standard deviations of the adjusted parameters λ, v and the number of basis functions. The values in parentheses indicate standard deviations for the means. The simulated results were obtained by averaging over 100 Monte Carlo trials.

It may be seen from the simulation results in both Example 4 and Example 5 that the models evaluated by BIC-type criteria are superior to those based on AIC-type criteria in almost all cases; they give smaller mean values with smaller variances for ASE. The standard deviations of λ determined by $\mathrm{BIC}_{\mathrm{P}}^{(\mathrm{N})}$ are smaller than the others in Gaussian regression models, while those determined by $\mathrm{BIC}_{\mathrm{M}}$ are smaller in non-Gaussian regression models. Criteria of the BIC-type tend to choose fewer basis functions and larger values of λ than those based on AIC-type criteria. It appears that AIC-type criteria are generally more variable and more likely to undersmooth than BIC-type criteria.

Example 6: *Robot arm data.* Andrieu et al. (2001) proposed a hierarchical full Bayesian model for radial basis function networks with Gaussian noise, in which the model dimension, model parameters, regularisation parameters and also noise parameters are treated as unknown random variables. They developed a reversible-jump Markov chain Monte Carlo simulation algorithm for radial basis networks for computing the joint posterior distribution of the parameters. Our method can be regarded as an approximate Bayesian methodology, and we compare the Gaussian version of our approach with the full Bayesian approach, by analysing the robot arm dataset which is often used as a benchmark dataset in the neural network literature (Andrieu et al., 2001; Holmes & Mallick, 1998; MacKay, 1992; Neal, 1996; Rios Insua & Müller, 1998). MacKay (1992) originally introduced the use of the Bayesian approach in the neural network literature. The dataset, created by D. J. C. MacKay and available at http://wol.ra.phy.cam.ac.uk/mackay/bigback/dat/, is a set of four-dimensional data $\{(x_{1\alpha}, x_{2\alpha}, y_{1\alpha}, y_{2\alpha}); \alpha = 1, \ldots, n\}$ generated from the following model:

$$y_{1\alpha} = 2\cos(x_{1\alpha}) + 1{\cdot}3\cos(x_{1\alpha} + x_{2\alpha}) + \varepsilon_{1\alpha}, \quad y_{2\alpha} = 2\sin(x_{1\alpha}) + 1{\cdot}3\sin(x_{1\alpha} + x_{2\alpha}) + \varepsilon_{2\alpha},$$

where $\varepsilon_{1\alpha}$ and $\varepsilon_{2\alpha}$ are normal noise variables with means 0 and variances $(0{\cdot}05)^2$.

The first 200 observations are used to estimate the model, and the last 200 observations are used to evalute the prediction accuracy. We fitted the radial basis function network Gaussian nonlinear regression model given in Example 1. The values of λ, v and p are chosen as the minimisers of $\mathrm{BIC}_{\mathrm{P}}^{(\mathrm{N})}$. As a result, we obtained $\hat{p} = 20$, $\hat{v} = 31{\cdot}81$ and $\hat{\lambda} = 2{\cdot}46 \times 10^{-7}$, and the corresponding average squared error was 0·00509.

Table 5 summarises the results obtained by various techniques. Our strategy and the full Bayesian approach yield almost the same results, and both give fitted functions that capture the true structure. An advantage of our procedure is that it is easily implemented in both its Gaussian and non-Gaussian versions.

The smoothness of a fitted curve or surface is mainly controlled by the hyperparameter v, and the regularisation parameter λ has the effect of reducing the variance of the weight parameters. Note that the Bayesian information criterion $\mathrm{BIC}_{\mathrm{P}}$ is not restricted to linear

Table 5: *Example* 6. *Comparison of the average squared errors*, ASE, *for the robot arm data. The results, except for our modelling strategy, are drawn from Andrieu et al.* (2001), *Holmes & Mallick* (1998), *MacKay* (1992), *Neal* (1996) *and Rios Insua & Müller* (1998)

Methods	ASE
MacKay's (1992) Gaussian approximation with highest evidence	0·00573
MacKay's (1992) Gaussian approximation with lowest test error	0·00557
Neal's (1996) hybrid Monte Carlo	0·00554
Neal's (1996) hybrid Monte Carlo with ARD	0·00549
Rios Insua & Müller's (1998) MLP with reversible-jump MCMC	0·00620
Holmes & Mallick's (1998) RBF with reversible-jump MCMC	0·00535
Andrieu et al.'s reversible-jump MCMC with Bayesian model	0·00502
Andrieu et al.'s reversible-jump MCMC with MDL	0·00512
Andrieu et al.'s reversible-jump MCMC with AIC	0·00520
Proposed modelling strategy	0·00509

ARD, automatic relevance determination; MLP, multilayer perceptron; MCMC, Markov chain Monte Carlo; RBF, radial basis function; MDL, minimum description length

estimators of regression functions, but may be applied to other nonlinear models such as multilayer perceptron neural networks. We conclude from the numerical studies that $\text{BIC}_{\text{P}}^{(\text{N})}$ and $\text{BIC}_{\text{P}}^{(\text{L})}$ perform well in practical situations.

5. Discussion

The criteria AIC and BIC have been widely used for variable selection, mainly in linear models such as autoregressive time series models. In this paper we concentrate on selection of smoothing parameters in nonlinear models. The proposed criterion may be used for selecting an optimal subset of variables in nonlinear modelling; optimal values of smoothing parameters are obtained as the minimisers of the criterion for each model, and then we choose a statistical model for which the value of the criterion is minimised over a set of competing models. Further work remains to be done towards constructing nonlinear modelling strategies of this nature in the context of areas such as neural network models, which are characterised by a large number of parameters.

Acknowledgement

The authors would like to thank the editor and anonymous reviewers for constructive and helpful comments that improved the quality of the paper considerably. We are also grateful to them for pointing out Andrieu et al. (2001), Spiegelhalter et al. (2002) and several other references.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd Int. Symp. on Information Theory*, Ed. B. N. Petrov and F. Csaki, pp. 267–81. Budapest: Akademiai Kiado.

Ando, T., Imoto, S. & Konishi, S. (2001). Estimating nonlinear regression models based on radial basis function networks (in Japanese). *Jap. J. Appl. Statist.* **30**, 19–35.

ANDRIEU, C., DE FREITAS, N. & DOUCET, A. (2001). Robust full Bayesian learning for radial basis networks. *Neural Comp.* **13**, 2359–407.
BISHOP, C. M. (1995). *Neural Networks for Pattern Recognition.* Oxford University Press.
BROOMHEAD, D. S. & LOWE, D. (1988). Multivariable functional interpolation and adaptive networks. *Complex Syst.* **2**, 321–35.
CLARKE, B. S. & BARRON, A. R. (1994). Jeffreys' proir is asymptotically least favorable under entropy risk. *J. Statist. Plan. Infer.* **41**, 37–60.
DAVISON, A. C. (1986). Approximate predictive likelihood. *Biometrika* **73**, 323–32.
EFRON, B. (1986). How biased is the apparent error rate of a prediction rule? *J. Am. Statist. Assoc.* **81**, 461–70.
GOOD, I. J. & GASKINS, R. A. (1971). Nonparametric roughness penalites for probability densities. *Biometrika* **58**, 255–77.
GREEN, P. J. & SILVERMAN, B. W. (1994). *Nonparametric Regression and Generalized Linear Models.* London: Chapman and Hall.
GREEN, P. J. & YANDELL, B. (1985). Semi-parametric generalized linear models. In *Generalized Linear Models*, Ed. R. Gilchrist, B. J. Francis and J. Whittaker, Lecture Notes in Statistics, **32**, pp. 44–55. Berlin: Springer.
HASTIE, T. J. & TIBSHIRANI, R. J. (1990). *Generalized Additive Models.* London: Chapman and Hall.
HOLMES, C. C. & MALLICK, B. K. (1998). Bayesian radial basis functions of variable dimension. *Neural Comp.* **10**, 1217–33.
HURVICH, C. M. & TSAI, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika* **76**, 297–307.
HURVICH, C. M., SIMONOFF, J. S. & TSAI, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *J. R. Statist. Soc.* B **60**, 271–93.
KASS, R. E. & RAFTERY, A. E. (1995). Bayes factors. *J. Am. Statist. Assoc.* **90**, 773–95.
KASS, R. E. & WASSERMAN, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. Am. Statist. Assoc.* **90**, 928–34.
KASS, R. E., TIERNEY, L. & KADANE, J. B. (1990). The validity of posterior expansions based on Laplace's method. In *Essays in Honor of George Barnard*, Ed. S. Geisser, J. S. Hodges, S. J. Press and A. Zellner, pp. 473–88. Amsterdam: North-Holland.
KONISHI, S. (1999). Statistical model evaluation and information criteria. In *Multivariate Analysis, Design of Experiments and Surrey Sampling*, Ed. S. Ghosh, pp. 369–99. New York: Marcel Dekker.
KONISHI, S. & KITAGAWA, G. (1996). Generalised information criteria in model selection. *Biometrika* **83**, 875–90.
KULLBACK, S. & LEIBLER, R. A. (1951). On information and sufficiency. *Ann. Math. Statist.* **22**, 79–86.
LANTERMAN, A. D. (2001). Schwarz, Wallace, and Rissanen: Intertwining themes in theories of model selection. *Int. Statist. Rev.* **69**, 185–212.
MACKAY, D. J. C. (1992). A practical Bayesian framework for backpropagation networks. *Neural Comp.* **4**, 448–72.
MCCULLAGH, P. & NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. London: Chapman and Hall.
MOODY, J. (1992). The *effective* number of parameters: an analysis of generalization and regularization in nonlinear learning systems. In *Advances in Neural Information Processing Systems 4*, Ed. J. E. Moody, S. J. Hanson and R. P. Lippmann, pp. 847–54. San Mateo, CA: Morgan Kaufmann.
MOODY, J. & DARKEN, C. J. (1989). Fast learning in networks of locally-tuned processing units. *Neural Comp.* **1**, 281–94.
MURATA, N., YOSHIZAWA, S. & AMARI, S. (1994). Network information criterion-determining the number of hidden units for an artificial neural network model. *IEEE Trans. Neural Networks* **5**, 865–72.
NEAL, R. M. (1996). *Bayesian Learning for Neural Networks*, Lecture Notes in Statistics 118. New York: Springer-Verlag.
NEATH, A. A. & CAVANAUGH, J. E. (1997). Regression and time series model selection using variants of the Schwarz information criterion. *Commun. Statist.* A **26**, 559–80.
NELDER, J. A. & WEDDERBURN, R. W. M. (1972). Generalized linear models. *J. R. Statist. Soc.* A **135**, 370–84.
O'HAGAN, A. (1995). Fractional Bayes factors for model comparison (with Discussion). *J. R. Statist. Soc.* B **57**, 99–138.
O'SULLIVAN, F., YANDELL, B. S. & RAYNOR, W. J. (1986). Automatic smoothing of regression functions in generalized linear models. *J. Am. Statist. Assoc.* **81**, 96–103.
PAULER, D. (1998). The Schwarz criterion and related methods for normal linear models. *Biometrika* **85**, 13–27.
POGGIO, T. & GIROSI, F. (1990). Networks for approximation and learning. *Proc. IEEE* **78**, 1484–7.
RIOS INSUA, D. & MÜLLER, P. (1998). Feedforward neural networks for nonparametric regression. In *Practical Nonparametric and Semiparametric Bayesian Statistics*, Ed. D. K. Dey, P. Müller and D. Sinha, pp. 181–91. New York: Springer Verlag.
RIPLEY, B. D. (1996). *Pattern Recognition and Neural Networks.* Cambridge University Press.
SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461–4.
SILVERMAN, B. W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with Discussion). *J. R. Statist. Soc.* B **47**, 1–52.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & van der Linde, A. (2002). Bayesian measures of model complexity and fit (with Discussion). *J. R. Statist. Soc.* B **64**, 583–639.
Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Commun. Statist.* A **7**, 13–26.
Takeuchi, K. (1976). Distributions of information statistics and criteria for adequacy of models (in Japanese). *Math. Sci.* **153**, 12–8.
Tanabe, K. & Tanaka, T. (1983). Estimation of curve and surface by Bayesian model (in Japanese). *Chikyu* **5**, 179–86.
Tierney, L. & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Am. Statist. Assoc.* **81**, 82–6.
Tierney, L., Kass, R. E. & Kadane, J. B. (1989). Fully exponential Laplace approximations to expectations and variances of nonpositive functions. *J. Am. Statist. Assoc.* **84**, 710–6.
Wahba, G. (1990). *Spline Models for Observational Data.* Philadelphia: Society for Industrial and Applied Mathematics.
Webb, A. (1999). *Statistial Pattern Recognition.* London: Arnold.
Whittaker, E. (1923). On a new method of graduation. *Proc. Edin. Math. Soc.* **41**, 63–75.

[*Received December* 2001. *Revised April* 2003]