Josai Mathematical Monographs
vol. 8 (2015), pp. 97-113

Predictive model selection criteria for relevance vector regression models

Kazuki MATSUDA

Abstract. We focus on a selection of kernel parameters in the framework of the relevance vector machine (RVM) for regression, called the relevance vector regression (RVR). The RVR can achieve a sparse model and utilize a kernel function similar to the support vector regression (SVR). A crucial issue in the model building process of the RVR is the selection of the optimal values for kernel parameters. In this paper, we derive a model selection criterion for evaluating the Bayesian predictive distribution of the RVR model from information-theoretic viewpoint. Monte Carlo experiments and real data analysis have been presented to demonstrate that the proposed modeling procedure performs well.

1. Introduction

In recent years, nonlinear regression modeling based on basis expansions provides an efficient tool for analyzing the data with complex structure and has been widely used in various fields of natural and social sciences (see, e.g., Bishop, 2006; Konishi and Kitagawa, 2008; Hastie *et al.*, 2009). The essential idea behind basis expansions is to express an unknown regression function through the linear combination of known nonlinear functions, called basis functions. According to the structure of data, various basis functions have been proposed; e.g., natural cubic splines (Green and Silverman, 1994), B-splines (Eilers and Marx, 1996; de Boor, 2001; Imoto and Konishi, 2003), radial basis functions (Bishop, 1995; Ripley, 1996; Kawano and Konishi, 2007; Ando *et al.*, 2008; Hastie *et al.*, 2009) and thin plate splines (Girosi *et al.*, 1995). These regression models are characterized by a large number of parameters to be estimated. However, maximum likelihood and least squares methods often yield unstable estimated models. In order to overcome this drawback, regularization methods and Bayesian approach have been widely used for the model estimation (see, e.g., Denison *et al.*, 2002; Figueiredo, 2003; Bishop, 2006).

The relevance vector regression (RVR; Tipping, 2000; Tipping, 2001) connects the strength of Bayesian approach and kernel-based methods to construct the regression model based on the relevance vector machine (RVM), whose form is similar to the support vector regression (SVR; Vapnik, 1995; Vapnik, 1998). The RVR can achieve a sparse model with good generalization capability, avoid over-fitting for the observed data and utilize a wide variety of kernel functions comparable to the SVR. In the RVR, the specific Bayesian approach with automatic relevance determination (ARD; Neal, 1996) leads to estimate the coefficient parameters automatically and sparsely. Moreover, the RVR has no need to determinate the trade-off parameter in the SVR. There are many successful examples by the RVR in various fields (see, e.g., Liu *et al.*, 2006; Cheng *et al.*, 2013; Bai *et al.*, 2014).

A crucial issue in the model building process based on the RVR is the selection of the optimal values for kernel parameters. Tripathi and Govindaraju (2007) noted the significance of this determination by using the bias-variance method. In order to overcome this problem, Tipping (2001) presented the method which maximizes the marginal likelihood of the model respect to the kernel parameters. It is, however, known that this method has the difficulty of the implementation, so the cross-validation (CV; Stone, 1974) method is suggested (Tipping, 2001; Quinonero-Candela and Hansen, 2002). The CV criterion was presented to evaluate the goodness of statistical models from a predictive point of view by separating the data into the training data and the test data. However, this method is known as a computationally expensive model selection procedure because of data separation. In order to establish a more effective criterion for the CV, Craven and Wahba (1979) introduced the generalized cross-validation (GCV) using the hat matrix of the estimated model. For this determination problem, Tripathi and Govindaraju (2007) suggested the Bayesian information criterion (BIC; Schwarz, 1978) as a stable model selection criterion. The BIC is a popular evaluation criterion for statistical models based on a Bayesian posterior probability. The Akaike information criterion (AIC; Akaike, 1973; Akaike, 1974) is also popular as a model selection criterion which is asymptotically consistent estimator of the Kullback-Leibler divergence (Kullback and Leibler, 1951) between a statistical model and an unknown true model. Hasite and Tibshirani (1990) presented the modified Akaike information criterion (mAIC) by replacing the number of the parameters with the trace of the hat matrix. However, the BIC and mAIC are derived as the model evaluation criteria for the statistical models estimated by the maximum likelihood methods (MLE) and thought of as the imperfect criteria for the RVR model theoretically.

Kitagawa (1997) introduced a predictive information criterion (PIC) for evaluating the goodness of Bayesian predictive distributions for Gaussian regression models, which was derived as an estimator of Kullback-Leibler information. The PIC has been applied to evaluate various statistical models; e.g., Bayesian regression models with unknown variance (Kim *et al.*, 2012) and Bayesian lasso by Park and Casella (2008) (Kawano *et al.*, 2014). However, there have been no researches on the selection of kernel parameters of the RVR models by the PIC. In this paper, we derive an information criterion to evaluate the Bayesian predictive distribution of the RVR. We select the optimal values of the kernel parameters that minimize our model selection criterion. Monte Carlo simulations and real data analysis are conducted to examine the performances of our model evaluation method in various situations.

The remainder of this paper is organized as follows. Section 2 describes a framework of the RVR models. In Section 3, we derive a model selection criterion for Bayesian predictive distributions of the RVR. Section 4 investigates the performances of the proposed criterion through Monte Carlo simulations and real data analysis. Some concluding remarks are given in Section 5.

2. Relevance vector regression models

Suppose that we have n observations $\{(y_i, \boldsymbol{x}_i); i = 1, \cdots, n\}$, where $\boldsymbol{x}_i \in \mathbb{R}^p$ are p-dimensional vectors of explanatory variables and $y_i \in \mathbb{R}$ are response variables. Here, we assume that the observed y_i are independently sampled from the regression model with noise ε_i as follows:

$$y_i = u(\boldsymbol{x}_i) + \varepsilon_i, \quad i = 1, \cdots, n,$$

where $u(\cdot)$ is an unknown regression function and ε_i are assumed to be the Gaussian noise with mean zero and variance σ^2. According to the basis expansion method, it is assumed that the regression function $u(\cdot)$ can be expressed as a linear combination of kernel functions as follows:

$$u(\boldsymbol{x}; \boldsymbol{w}) = \sum_{j=1}^{n} w_j K(\boldsymbol{x}, \boldsymbol{x}_j), \tag{1}$$

where $K(\cdot,\cdot)$ is a basis function defined for each observation and $\boldsymbol{w} = (w_1, \cdots, w_n)^T$ is an unknown coefficient vector. In the RVR, various kernel functions have been used as a basis function $K(\cdot,\cdot)$; e.g., Gaussian kernel ($K(\boldsymbol{x}, \boldsymbol{x}') = \exp\{-||\boldsymbol{x} - \boldsymbol{x}'||^2/2h^2\}$), polynomial kernel ($K(\boldsymbol{x}, \boldsymbol{x}') = (\boldsymbol{x}^T\boldsymbol{x}' + c)^d$) and sigmoid kernel ($K(\boldsymbol{x}, \boldsymbol{x}') = \tanh(b\boldsymbol{x}^T\boldsymbol{x}' + c)$). These kernel functions are characterized by some adjusted parameters to be optimized. The optimization of the adjusted parameters can be viewed as a model selection problem. For the detail of kernel functions, we refer to Scholkopf and Smola (2001) and Bishop (2006).

Since the model (1) and the assumption of Gaussian noise ε_i, we have the Gaussian density over y_i with mean $u(\boldsymbol{x}_i; \boldsymbol{w})$ and variance σ^2, namely $N(y_i; u(\boldsymbol{x}_i; \boldsymbol{w}), \sigma^2)$. For convenience, a hyperparameter β is defined as $\beta = 1/\sigma^2$. Thus, the likelihood function can be expressed as

$$p(\boldsymbol{y}|\boldsymbol{w}, \beta) = \left(\frac{\beta}{2\pi}\right)^{\frac{n}{2}} \exp\left\{ -\frac{\beta}{2}(\boldsymbol{y} - \Phi\boldsymbol{w})^T(\boldsymbol{y} - \Phi\boldsymbol{w})\right\}, \tag{2}$$

where $\boldsymbol{y} = (y_1, \cdots, y_n)^T$ and $\Phi = (\boldsymbol{\phi}(\boldsymbol{x}_1), \cdots, \boldsymbol{\phi}(\boldsymbol{x}_n))^T$ is the design matrix with $\boldsymbol{\phi}(\boldsymbol{x}_j) = (K(\boldsymbol{x}_j, \boldsymbol{x}_1), \cdots, K(\boldsymbol{x}_j, \boldsymbol{x}_n))^T$ $(j = 1, \cdots, n)$.

Next, we suppose the Gaussian prior distribution over each coefficient w_j with zero-mean and variance α_j^{-1}. Therefore the prior distribution over $\boldsymbol{w}$ is given by

$$p(\boldsymbol{w}|\boldsymbol{\alpha}) = \prod_{j=1}^{n} N(w_j; 0, \alpha_j^{-1}), \tag{3}$$

where $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_n)^T$ is an n-dimensional hyperparameter vector and α_j is an individual parameter associated independently with each coefficient w_j. Since the likelihood function in (2) and the prior distribution over $\boldsymbol{w}$ in (3) are both Gaussian distributions, the posterior distribution over $\boldsymbol{w}$ can also be obtained by Bayes' theorem as Gaussian distribution

$$p(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{\alpha}, \beta) = N_n(\boldsymbol{w}; \boldsymbol{\mu}, \Sigma),$$

where the n-dimensional posterior mean vector $\boldsymbol{\mu}$ and $n \times n$ covariance matrix Σ are respectively given by

$$\boldsymbol{\mu} = \beta \Sigma \Phi^T \boldsymbol{y}, \quad \Sigma = (A + \beta \Phi^T \Phi)^{-1},$$

with $A = \mathrm{diag}(\alpha_1, \cdots, \alpha_n)$. The estimated value of the coefficient vector $\boldsymbol{w}$ is given by the posterior mean $\boldsymbol{\mu}$ which is the maximum a posterior (MAP) estimator of $\boldsymbol{w}$. We see that $\boldsymbol{\mu}$ and Σ depend on the value of hyperparameters $\boldsymbol{\alpha}$ and β. Thus, the hyperparameters are needed to be optimized.

In the RVR model building process, the optimal values of hyperparameters $\boldsymbol{\alpha}$ and β are determined by using Bayesian evidence procedure which is maximization of marginal likelihood or type-II maximum likelihood method (Tipping, 2001). By integrating out the coefficients $\boldsymbol{w}$, the marginal likelihood is computed as

$$p(\boldsymbol{y}|\boldsymbol{\alpha}, \beta) = \int p(\boldsymbol{y}|\boldsymbol{w}, \beta) p(\boldsymbol{w}|\boldsymbol{\alpha}) \mathrm{d}\boldsymbol{w} = N_n(\boldsymbol{y}; \boldsymbol{0}, C), \tag{4}$$

where $C = \beta^{-1} I_n + \Phi A^{-1} \Phi^T$ and I_n is an $n \times n$ identity matrix. Here, we define the estimated parameters $\boldsymbol{\alpha}$ and β as $\hat{\boldsymbol{\alpha}}$ and $\hat{\beta}$. Setting the derivatives of the marginal likelihood (4) with respect to $\boldsymbol{\alpha}$ and β to zero, we obtain the following estimated formulae

$$\hat{\alpha}_j = \frac{\gamma_j}{\mu_j^2} \ (j = 1, \cdots, n), \quad \hat{\beta}^{-1} = \frac{(\boldsymbol{y} - \Phi \boldsymbol{\mu})^T (\boldsymbol{y} - \Phi \boldsymbol{\mu})}{n - \sum_k \gamma_k}, \tag{5}$$

where μ_j is the j-th element of the posterior mean $\boldsymbol{\mu}$, and $\gamma_j = 1 - \alpha_j \Sigma_{jj}$. Here Σ_{jj} is the j-th diagonal element of the posterior covariance matrix Σ. With the mutuality of formulae (5), we need to re-estimate and set the initialization of hyperparameters.

For this optimization, Tipping and Faul (2003) proposed a more fast algorithm, called the *Sequential Sparse Bayesian Learning Algorithm*, as follows.

```
Algorithm 1: A Sequential Sparse Bayesian Learning Algorithm
1. Initialize a $\beta$.
2. Initialize a single $\alpha_j$ with basis $\boldsymbol{\phi}(\boldsymbol{x}_j)$ by (5) and others are set to infinity.
3. Compute $\boldsymbol{\mu}$ and $\Sigma$, along with the following parameters $s_j$ and $q_j$ given by
       $$s_j = \frac{\alpha_j S_j}{\alpha_j - S_j}, \quad q_j = \frac{\alpha_j Q_j}{\alpha_j - S_j},$$
   where $S_j = \boldsymbol{\phi}(\boldsymbol{x}_j)^T C^{-1} \boldsymbol{\phi}(\boldsymbol{x}_j)$ and $Q_j = \boldsymbol{\phi}(\boldsymbol{x}_j)^T C^{-1} \boldsymbol{y}$.
4. Select an index $k$ from the set of $\{1, 2, \cdots, n\}$.
5. Compute $\theta_k = q_k^2 - s_k$.
   5-1. If $\theta_k > 0$ and $\alpha_k < \infty$ (i.e., $\boldsymbol{\phi}(\boldsymbol{x}_k)$ is in the model), re-estimate $\alpha_k$.
   5-2. If $\theta_k > 0$ and $\alpha_k = \infty$, add $\boldsymbol{\phi}(\boldsymbol{x}_k)$ to the model and compute $\alpha_k$.
   5-3. If $\theta_k \leqslant 0$ and $\alpha_k < \infty$, remove $\boldsymbol{\phi}(\boldsymbol{x}_k)$ from the model and $\alpha_k$ set to infinity.
6. Update the $\beta$.
7. Recompute $\Sigma$, $\boldsymbol{\mu}$, all $s_j$ and $q_j$.
8. Repeat steps 4.-7. until convergence.
```

Through this optimization, if the hyperparameter α_k is estimated to be infinity, the corresponding regression coefficient w_k can be considered to be exactly zero because of the form of the prior distribution (3). Most of the regression coefficients are typically estimated to be zero and corresponding basis functions are removed from the model. Consequently, the sparse model is built based on a few basis functions, called relevance vectors (RVs).

3. Model selection criterion

In the model building process based on the RVR, a crucial issue is the selection of the optimal values for kernel parameters. The selection can be viewed as a model selection and evaluation problem. In this section, in order to overcome this problem more effectively, we derive a model selection criterion to evaluate the Bayesian predictive distribution of the RVR.

3.1. Bayesian predictive distribution

With estimated values $\hat{\boldsymbol{\alpha}}$ and $\hat{\beta}$ by the maximization of the marginal likelihood, the Bayesian predictive distribution in the framework of the RVR models is given by

$$h(\boldsymbol{z}|\boldsymbol{y}, \hat{\boldsymbol{\alpha}}, \hat{\beta}) = \int p(\boldsymbol{z}|\boldsymbol{w}, \hat{\beta}) p(\boldsymbol{w}|\boldsymbol{y}, \hat{\boldsymbol{\alpha}}, \hat{\beta}) \mathrm{d}\boldsymbol{w} = N_n(\boldsymbol{z}; \boldsymbol{\mu}^\star, \Sigma^\star),$$

where $\boldsymbol{z} = (z_1, \cdots, z_n)^T$ is a vector of future data generated independently of the observed $\boldsymbol{y}$, and the mean $\boldsymbol{\mu}^\star$ and covariance matrix $\Sigma^\star$ of the Bayesian predictive distribution are respectively given by

$$\boldsymbol{\mu}^{\star} = \hat{\beta}\Phi\hat{\Sigma}\Phi^{T}\boldsymbol{y}, \quad \Sigma^{\star} = \hat{\beta}^{-1}I_n + \Phi\hat{\Sigma}\Phi^{T},$$

where $\hat{\Sigma} = (\hat{A} + \hat{\beta}\Phi^{T}\Phi)^{-1}$ and $\hat{A} = \mathrm{diag}(\hat{\alpha}_1, \cdots, \hat{\alpha}_n)$. Hereafter, for simplicity, we denote $h(\boldsymbol{z}|\boldsymbol{y}, \hat{\boldsymbol{\alpha}}, \hat{\beta})$ as $h(\boldsymbol{z}|\boldsymbol{y})$.

3.2. Proposed criterion

Kitagawa (1997) proposed an information criterion to evaluate the goodness of the Bayesian predictive distribution based on the Kullback-Leibler divergence, called predictive information criterion (PIC). In this paper, according to Kitagawa (1997), we derive the PIC for the RVR models. The PIC for Bayesian models is, in general, given by

$$\mathrm{PIC} = -2\log h(\boldsymbol{y}|\boldsymbol{y}) + 2\mathrm{Bias},$$

where Bias is a bias term between the log-likelihood and the expected log-likelihood as follows:

$$\mathrm{Bias} = E_{q(\boldsymbol{y})}\left[\log h(\boldsymbol{y}|\boldsymbol{y}) - E_{q(\boldsymbol{z})}\left[\log h(\boldsymbol{z}|\boldsymbol{y})\right]\right], \tag{6}$$

where $q(\cdot)$ is an unknown true distribution and we assume that

$$q(\boldsymbol{z}) = p(\boldsymbol{z}|\tilde{\boldsymbol{w}}, \tilde{\beta}) = N_n(\boldsymbol{z}; \Phi\tilde{\boldsymbol{w}}, \tilde{\beta}^{-1}I_n) = N_n(\boldsymbol{z}; \tilde{\boldsymbol{\mu}}, \tilde{\Sigma}),$$

where the true mean $\tilde{\boldsymbol{\mu}}$ and covariance matrix $\tilde{\Sigma}$ are given by $\tilde{\boldsymbol{\mu}} = \Phi\tilde{\boldsymbol{w}}$ and $\tilde{\Sigma} = \tilde{\beta}^{-1}I_n$, respectively. Under this assumption, the bias term (6) can be written as

$$\begin{aligned}
\mathrm{Bias} &= E_{p(\boldsymbol{y}|\tilde{\boldsymbol{\mu}},\tilde{\beta})}\left[\log h(\boldsymbol{y}|\boldsymbol{y}) - E_{p(\boldsymbol{z}|\tilde{\boldsymbol{\mu}},\tilde{\beta})}\left[\log h(\boldsymbol{z}|\boldsymbol{y})\right]\right] \\
&= -\frac{1}{2}E_{p(\boldsymbol{y}|\tilde{\boldsymbol{\mu}},\tilde{\beta})}\left[(\boldsymbol{y}-\boldsymbol{\mu}^{\star})^{T}\Sigma^{\star-1}(\boldsymbol{y}-\boldsymbol{\mu}^{\star}) - E_{p(\boldsymbol{z}|\tilde{\boldsymbol{\mu}},\tilde{\beta})}\left[(\boldsymbol{z}-\boldsymbol{\mu}^{\star})^{T}\Sigma^{\star-1}(\boldsymbol{z}-\boldsymbol{\mu}^{\star})\right]\right] \\
&= -\frac{1}{2}\mathrm{tr}\left\{\Sigma^{\star-1}E_{p(\boldsymbol{y}|\tilde{\boldsymbol{\mu}},\tilde{\beta})}\left[(\boldsymbol{y}-\boldsymbol{\mu}^{\star})(\boldsymbol{y}-\boldsymbol{\mu}^{\star})^{T} - E_{p(\boldsymbol{z}|\tilde{\boldsymbol{\mu}},\tilde{\beta})}\left[(\boldsymbol{z}-\boldsymbol{\mu}^{\star})(\boldsymbol{z}-\boldsymbol{\mu}^{\star})^{T}\right]\right]\right\} \\
&= -\frac{1}{2}\mathrm{tr}\left\{\Sigma^{\star-1}E_{p(\boldsymbol{y}|\tilde{\boldsymbol{\mu}},\tilde{\beta})}\left[(\boldsymbol{y}-\tilde{\boldsymbol{\mu}})(\tilde{\boldsymbol{\mu}}-\boldsymbol{\mu}^{\star})^{T} + (\tilde{\boldsymbol{\mu}}-\boldsymbol{\mu}^{\star})(\boldsymbol{y}-\tilde{\boldsymbol{\mu}})^{T}\right]\right\}.
\end{aligned} \tag{7}$$

Here, the first term of (7) is

$$\begin{aligned}
&E_{p(\boldsymbol{y}|\tilde{\boldsymbol{\mu}},\tilde{\beta})}\left[(\boldsymbol{y}-\tilde{\boldsymbol{\mu}})(\tilde{\boldsymbol{\mu}}-\boldsymbol{\mu}^{\star})^{T}\right] \\
&= E_{p(\boldsymbol{y}|\tilde{\boldsymbol{\mu}},\tilde{\beta})}\left[(\boldsymbol{y}-\tilde{\boldsymbol{\mu}})(\tilde{\boldsymbol{\mu}}-\boldsymbol{y})^{T} + (\boldsymbol{y}-\tilde{\boldsymbol{\mu}})(\boldsymbol{y}-H\boldsymbol{y})^{T}\right]
\end{aligned}$$

$$
\begin{aligned}
&= -\tilde{\beta}^{-1} I_n + \tilde{\beta}^{-1} I_n (I_n - H)^T \\
&= -\tilde{\beta}^{-1} H,
\end{aligned}
$$

where $H = \hat{\beta}\Phi\hat{\Sigma}\Phi^T$ is the hat matrix. Similarly, the second term is given by

$$
E_{p(\boldsymbol{y}|\tilde{\boldsymbol{\mu}},\tilde{\beta})}\left[(\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}^\star)(\boldsymbol{y} - \tilde{\boldsymbol{\mu}})^T\right] = -\tilde{\beta}^{-1} H.
$$

Therefore, the bias term (6) is evaluated as

$$
\begin{aligned}
\text{Bias} &= -\frac{1}{2}\text{tr}\left\{\Sigma^{\star -1} E_{p(\boldsymbol{y}|\tilde{\boldsymbol{\mu}},\tilde{\beta})}\left[(\boldsymbol{y} - \tilde{\boldsymbol{\mu}})(\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}^\star)^T + (\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}^\star)(\boldsymbol{y} - \tilde{\boldsymbol{\mu}})^T\right]\right\} \\
&= -\frac{1}{2}\text{tr}\left\{\Sigma^{\star -1}(-\tilde{\beta}^{-1}H - \tilde{\beta}^{-1}H)\right\} \\
&= \tilde{\beta}^{-1}\text{tr}\{\Sigma^{\star -1}H\}. \qquad (8)
\end{aligned}
$$

Consequently, by replacing the bias term in (8), we obtain

$$
\text{PIC} = n\log(2\pi) + \log|\Sigma^\star| + (\boldsymbol{y} - \boldsymbol{\mu}^\star)^T\Sigma^{\star -1}(\boldsymbol{y} - \boldsymbol{\mu}^\star) + 2\tilde{\beta}^{-1}\text{tr}\{\Sigma^{\star -1}H\}.
$$

After estimating the unknown true variance $\tilde{\beta}^{-1}$ by the maximum marginal likelihood estimator $\hat{\beta}^{-1}_{\text{MMLE}} = \hat{\beta}^{-1}$, we have the PIC in the framework of the RVR models as follows:

$$
\text{PIC} = n\log(2\pi) + \log|\Sigma^\star| + (\boldsymbol{y} - \boldsymbol{\mu}^\star)^T\Sigma^{\star -1}(\boldsymbol{y} - \boldsymbol{\mu}^\star) + 2\hat{\beta}^{-1}_{\text{MMLE}}\text{tr}\{\Sigma^{\star -1}H\}.
$$

We select the optimal values of the adjusted parameters of kernel functions that minimize the PIC.

3.3. Other selection methods

3.3.1 Maximization of marginal likelihood

For determining the optimal values for the kernel parameters, Tipping (2001, p.235) introduced the method by maximizing the marginal likelihood

$$
p(\boldsymbol{y}|\boldsymbol{\alpha},\beta) = \int p(\boldsymbol{y}|\boldsymbol{w},\beta)p(\boldsymbol{w}|\boldsymbol{\alpha})\text{d}\boldsymbol{w} = N_n(\boldsymbol{y};\boldsymbol{0},C),
$$

with respect to kernel parameter. It is, however, pointed out that this technique has the difficulty of the implementation, so the cross-validation (CV) is suggested as a better method for this selection (Tipping, 2001; Quinonero-Candela and Hansen, 2002). Instead, we consider that the minimization of the negative value of log marginal likelihood as follows:

$$
\text{ML} = n\log(2\pi) + \log|\hat{\beta}^{-1}I_n + \Phi\hat{A}^{-1}\Phi^T| + \boldsymbol{y}^T(\hat{\beta}^{-1}I_n + \Phi\hat{A}^{-1}\Phi^T)^{-1}\boldsymbol{y}.
$$

3.3.2 Generalized cross-validation

From a predictive viewpoint, the goodness of a fitted regression model is ordinary measured by the predictive squared error (PSE) $= \sum_{i=1}^{n}\{z_i - \hat{u}(\boldsymbol{x}_i)\}^2/n$, where z_i are future observations and $\hat{u}(\boldsymbol{x})$ is the estimated regression function. However, it is difficult to consider situations in which future observations z_i are observed.

The cross-validation (Stone, 1974) is a method to evaluate a statistical model from a predictive point of view. The leave-one-out CV or n-fold CV $(= \sum_{i=1}^{n}\{y_i - \hat{u}^{(-i)}(\boldsymbol{x}_i)\}^2/n)$ was introduced as the estimator of the PSE by separating the data used for model estimation and model evaluation, where $\hat{u}^{(-i)}(\boldsymbol{x})$ is the regression function constructed by the $n-1$ observations removed the i-th observation $(y_i, \boldsymbol{x}_i)$. However, this method is known as the evaluation method with computationally expensive because of the data separation. In order to overcome this difficulty, Craven and Wahba (1979) introduced the more generalized cross-validation (GCV) given by

$$\mathrm{GCV} = \frac{1}{n}\frac{\sum_{i=1}^{n}\{y_i - \hat{u}(\boldsymbol{x}_i)\}^2}{\{1 - \mathrm{tr}(H)/n\}^2},$$

where H is the hat matrix. The GCV can alleviate the high computational cost of the CVs. For the detail of the cross-validation methods, we refer to Stone (1974), Geisser (1975), Efron (1982) and Konishi and Kitagawa (2008).

3.3.3 Information criteria

Akaike information criterion (AIC; Akaike, 1973; Akaike, 1974) was proposed for evaluating the goodness of statistical models from a predictive point of view through the Kullback-Leibler information (Kullback and Leibler, 1951) between the statistical models and a true model.

Hastie and Tibshirani (1990) introduced the modified Akaike information criterion (mAIC) by replacing the number of efficient parameters in AIC with the trace of the hat matrix H. The mAIC for the RVR model is given by

$$\begin{aligned}\mathrm{mAIC} = {} & n\log(2\pi) - n\log\hat{\beta} + \hat{\beta}(\boldsymbol{y} - \Phi\hat{\boldsymbol{w}})^T(\boldsymbol{y} - \Phi\hat{\boldsymbol{w}}) \\ & +2\mathrm{tr}\{\hat{\beta}\Phi(\hat{A} + \hat{\beta}\Phi^T\Phi)^{-1}\Phi^T\}.\end{aligned}$$

The Bayesian information criterion (BIC) proposed by Schwarz (1978) is a model evaluation criterion of statistical models obtained from a Bayesian viewpoint. In a similar way to obtain the mAIC, we consider the BIC given by

$$\begin{aligned}\mathrm{BIC} = {} & n\log(2\pi) - n\log\hat{\beta} + \hat{\beta}(\boldsymbol{y} - \Phi\hat{\boldsymbol{w}})^T(\boldsymbol{y} - \Phi\hat{\boldsymbol{w}}) \\ & + \log(n)\mathrm{tr}\{\hat{\beta}\Phi(\hat{A} + \hat{\beta}\Phi^T\Phi)^{-1}\Phi^T\}.\end{aligned}$$

We select the optimal values for kernel parameters that minimize these model selection and evaluation criteria.

4. Numerical examples

4.1. Monte Carlo simulations

For the simulated study, the repeated random samples $\{(y_i, x_i); i = 1, \cdots, n\}$ with $n = 25, 50$ or 100 were generated from a true regression model $y_i = u(x_i) + \varepsilon_i$. Here, we considered the true regression models

$$u_1(x) = \sin(2\pi x^3), \tag{9}$$
$$u_2(x) = 3\exp(-3x)\sin(3\pi x), \tag{10}$$

where the design points x_i are randomly distributed in $[0, 1]$ and the error ε_i are independently, normally distributed with mean zero and variance $\sigma^2 = 0.1^2, 0.15^2$ or 0.2^2. Figure 1 shows the true regression curves.

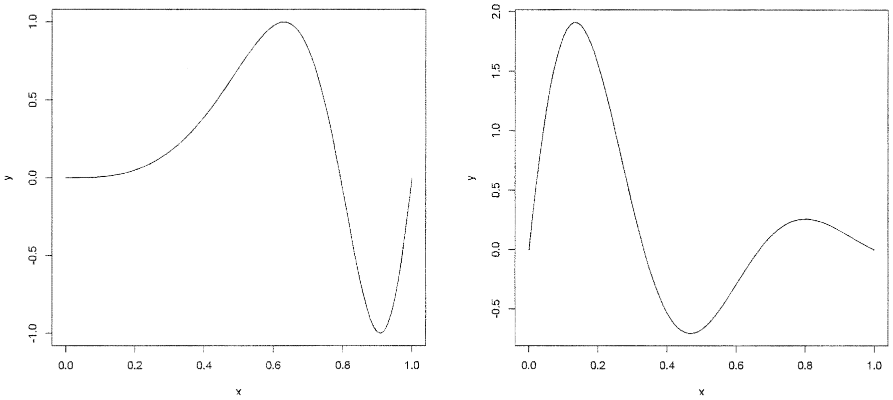

Figure 1: The true regression curves for Monte Carlo simulations (left, $u_1(x)$; right, $u_2(x)$).

For the analyses of the simulated data, we considered the RVR models with the Gaussian kernel functions

$$k(\boldsymbol{x}, \boldsymbol{x}_j) = \exp\left\{-\frac{||\boldsymbol{x} - \boldsymbol{x}_j||^2}{2h^2}\right\}, \quad j = 1, \cdots, n,$$

where h^2 is the adjusted parameter of dispersion. We set the candidate values of the width parameter h^2 to $\{0.10^2, 0.11^2, \cdots, 0.30^2\}$ and chose the optimal value of this parameter that minimizes the evaluation criteria; i.e., ML, GCV, mAIC, BIC and our proposed PIC.

We repeated 1000 times in each situation and then computed the number of relevance vectors (RVs) as the measure of the sparsity, the mean squared error (MSE) defined by MSE $= \sum_{i=1}^{n}\{y_i - \hat{u}(x_i)\}^2/n$ as the goodness of fitting for the observed data and the PSE $= \sum_{i=1}^{n}\{z_i - \hat{u}(x_i)\}^2/n$ as the goodness of predictive capability, where z_i are the future observations generated from true regression models (9) and (10). Tables 1 to 6 show the simulation results with the mean values (MEAN) and standard deviations (SD) for RVs, MSE and PSE, respectively.

From simulation results, our proposed PIC criterion has the best fitting and prediction accuracies (low MSE and PSE) especially in the cases such as small n, and tends to choose the model with the low sparsity (i.e., large value of the number of RVs) than other criteria. The GCV and mAIC are the second best criteria on the MSE and PSE. However, the ML and BIC criteria often fail to select a good model with unstable SD and tend to obtain models with high sparsity.

Table 1: Comparison of simulation results for the true function $u_1(x)$ with $\sigma = 0.1$.

	RVs	**MSE**	**PSE**
Criterion	**MEAN(SD)**	**MEAN(SD)**$\times 10^{-3}$	**MEAN(SD)**$\times 10^{-2}$
($n = 25$)			
ML	4.043 (1.193)	8.149 (3.392)	1.444 (0.463)
GCV	4.364 (1.060)	7.543 (2.945)	1.398 (0.442)
mAIC	4.382 (1.058)	7.534 (2.938)	1.397 (0.442)
BIC	4.209 (1.089)	7.655 (2.998)	1.408 (0.448)
PIC	4.412 (1.052)	**7.511** (2.929)	**1.395** (0.442)
($n = 50$)			
ML	4.972 (1.132)	8.786 (2.059)	1.256 (0.260)
GCV	5.092 (1.092)	8.551 (1.964)	1.226 (0.254)
mAIC	5.108 (1.091)	8.547 (1.964)	1.226 (0.254)
BIC	4.821 (1.076)	8.666 (1.998)	1.233 (0.258)
PIC	5.151 (1.112)	**8.536** (1.962)	**1.225** (0.254)
($n = 100$)			
ML	5.567 (1.185)	9.307 (1.375)	1.134 (0.173)
GCV	6.356 (3.017)	9.188 (1.345)	**1.118** (0.166)
mAIC	6.388 (3.067)	9.185 (1.343)	**1.118** (0.166)
BIC	5.441 (1.286)	9.302 (1.367)	1.121 (0.169)
PIC	6.621 (3.652)	**9.174** (1.344)	**1.118** (0.166)

4.2. Real data analyses

We analyzed the fossil data (Bralower *et al.*, 1997), using nonlinear modeling procedures. The fossil data from **SemiPar** package (Ruppert *et al.*, 2003) in R has 106 observations on fossil shells. The data consist of the age in millions of years as x and the ratios of strontium isotopes as y. At first, we demonstrated the performances of our proposed PIC and other model selection criteria by using a subset of the data for testing and the remainders for training. We fitted our

Table 2: Comparison of simulation results for the true function $u_1(x)$ with $\sigma = 0.15$.

Criterion	**RVs** **MEAN(SD)**	**MSE** **MEAN(SD)**$\times 10^{-2}$	**PSE** **MEAN(SD)**$\times 10^{-2}$
$(n = 25)$			
ML	3.552 (1.144)	1.824 (0.699)	3.191 (0.964)
GCV	3.903 (1.053)	1.703 (0.638)	3.109 (0.923)
mAIC	3.950 (1.042)	1.697 (0.633)	3.107 (0.921)
BIC	3.736 (1.075)	1.727 (0.648)	3.121 (0.927)
PIC	3.975 (1.029)	**1.694** (0.633)	**3.104** (0.920)
$(n = 50)$			
ML	4.426 (1.149)	1.967 (0.446)	2.802 (0.601)
GCV	4.607 (1.053)	1.923 (0.426)	2.746 (0.587)
mAIC	4.622 (1.050)	1.922 (0.425)	2.746 (0.586)
BIC	4.353 (1.078)	1.944 (0.431)	2.765 (0.591)
PIC	4.631 (1.054)	**1.920** (0.425)	**2.744** (0.584)
$(n = 100)$			
ML	5.219 (1.116)	2.084 (0.329)	2.547 (0.374)
GCV	5.419 (1.624)	2.062 (0.324)	2.510 (0.369)
mAIC	5.422 (1.622)	2.062 (0.324)	2.510 (0.369)
BIC	5.074 (1.212)	2.079 (0.326)	2.518 (0.371)
PIC	5.507 (1.713)	**2.060** (0.324)	**2.509** (0.369)

Table 3: Comparison of simulation results for the true function $u_1(x)$ with $\sigma = 0.2$.

Criterion	**RVs** **MEAN(SD)**	**MSE** **MEAN(SD)**$\times 10^{-2}$	**PSE** **MEAN(SD)**$\times 10^{-2}$
$(n = 25)$			
ML	3.206 (1.056)	3.238 (1.281)	5.432 (1.685)
GCV	3.574 (1.010)	3.039 (1.188)	5.344 (1.647)
mAIC	3.599 (1.010)	3.034 (1.186)	5.339 (1.645)
BIC	3.379 (1.021)	3.087 (1.209)	5.364 (1.663)
PIC	3.635 (1.005)	**3.026** (1.182)	**5.334** (1.642)
$(n = 50)$			
ML	3.993 (1.135)	3.575 (0.849)	4.960 (1.035)
GCV	4.251 (1.026)	3.478 (0.809)	4.874 (1.007)
mAIC	4.261 (1.027)	3.477 (0.808)	4.873 (1.007)
BIC	3.976 (1.030)	3.522 (0.827)	4.912 (1.023)
PIC	4.275 (1.018)	**3.474** (0.806)	**4.870** (1.008)
$(n = 100)$			
ML	4.843 (1.258)	3.757 (0.576)	4.486 (0.645)
GCV	5.029 (1.312)	3.716 (0.562)	4.432 (0.630)
mAIC	5.029 (1.312)	3.716 (0.562)	4.432 (0.630)
BIC	4.660 (1.203)	3.746 (0.569)	4.458 (0.639)
PIC	5.068 (1.339)	**3.714** (0.561)	**4.430** (0.628)

Table 4: Comparison of simulation results for the true function $u_2(x)$ with $\sigma = 0.1$.

Criterion	**RVs** **MEAN(SD)**	**MSE** **MEAN(SD)**$\times 10^{-3}$	**PSE** **MEAN(SD)**$\times 10^{-2}$
$(n = 25)$			
ML	4.211 (0.938)	7.940 (3.035)	1.445 (0.453)
GCV	4.478 (1.049)	7.162 (2.646)	1.397 (0.422)
mAIC	4.517 (1.071)	7.147 (2.642)	1.396 (0.421)
BIC	4.326 (1.019)	7.264 (2.674)	1.404 (0.424)
PIC	4.572 (1.084)	**7.116** (2.622)	**1.394** (0.421)
$(n = 50)$			
ML	4.708 (0.962)	9.010 (2.143)	1.274 (0.269)
GCV	5.766 (1.949)	8.409 (1.944)	1.228 (0.258)
mAIC	5.802 (1.950)	8.402 (1.942)	1.228 (0.258)
BIC	5.252 (1.691)	8.540 (1.984)	1.235 (0.260)
PIC	5.898 (2.062)	**8.387** (1.940)	**1.226** (0.257)
$(n = 100)$			
ML	5.157 (1.143)	9.545 (1.460)	1.142 (0.166)
GCV	7.490 (3.967)	9.169 (1.369)	1.111 (0.161)
mAIC	7.504 (3.962)	9.166 (1.368)	1.111 (0.161)
BIC	6.338 (2.862)	9.302 (1.384)	1.115 (0.164)
PIC	7.750 (4.155)	**9.150** (1.366)	**1.110** (0.160)

Table 5: Comparison of simulation results for the true function $u_2(x)$ with $\sigma = 0.15$.

Criterion	**RVs** **MEAN(SD)**	**MSE** **MEAN(SD)**$\times 10^{-2}$	**PSE** **MEAN(SD)**$\times 10^{-2}$
$(n = 25)$			
ML	3.799 (0.955)	1.732 (0.660)	3.126 (0.922)
GCV	4.008 (0.993)	1.610 (0.595)	3.064 (0.895)
mAIC	4.032 (0.993)	1.606 (0.593)	3.063 (0.893)
BIC	3.843 (0.986)	1.638 (0.610)	3.069 (0.889)
PIC	4.062 (0.987)	**1.601** (0.590)	**3.057** (0.893)
$(n = 50)$			
ML	4.289 (0.908)	2.025 (0.454)	2.754 (0.568)
GCV	4.844 (1.584)	1.932 (0.424)	2.686 (0.540)
mAIC	4.853 (1.582)	1.931 (0.424)	2.685 (0.540)
BIC	4.396 (1.194)	1.960 (0.434)	2.704 (0.548)
PIC	5.017 (1.718)	**1.927** (0.423)	**2.684** (0.540)
$(n = 100)$			
ML	4.738 (1.098)	2.125 (0.315)	2.582 (0.374)
GCV	6.228 (2.902)	2.057 (0.299)	2.522 (0.359)
mAIC	6.253 (2.908)	2.057 (0.299)	2.521 (0.359)
BIC	5.426 (2.098)	2.078 (0.304)	2.530 (0.361)
PIC	6.438 (3.010)	**2.054** (0.298)	**2.519** (0.359)

Table 6: Comparison of simulation results for the true function $u_2(x)$ with $\sigma = 0.2$.

Criterion	RVs MEAN(SD)	MSE MEAN(SD)$\times 10^{-2}$	PSE MEAN(SD)$\times 10^{-2}$
($n = 25$)			
ML	3.542 (0.975)	3.044 (1.102)	5.413 (1.613)
GCV	3.694 (0.954)	2.884 (1.055)	5.363 (1.591)
mAIC	3.713 (0.965)	2.880 (1.054)	5.365 (1.591)
BIC	3.565 (0.964)	2.919 (1.070)	5.385 (1.600)
PIC	3.748 (0.968)	**2.870** (1.047)	**5.356** (1.594)
($n = 50$)			
ML	4.029 (0.950)	3.544 (0.785)	4.850 (1.027)
GCV	4.307 (1.275)	3.431 (0.759)	4.772 (1.010)
mAIC	4.348 (1.318)	3.429 (0.758)	4.771 (1.010)
BIC	3.937 (1.008)	3.472 (0.767)	4.801 (1.014)
PIC	4.430 (1.433)	**3.423** (0.756)	**4.768** (1.007)
($n = 100$)			
ML	4.434 (1.060)	3.788 (0.564)	4.540 (0.657)
GCV	5.356 (2.215)	3.690 (0.544)	4.462 (0.643)
mAIC	5.363 (2.218)	3.690 (0.544)	4.462 (0.643)
BIC	4.643 (1.594)	3.729 (0.550)	4.487 (0.647)
PIC	5.624 (2.482)	**3.685** (0.544)	**4.456** (0.643)

modeling procedure for the complete data at the second.

We selected a random subset of size m ($= 10$ or 20) from the fossil data for testing, and fitted our modeling procedure based on the RVR with Gaussian kernel functions to the remainders. Here, we set the candidate values of the kernel parameter h^2 to $\{4.0^2, 4.1^2, \cdots, 8.0^2\}$ and chose the optimal value of h^2 that minimizes the model selection criteria. We repeated this procedure 1000 times in each situation and then computed the mean predictive error for the test data. Table 7 shows the result with the mean values (MEAN) and standard deviations (SD) for the mean predictive error. From this results, we pointed out that the PIC based modeling gives a stable model estimate as compared with other techniques.

Secondly, we fitted our modeling procedure to the complete fossil data using the same candidate values for the width parameter h^2 as described above. Then, PIC, GCV and mAIC selected $h^2 = 4.9^2$, ML selected $h^2 = 5.3^2$, and BIC selected $h^2 = 6.0^2$, respectively. Figure 2 shows the fitted curves for the fossil data.

5. Concluding remarks

This paper considered the model selection and evaluation problem of the relevance vector regression (RVR) models, which is the determination of the optimal values for the kernel parameters. In order to evaluate the RVR models more effectively, we derived the model selection criterion for evaluating a Bayesian predictive

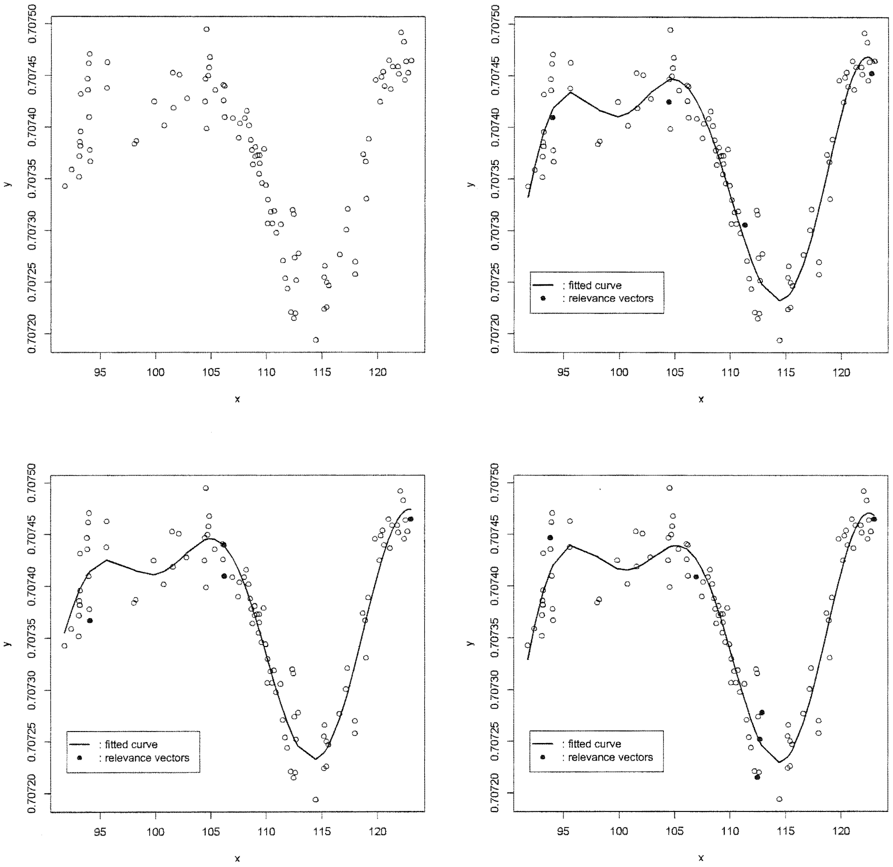

Figure 2: The fossil data (upper left) and fitted curves (upper right, PIC, GCV and mAIC; lower left, ML; lower right, BIC).

Table 7: Comparison results for the fossil data with data separation.

Criterion	Mean predictive error MEAN(SD)×10^{-10}
$(m = 10)$	
ML	7.761 (3.444)
GCV	7.331 (3.215)
mAIC	7.400 (3.223)
BIC	7.439 (3.320)
PIC	**7.302 (3.203)**
$(m = 20)$	
ML	7.939 (2.306)
GCV	7.363 (2.102)
mAIC	7.402 (2.114)
BIC	7.465 (2.185)
PIC	**7.355 (2.098)**

distribution in the framework of the RVR. Monte Carlo experiments and real data analysis showed that our proposed modeling procedure performs well in various situations. The simulation results suggested that our PIC has better fitting and predictive accuracies compared with other model selection criteria.

Acknowledgements The author would like to thank Professor Sadanori Konishi of Chuo University for his valuable advice and encouragement. The author is also grateful to the Editor and two anonymous reviewers for their helpful suggestions.

References

[1] Akaike, H. (1973). Maximum likelihood identification of gaussian autoregressive moving average models. *Biometrika*, **60**, 255–265.

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

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

[4] Bai, Y., Wang, P., Li, C., Xie, J. and Wang, Y. (2014). A multi-scale relevance vector regression approach for daily urban water demand forecasting. *Journal of Hydrology*, **517**, 236-245.

[5] Bishop, C.M. (1995). *Neural Networks for Pattern Recognition.* Oxford University Press.

[6] Bishop, C.M. (2006). *Pattern Recognition and Machine Learning.* Springer.

[7] de Boor, C. (2001). *A Practical Guide to Splines.* Springer.

[8] Bralower, T.J., Fullagar, P.D., Paull, C.K., Dwyer, G.S. and Leckie, R.M. (1997). Mid-cretaceous strontium-isotope stratigraphy of deep-sea sections. *Geological Society of America Bulletin*, **109**, 1421–1442.

[9] Cheng, B., Zhang, D., Chen, S., Kaufer, D.I. and Shen, D. (2013). Semi-Supervised Multimodal Relevance Vector Regression Improves Cognitive Performance Estimation from

Imaging and Biological Biomarkers. *Neuroinformatics*, **11**(3), 339–353.

[10] Craven, P. and Wahba, G. (1979). Optimal smoothing of noisy data with spline functions. *Numerische Mathematik*, **31**, 377–403.

[11] Denison, D.G.T., Holmes, C.C., Mallick, B.K. and Smith, A.F. (2002). *Bayesian Method for Nonlinear Classification and Regression.* Wiley.

[12] Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans.* Society for Industrial and Applied Mathematics.

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

[14] Figueiredo, M.A.T. (2003). Adaptive sparseness for supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **25**, 1150–1159.

[15] Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association*, **70**, 320–328.

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

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

[18] Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models.* Chapman and Hall.

[19] Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning (2nd ed.).* Springer.

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

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

[22] Kawano, S., Hoshina, I., Matsuda, K. and Konishi, S. (2014). Predictive model selection criteria for Bayesian lasso. *MI Preprint Series*, Kyushu University.

[23] Kim, D., Kawano, S. and Konishi, S. (2012). Predictive information criteria for Bayesian nonlinear regression models. *Bulletin of Informatics and Cybernetics*, **44**, 17–28.

[24] Kitagawa, G. (1997). Information criteria for the predictive evaluation of Bayesian models. *Communications in Statistics-Theory and Methods*, **26**, 2223–2246.

[25] Konishi, S. and Kitagawa, G. (2008). *Information Criteria and Statistical Modeling.* Springer.

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

[27] Liu, F., Zhou, J.Z., Qiu, F.P., Yang, J.J. and Liu, L. (2006). Nonlinear hydrological time series forecasting based on the relevance vector regression. *Lecture Notes in Computer Science*, **4233**, 880–889.

[28] Neal, R.M. (1996). *Bayesian Learning for Neural Networks.* Springer.

[29] Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, **103**, 681–686.

[30] Quinonero-Candela, J. and Hansen, L.K. (2002). Time series prediction based on the relevance vector machine with adaptive kernels. *IEEE international conference on acoustics, speech and signal processing*, 985–988.

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

[32] Ruppert, D., Wand, M.P. and Carroll, R.J. (2003). *Semiparametric Regression.* Cambridge University Press.

[33] Scholkopf, B. and Smola, A.J. (2001). *Learning with kernels: support vector machines, regularization, optimization, and beyond.* MIT press.

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

[35] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society*, **B–36**, 111–147.

[36] Tipping, M.E. (2000). The relevance vector machine. *Advances in Neural Information*

Processing Systems, **12**, 652–658.

[37] Tipping, M.E. (2001). Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, **1**, 211–244.

[38] Tipping, M.E. and Faul, A. (2003). Fast marginal likelihood maximization for sparse Bayesian models. In C. M. Bishop and B. Frey (Eds.), *Proceedings Ninth International Workshop on Artificial Intelligence and Statistics, Key West, Florida.*

[39] Tripathi, S. and Govindaraju, R.S. (2007). On selection of kernel parameters in relevance vector machines for hydrologic applications. *Stochastic Environmental Research and Risk Assessment*, **21**(6), 747–764.

[40] Vapnik, V.N. (1995). *The nature of statistical learning theory.* Springer.

[41] Vapnik, V.N. (1998). *Statistical learning theory.* Wiley.

Kazuki MATSUDA

Department of Mathematics, Graduate School of Science and Engineering, Chuo University,
1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan
kazuki@gug.math.chuo-u.ac.jp