PREDICTIVE INFORMATION CRITERIA FOR ROBUST RELEVANCE VECTOR REGRESSION MODELS

Matsuda, Kazuki
Department of Mathematics, Graduate School of Science and Engineering, Chuo University

Kawano, Shuichi
Department of Computer and Network Engineering, Graduate School of Informatics and Engineering, The University of Electro-Communications

Konishi, Sadanori
Department of Mathematics, Faculty of Science and Engineering, Chuo University

https://doi.org/10.5109/2233862

出版情報：Bulletin of informatics and cybernetics. 50, pp.65-80, 2018-12. 統計科学研究会
バージョン：
権利関係：

PREDICTIVE INFORMATION CRITERIA FOR ROBUST RELEVANCE VECTOR REGRESSION MODELS

by

Kazuki Matsuda, **Shuichi Kawano** and **Sadanori Konishi**

Reprinted from the Bulletin of Informatics and Cybernetics
Research Association of Statistical Sciences, Vol.50

FUKUOKA, JAPAN
2018

Bulletin of Informatics and Cybernetics, Vol. 50, 2018

PREDICTIVE INFORMATION CRITERIA FOR ROBUST RELEVANCE VECTOR REGRESSION MODELS

By

Kazuki Matsuda[*], **Shuichi Kawano**[†] and **Sadanori Konishi**[‡]

Abstract

The relevance vector regression (RVR) is a Bayesian nonlinear regression procedure whose model is expressed in terms of kernel functions, like the support vector regression (SVR). In order to overcome the sensitivity to outliers of the RVR, the robust relevance vector regression (RRVR) procedures have been proposed. A crucial issue in the model building process of the RRVR is the choice of kernel parameters. The selection of these parameters can be viewed as a model selection and evaluation problem. In this paper, we derive a model selection criterion for the Bayesian predictive distribution of the RRVR models. Monte Carlo experiments and real data analysis show that our model selection criterion performs well in various situations.

Key Words and Phrases: Basis expansions, Bayesian predictive distribution, Model selection, Robustness.

1. Introduction

Nonlinear regression modeling procedures are useful for analyzing data with a complex structure and these procedures have been widely developed in various fields of statistical science (see, e.g., Bishop, 2006; Hastie *et al.*, 2009). Basis expansion methods are known to be efficient for constructing nonlinear regression models. The essential idea of basis expansions is to express a regression function as a linear combination of specified nonlinear functions, called basis functions (Konishi and Kitagawa, 2008; Hastie *et al.*, 2009). In order to capture the nonlinear structure of the data, various types of 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; Ando *et al.*, 2008; Hastie *et al.*, 2009), and thin plate splines (Girosi *et al.*, 1995). These regression models are characterized by parameters that need to be estimated. However, maximum likelihood and least squares

[*] Department of Mathematics, Graduate School of Science and Engineering, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan. tel +81–3–3817–1740 kazuki@gug.math.chuo-u.ac.jp

[†] Department of Computer and Network Engineering, Graduate School of Informatics and Engineering, The University of Electro-Communications, 1-5-1 Chofugaoka, Chofu-shi, Tokyo 182-8585, Japan. tel +81–42–443–5620 skawano@ai.lab.uec.ac.jp

[‡] Department of Mathematics, Faculty of Science and Engineering, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan. tel +81–3–3817–1715 konishi@math.chuo-u.ac.jp

methods often yield unstable estimated models. To overcome this difficulty, regularization methods and the Bayesian approach have been widely used in model estimation (see, e.g., Denison *et al.*, 2002; Figueiredo, 2003; Bishop, 2006).

The relevance vector regression (RVR; Tipping, 2000; Tipping, 2001) is a nonlinear regression procedure which expresses a model as a linear combination of kernel functions with forms similar to those in the support vector regression (SVR; Vapnik, 1995; Vapnik, 1998). The coefficient parameters of the RVR model are automatically estimated, and most of these are determined to be zero by a Bayesian methodology using automatic relevance determination (ARD; Neal, 1996). It is known that the RVR overcomes some weak points of the SVR, e.g., a selection problem of a trade-off parameter and a lack of probability structure, and many successful examples by the RVR in various fields have been reported (see, e.g., Liu *et al.*, 2006; Cheng *et al.*, 2013; Bai *et al.*, 2014).

However, the RVR is sensitive to outliers because of the assumption that the errors are Gaussian distributed with same variance parameter. This is important since it is known that the goodness of fit for the data and predictive accuracy decrease when the data include outliers. To overcome this difficulty, Han and Zhao (2010) proposed the robust relevance vector regression (RRVR) procedure, in which the errors are assumed to be Gaussian distributed with different variance parameters. Mitra *et al.* (2010) also introduced the RRVR procedure using the sparse outlier modeling approach.

A crucial issue for the RRVR model building process is the selection of the optimal values of the parameters of the kernel basis functions. This selection can be viewed as a model selection and evaluation problem. In order to solve the model selection problem, Stone (1974) presented the cross-validation (CV) method, which is one of the popular model evaluation criteria. The CV evaluates the goodness of statistical models from a predictive point of view by separating the data into training data and test data. Craven and Wahba (1979) subsequently proposed the generalized cross-validation (GCV) which approximates the leave-one-out CV criterion without data separation. The Akaike information criterion (AIC; Akaike, 1973; Akaike, 1974) is another well-known model selection criterion; the AIC evaluates the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951) between a statistical model and an unknown true model. In the modified Akaike information criterion (mAIC) method proposed by Hastie and Tibshirani (1990), an offset term in the AIC is replaced by the trace value of the prediction matrix, called the hat matrix. Schwarz (1978) presented the Bayesian information criterion (BIC) for evaluating the goodness of a statistical model estimated by the maximum likelihood method from a Bayesian point of view. However, the RRVR is a Bayesian regression procedure, and these information criteria were not derived to evaluate Bayesian statistical models. We consider here how to solve this model selection problem from a theoretical point of view. In Bayesian regression analysis, it is important to obtain a Bayesian predictive distribution from a predictive viewpoint. However, there has been little research on choosing the optimal values of kernel parameters in the RRVR framework by evaluating the Bayesian predictive distribution.

In this paper, we derive a model selection and evaluation criterion for the Bayesian predictive distribution in the RRVR framework. In deriving our proposed procedure, we referred the predictive information criterion (PIC) proposed by Kitagawa (1997). In Kitagawa (1997), the PIC was derived as an estimator of the KL divergence between the Bayesian predictive distribution and the true model in the framework of the Bayesian regression model based on Gaussian noise with known variance. The PIC has been used

to evaluate various statistical models, e.g., Bayesian regression models with unknown variance (Kim *et al.*, 2012), the Bayesian lasso by Park and Casella (2008) (Kawano *et al.*, 2014), and the RVR models (Matsuda, 2015). In the model selection process of the RRVR, the values of the kernel parameters are selected to minimize a proposed criterion. The performance of this proposed model selection criterion is investigated in various situations by using Monte Carlo simulations and real data examples.

The remainder of the paper is organized as follows. Section 2 describes the nonlinear regression model building process of the RVR. In Section 3, we describe the framework of the RRVR models. We derive a model selection criterion for the Bayesian predictive distribution of the RRVRs in Section 4. Monte Carlo simulations and a real data analysis are presented in Section 5 to examine the performance of our model selection strategy. Concluding remarks are given in Section 6. Detailed derivations of the Bayesian predictive distribution and the PIC are relegated to the Appendix.

2. Relevance vector regression

Suppose that n observations $\{(y_i, x_i); i = 1, \cdots, n\}$ are obtained in terms of the response variable y and explanatory variable x. In order to analyze data with a nonlinear structure, we consider the regression model

$$y_i = u(x_i) + \varepsilon_i, \quad i = 1, \cdots, n,$$

where $u(x)$ is an unknown regression function and errors ε_i are independently distributed according to $N(\varepsilon_i; 0, \beta^{-1})$, where the notation indicates a Gaussian distribution over ε_i with mean 0 and variance β^{-1} $(\beta > 0)$.

Tipping (2000, 2001) presented a sparse Bayesian procedure for constructing regression models called the relevance vector regression (RVR). In model building processes based on the RVR, kernel functions are used as basis functions as follows:

$$u(x_i; \mathbf{w}) = w_0 + \sum_{j=1}^{n} w_j k_j(x_i, x_j) = \mathbf{w}^T \phi(x_i),$$

where $k_j(x, x')$ are the kernel functions and $\phi(x) = (1, k_1(x, x_1), \cdots, k_n(x, x_n))^T$ is a vector of basis functions. For the kernel function, Gaussian, polynomial, and sigmoid types are widely used. For more details about kernel functions, we refer the reader to Bishop (2006). If the error terms ε_i are independently distributed according to $N(\varepsilon_i; 0, \beta^{-1})$, then the model has the likelihood function

$$f(\mathbf{y}|\mathbf{w}, \beta) = (2\pi\beta^{-1})^{-\frac{n}{2}} \exp\left\{-\frac{\beta}{2}(\mathbf{y} - \Phi\mathbf{w})^T(\mathbf{y} - \Phi\mathbf{w})\right\},$$

where $\mathbf{y} = (y_1, \cdots, y_n)^T$ and $\Phi = (\phi(x_1), \cdots, \phi(x_n))^T$.

Next, we assume that the coefficient vector $\mathbf{w}$ has an automatic relevance determination (ARD; Neal, 1996) prior

$$p(\mathbf{w}|\alpha) = \prod_{j=0}^{n} N(w_j; 0, \alpha_j^{-1}) = (2\pi)^{-\frac{n}{2}} |A|^{\frac{1}{2}} \exp\left\{-\frac{1}{2}\mathbf{w}^T A \mathbf{w}\right\}, \tag{1}$$

where $\alpha = (\alpha_0, \alpha_1, \cdots, \alpha_n)^T$ is a vector of hyperparameters and $A = \mathrm{diag}(\alpha_0, \alpha_1, \cdots, \alpha_n)$. Then, from Bayes' theorem, the posterior distribution of $\mathbf{w}$ is

$$p(\mathbf{w}|\mathbf{y}, \alpha, \beta) = (2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{w} - \mu)^T \Sigma^{-1} (\mathbf{w} - \mu)\right\},$$

where the posterior mean μ and covariance matrix Σ are, respectively,

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

The estimated values of hyperparameters α and variance parameter β are obtained by the type-Ⅱ maximum likelihood method (Berger, 1985), which maximizes the marginal likelihood function

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

where $C = \beta^{-1} I_n + \Phi A^{-1} \Phi^T$ and I_n is an $n \times n$ identity matrix. Setting the derivatives of the marginal likelihood to zero, we obtain the estimators of α and β as

$$\hat{\alpha}_j = \frac{\gamma_j}{\mu_j^2} \quad (j = 0, 1, \cdots, n),$$
$$\hat{\beta}^{-1} = \frac{||\mathbf{y} - \Phi\mu||^2}{n - \sum_{j=0}^{n} \gamma_j},$$

where μ_j is the j-th element of the posterior mean μ, $\gamma_j = 1 - \alpha_j \Sigma_{jj}$, Σ_{jj} is the j-th diagonal element of the posterior covariance matrix Σ, and $||\cdot||$ is the Euclidean norm.

3. Robust relevance vector regression

3.1. RRVR$_1$: An approach using different variance parameters

The RVR is sensitive to outliers because of the assumption that the Gaussian errors are distributed with same variance parameter. To overcome this difficulty, Han and Zhao (2010) proposed the robust relevance vector regression (RRVR) procedure, which has the errors distributed with different variance parameters. Henceforth, we call this method RRVR$_1$.

For n independent observations $\{(y_i, x_i); i = 1, \cdots, n\}$, the RRVR$_1$ model is given by

$$y_i = \mathbf{w}^T \phi(x_i) + \varepsilon_i, \quad i = 1, \cdots, n,$$

where ε_i are independently distributed as $N(\varepsilon_i; 0, \sigma^2/\beta_i)$, in which σ^2 is the average variance parameter and β_i are the individual variance parameters. Here, the prior distribution of β_i is assumed to be the gamma distribution $Gamma(a_i, b_i)$, where $a_i > 0$ and $b_i > 0$. Then, the RRVR$_1$ model has the following likelihood function:

$$f(\mathbf{y}|\mathbf{w}, \beta, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} |B|^{\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \Phi\mathbf{w})^T B (\mathbf{y} - \Phi\mathbf{w})\right\},$$

where $B = \mathrm{diag}(\beta_1, \cdots, \beta_n)$. Assuming ARD prior distribution (1) for coefficients $\mathbf{w}$ the posterior distribution of $\mathbf{w}$ is given by

$$p(\mathbf{w}|\mathbf{y}, \alpha, \beta, \sigma^2) = (2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(\mathbf{w} - \mu)^T \Sigma^{-1} (\mathbf{w} - \mu) \right\},$$

where the posterior mean μ and covariance matrix Σ are, respectively,

$$\mu = \sigma^{-2} \Sigma \Phi^T B \mathbf{y}, \quad \Sigma = (A + \sigma^{-2} \Phi^T B \Phi)^{-1}.$$

Then, the marginal likelihood function is

$$p(\mathbf{y}|\alpha, \beta, \sigma^2) = \int p(\mathbf{y}|\mathbf{w}, \beta, \sigma^2) p(\mathbf{w}|\alpha) \mathrm{d}\mathbf{w} = N_n(\mathbf{y}; \mathbf{0}, C),$$

where $C = \sigma^2 B^{-1} + \Phi A^{-1} \Phi^T$. The estimators of α, β, and σ^2 are obtained by maximizing the logarithm of the product of $p(\mathbf{y}|\alpha, \beta, \sigma^2)$ and $p(\beta)$ as follows:

$$\begin{aligned}
\hat{\alpha}_j &= \frac{\gamma_j}{\mu_j^2} \quad (j = 0, 1, \cdots, n), \\
\hat{\beta}_i &= \frac{2a_i + 1}{2b_i + \sigma^{-2}[(y_i - \phi(x_i)^T \mu)^2 + \phi(x_i)^T \Sigma \phi(x_i)]} \quad (i = 1, \cdots, n), \\
\hat{\sigma^2} &= \frac{(\mathbf{y} - \Phi\mu)^T B (\mathbf{y} - \Phi\mu)}{n - \sum_{j=0}^{n} \gamma_j},
\end{aligned}$$

where $\gamma_j = 1 - \alpha_j \Sigma_{jj}$, and Σ_{jj} is the j-th diagonal element of the posterior covariance matrix Σ.

Since these formulas depend on each other, the calculation must be repeated until a convergence condition is satisfied. In this optimization, most coefficient parameters are estimated to be zero and the corresponding kernel functions are removed from the model. The estimated model is then constructed using some design points corresponding to non-zero coefficients, where these design points are referred to as relevance vectors (RV; Tipping, 2001).

3.2. RRVR$_2$: An approach using sparse outlier model

Mitra *et al.* (2010) proposed the RRVR procedure using the sparse outlier modeling approach. Henceforth, we call this method RRVR$_2$. In the framework of the RRVR$_2$, the model is given by

$$\mathbf{y} = \Phi\mathbf{w} + \boldsymbol{\varepsilon} + \mathbf{s},$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_n)^T$ is a Gaussian component of the noise, ε_i is distributed according to $N(\varepsilon_i; 0, \beta^{-1})$, and $\mathbf{s} = (s_1, \cdots, s_n)^T$ is a component due to outliers. The above equation can be written as

$$\mathbf{y} = \Psi\mathbf{w}_s + \boldsymbol{\varepsilon},$$

where $\Psi = [\Phi; I_n]$ is an $n \times (2n+1)$ matrix and $\mathbf{w}_s = [\mathbf{w}^T, \mathbf{s}^T]^T$ is the $(2n+1)$-dimensional coefficient vector.

Here, we consider an ARD prior over the coefficients $\mathbf{w}_s$, and then the posterior distribution can be obtained by a Gaussian distribution $N(\mathbf{w}_s; \mu, \Sigma)$, where the posterior mean and covariance matrix are given by, respectively,

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

where $A = \mathrm{diag}(\alpha_0, \cdots, \alpha_{2n})$. The updates of the parameters are similar with those of RVR in Section 2.

4. Model selection criterion

4.1. Predictive information criterion for the RRVR$_1$

The Bayesian predictive distribution in the framework of the RRVR$_1$ model is given by

$$h(\mathbf{z}|\mathbf{y}, \hat{\alpha}, \hat{\beta}, \hat{\sigma}^2) = N(\mathbf{z}; \mu_p, \Sigma_p),$$

where $\mathbf{z} = (z_1, \cdots, z_n)^T$ is a vector of future data generated independently of the observed $\mathbf{y}$, $\mu_p = \Phi \hat{\mu}$, $\Sigma_p = \hat{\sigma}^2 \hat{B}^{-1} + \Phi \hat{\Sigma} \Phi^T$, $\hat{\mu} = \hat{\sigma}^{-2} \hat{\Sigma} \Phi^T \hat{B} \mathbf{y}$, $\hat{\Sigma} = (\hat{A} + \hat{\sigma}^{-2} \Phi^T \hat{B} \Phi)^{-1}$, $\hat{A} = \mathrm{diag}(\hat{\alpha}_0, \hat{\alpha}_1, \cdots, \hat{\alpha}_n)$ and $\hat{B} = \mathrm{diag}(\hat{\beta}_1, \cdots, \hat{\beta}_n)$. The derivation of the Bayesian predictive distribution is given in Appendix A. Here, we denote $h(\mathbf{z}|\mathbf{y}, \hat{\alpha}, \hat{\beta}, \hat{\sigma}^2)$ as $h(\mathbf{z}|\mathbf{y})$.

The predictive information criterion (PIC) was proposed by Kitagawa (1997) for evaluating the goodness of the Bayesian predictive distribution. The PIC of the RRVR$_1$ model is

$$\mathrm{PIC} = -2 \log h(\mathbf{y}|\mathbf{y}) + 2 E_{q(\mathbf{y})} \left[\log h(\mathbf{y}|\mathbf{y}) - E_{q(\mathbf{z})} \left[\log h(\mathbf{z}|\mathbf{y}) \right] \right],$$

where $q(\cdot)$ is the unknown true distribution. Then, we assume that

$$q(\mathbf{z}) = p(\mathbf{z}|\theta_q) = p(\mathbf{z}|\mathbf{w}_q, \beta_q, \sigma_q^2) = N(\mathbf{z}|\mu_q, \Sigma_q),$$

where the true mean μ_q and the covariance matrix Σ_q are given by, respectively,

$$\mu_q = \Phi \mathbf{w}_q, \quad \Sigma_q = \sigma_q^2 B_q^{-1},$$

where $B_q = \mathrm{diag}(\beta_{q1}, \cdots, \beta_{qn})$.

Then, we obtain

$$\begin{aligned}
&E_{q(\mathbf{y})} \left[\log h(\mathbf{y}|\mathbf{y}) - E_{q(\mathbf{z})} \left[\log h(\mathbf{z}|\mathbf{y}) \right] \right] \\
&= -\frac{1}{2} E_{p(\mathbf{y}|\theta_q)} \left[(\mathbf{y} - \mu_p)^T \Sigma_p^{-1} (\mathbf{y} - \mu_p) - E_{p(\mathbf{z}|\theta_q)} \left[(\mathbf{z} - \mu_p)^T \Sigma_p^{-1} (\mathbf{z} - \mu_p) \right] \right] \\
&= -\frac{1}{2} \mathrm{tr} \left\{ \Sigma_p^{-1} E_{p(\mathbf{y}|\theta_q)} \left[(\mathbf{y} - \mu_q)(\mu_q - \mu_p)^T + (\mu_q - \mu_p)(\mathbf{y} - \mu_q)^T \right] \right\}.
\end{aligned}$$

If we define a matrix H_p such that $\mu_p = \hat{\sigma}^{-2} \Phi \hat{\Sigma} \Phi^T \hat{B} \mathbf{y} = H_p \mathbf{y}$, then

$$E_{p(\mathbf{y}|\theta_q)} \left[(\mathbf{y} - \mu_q)(\mu_q - \mu_p)^T \right] = -\sigma_q^2 B_q^{-1} H_p^T. \tag{2}$$

The derivation of (2) is given in Appendix B. Similarly, we can obtain

$$E_{p(\mathbf{y}|\theta_q)}\left[(\mu_q-\mu_p)(\mathbf{y}-\mu_q)^T\right]=E_{p(\mathbf{y}|\theta_q)}\left[\left\{(\mathbf{y}-\mu_q)(\mu_q-\mu_p)^T\right\}^T\right]=-\sigma_q^2H_pB_q^{-1}.$$

Consequently, the PIC in the framework of the RRVR_1 model is obtained as follows:

$$\begin{aligned}\mathrm{PIC}&=n\log(2\pi)+\log|\Sigma_p|+(\mathbf{y}-\mu_p)^T\Sigma_p^{-1}(\mathbf{y}-\mu_p)+\sigma_q^2\mathrm{tr}\{\Sigma_p^{-1}(B_q^{-1}H_p^T+H_pB_q^{-1})\}\\&=n\log(2\pi)+\log|\Sigma_p|+(\mathbf{y}-\mu_p)^T\Sigma_p^{-1}(\mathbf{y}-\mu_p)+2\sigma_q^2\mathrm{tr}(\Sigma_p^{-1}H_pB_q^{-1}).\end{aligned}$$

However, this criterion depends on the unknown true values β_q and σ_q^2. We replace these values by their estimators as follows:

$$\mathrm{PIC}=n\log(2\pi)+\log|\Sigma_p|+(\mathbf{y}-\mu_p)^T\Sigma_p^{-1}(\mathbf{y}-\mu_p)+2\hat{\sigma}^2\mathrm{tr}(\Sigma_p^{-1}H_p\hat{B}^{-1}).$$

We select the optimal kernel parameters so that they minimize the PIC.

4.2. Predictive information criterion for the RRVR_2

In the framework of the RRVR_2 model, the Bayesian predictive distribution is the Gaussian distribution with mean μ_p and Σ_p given by

$$\mu_p=\hat{\beta}\Phi\hat{\Sigma}_w\Phi^T\mathbf{y},\quad \Sigma_p=\hat{\beta}^{-1}I_n+\Phi\hat{\Sigma}_w\Phi^T,$$

where $\hat{\Sigma}_w$ is the $(n+1)\times(n+1)$ matrix whose (i,j)-th element is $(\hat{\Sigma})_{ij}$. Here $(A)_{ij}$ means the (i,j)-th element of the matrix A and $\hat{\Sigma}=(\hat{A}+\hat{\beta}\Psi^T\Psi)^{-1}$.

According to Matsuda (2015), the PIC in the framework of the RRVR_2 model is obtained by

$$\mathrm{PIC}=n\log(2\pi)+\log|\Sigma_p|+(\mathbf{y}-\mu_p)^T\Sigma_p^{-1}(\mathbf{y}-\mu_p)+2\hat{\beta}^{-1}\mathrm{tr}\Sigma_p^{-1}H,$$

where $H=\hat{\beta}\Phi\hat{\Sigma}\Phi^T$ is the hat matrix.

4.3. Other model selection criteria

4.3.1. Cross-validation

The cross-validation (CV; Stone, 1974) method evaluates a statistical model from a predictive point of view. The leave-one-out CV ($=\sum_{i=1}^n\{y_i-\hat{u}^{(-i)}(x_i)\}^2/n$) was introduced as an estimator of the predictive squared error by separating the data used for model estimation and evaluation, where $\hat{u}^{(-i)}(x)$ is the regression function constructed from the $(n-1)$ observations with the i-th observation (y_i,x_i) removed. The leave-one-out CV is equivalent to the n-fold CV, in which the data is partitioned into n dataset for training and testing. K-fold CV method is defined by dividing the data into K parts $\{S_1,\cdots,S_K\}$ as follows

$$\frac{1}{K}\sum_{k=1}^K\frac{1}{n_k}\sum_{i\in S_k}\left\{y_i-\hat{u}^{(-S_k)}(x_i)\right\}^2.$$

4.3.2. Generalized cross-validation

The CV method is known to be computationally expensive because of the data separation. In order to overcome this difficulty, Craven and Wahba (1979) introduced the generalized cross-validation (GCV) given by

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

where $H = \sigma^{-2}\Sigma\Phi^T B$ is the hat matrix. The GCV can avoid the high computational cost of the conventional CV. For the details of the cross-validation methods, we refer the reader to Stone (1974), Geisser (1975), Efron (1982), and Konishi and Kitagawa (2008).

4.3.3. Akaike information criterion

The Akaike information criterion (AIC; Akaike, 1973; Akaike, 1974) was proposed for evaluating the goodness of statistical models from a predictive point of view based on the KL divergence (Kullback and Leibler, 1951) between the statistical models estimated by the maximum likelihood method and the true model. Replacing the number of free parameters in AIC with the trace of the hat matrix H, Hastie and Tibshirani (1990) introduced the modified Akaike information criterion (mAIC). The mAIC for the RRVR_1 model is given by

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

4.3.4. Bayesian information criterion

Schwarz (1978) presented the Bayesian information criterion (BIC), which is a model evaluation criterion for statistical models obtained from a Bayesian point of view. In a similar way to obtaining the mAIC, the mBIC of the RRVR_1 model is can be obtained as

$$\begin{aligned}\mathrm{mBIC} = {} & n\log(2\pi\hat{\sigma}^2) - \log|\hat{B}| + \hat{\sigma}^{-2}(\mathbf{y} - \Phi\hat{\mathbf{w}})^T\hat{B}(\mathbf{y} - \Phi\hat{\mathbf{w}}) \\ & + \log(n)\mathrm{tr}\{\hat{\sigma}^{-2}(\hat{A} + \hat{\sigma}^{-2}\Phi^T\hat{B}\Phi)^{-1}\Phi^T\hat{B}\}.\end{aligned}$$

The optimal values for the kernel parameters are then those that minimize these model selection and evaluation criteria.

5. Numerical examples

We demonstrated our model selection and evaluation criterion for the RRVR_1.

5.1. Monte Carlo simulations

For the numerical simulations, we assumed that repeated random samples $\{(x_i, y_i); i = 1, \cdots, n\}$ with $n = 50$ or 100 were generated from a true regression model $y_i = u(x_i) + \varepsilon_i$, where we considered the following true regression models:

$$u_1(x) = \frac{\sin(x)}{x},$$

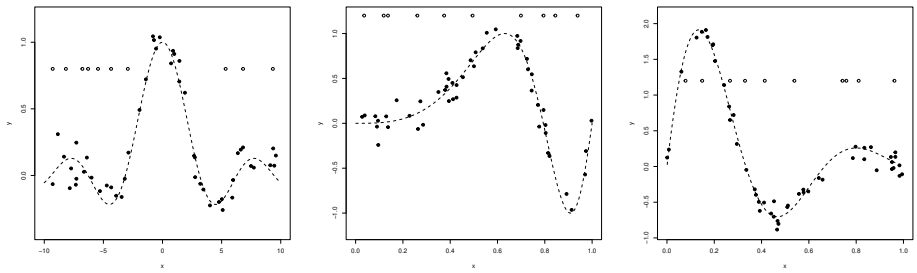

Figure 1: True curves (dotted lines; left, u_1; middle, u_2; right, u_3) and generated data (black dots) with outliers (white circles).

$$u_2(x) = \sin(2\pi x^3),$$
$$u_3(x) = 3\exp(-3x)\sin(3\pi x).$$

Here, the design points x_i were randomly distributed over the interval $[-10, 10]$ in the case of $u_1(x)$ and $[0, 1]$ in the other cases, and the errors ε_i were independently and normally distributed with mean zero and variance $\sigma^2 = 0.1^2, 0.2^2$, or 0.3^2. In order to consider the situation that the data include outliers, we add random samples $\{(x_k, y_k = 0.8); k = 1, \cdots, l\}$ in the case of $u_1(x)$ and $\{(x_k, y_k = 1.2); k = 1, \cdots, l\}$ in the other cases with $l = 5$ or 10, where x_k were also randomly distributed in the same interval as each of the design points. Figure 1 shows the true curves along with the generated data with outliers.

In constructing a regression model, we consider the Gaussian kernel function

$$k(x, x') = \exp\left\{-\frac{(x - x')^2}{2h^2}\right\}, \tag{3}$$

where h^2 is the width parameter and we set the candidate values $\{1.5^2, 1.6^2, \cdots, 3.5^2\}$ in the case of $u_1(x)$ and $\{0.30^2, 0.31^2, \cdots, 0.50^2\}$ in the other cases, and set the hyperparameters a and b to 1. The optimal value of the parameter of the kernel function is then the one that minimizes the model evaluation criterion.

We performed 200 repetitions for each model situation and then calculated the mean squared error between the estimated value and the value of the true model, namely, $\sum_{i=1}^{n}\{u(x_i) - \hat{u}(x_i)\}^2/n$, as the goodness of the predictive accuracy. Tables 1 to 3 show the simulation results for the mean value (MEAN) and standard deviation (SD) of the mean squared error in each situation.

The simulation results are summarized as follows. First, in many cases, the PIC performs well for the model selection problem from a predictive point of view (i.e., it provides a small mean squared error). The mAIC and mBIC have inferior results in all cases. In several situations, the 10-fold CV (i.e., we use $K = 10$) and the GCV perform better than the PIC, but there are cases when the CV methods perform significantly worse than the PIC. Therefore, the PIC is considered to be better and more stable model selection and evaluation criterion than other methods.

Table 1: Simulation results for the true function $u_1(x)$. The bold values correspond to the smallest means.

			Mean squared error (MEAN(SD)$\times 10^{-2}$)				
n	σ	l	10-fold CV	GCV	mAIC	mBIC	PIC
50	0.1	5	0.370(0.976)	0.462(0.875)	0.752(0.530)	0.645(0.161)	**0.298**(0.338)
		10	0.853(1.885)	1.065(1.882)	1.925(1.370)	1.797(1.075)	**0.569**(1.295)
	0.2	5	1.256(1.787)	1.242(1.858)	2.086(0.698)	2.108(0.711)	**1.177**(0.777)
		10	1.885(2.542)	1.872(2.594)	3.722(1.205)	3.586(1.156)	**1.599**(1.284)
	0.3	5	2.146(1.952)	**2.085**(1.977)	3.140(1.131)	3.154(1.258)	2.187(1.178)
		10	2.981(2.467)	**2.861**(2.600)	4.384(1.573)	4.494(1.910)	3.006(1.686)
100	0.1	5	**0.154**(0.069)	0.170(0.066)	0.189(0.074)	0.182(0.066)	0.155(0.070)
		10	0.171(0.098)	0.193(0.100)	0.211(0.090)	0.204(0.066)	**0.145**(0.082)
	0.2	5	**0.550**(0.309)	0.576(0.310)	0.618(0.260)	0.623(0.269)	0.582(0.259)
		10	**0.581**(0.35)	0.640(0.337)	0.683(0.318)	0.662(0.259)	0.582(0.284)
	0.3	5	**1.092**(0.548)	1.109(0.569)	1.204(0.447)	1.220(0.485)	1.154(0.441)
		10	1.363(1.127)	**1.358**(1.131)	1.591(0.601)	1.599(0.653)	1.368(0.661)

Table 2: Simulation results for the true function $u_2(x)$. The bold values correspond to the smallest means.

			Mean squared error (MEAN(SD)$\times 10^{-2}$)				
n	σ	l	10-fold CV	GCV	mAIC	mBIC	PIC
50	0.1	5	1.140(2.019)	1.151(2.054)	2.437(1.685)	2.546(1.673)	**0.806**(1.803)
		10	1.805(4.075)	2.170(3.681)	4.25(3.176)	3.782(3.265)	**1.173**(2.674)
	0.2	5	1.532(1.760)	1.743(1.781)	2.873(1.327)	2.901(0.687)	**1.235**(1.037)
		10	2.856(3.292)	3.097(3.314)	5.330(2.793)	5.282(3.400)	**1.956**(3.372)
	0.3	5	2.361(1.813)	2.469(1.880)	3.483(1.284)	3.594(1.431)	**2.269**(1.360)
		10	3.884(3.685)	4.073(3.678)	5.591(2.781)	5.727(3.479)	**3.191**(3.006)
100	0.1	5	0.306(1.075)	0.276(1.433)	0.905(0.319)	1.488(0.116)	**0.192**(0.817)
		10	0.318(1.521)	0.367(1.534)	1.582(0.405)	2.017(0.141)	**0.220**(0.383)
	0.2	5	0.665(1.000)	0.689(1.029)	1.857(0.324)	2.148(0.249)	**0.628**(0.301)
		10	0.823(1.147)	0.926(1.130)	2.263(0.691)	2.411(0.313)	**0.660**(0.581)
	0.3	5	1.209(0.951)	1.223(1.015)	2.286(0.575)	2.481(0.519)	**1.192**(0.589)
		10	1.520(1.294)	1.557(1.315)	2.782(0.897)	2.843(0.594)	**1.302**(0.973)

Table 3: Simulation results for the true function $u_3(x)$. The bold values correspond to the smallest means.

			Mean squared error (MEAN(SD)$\times 10^{-2}$)				
n	σ	l	10-fold CV	GCV	mAIC	mBIC	PIC
50	0.1	5	0.750(3.347)	0.650(3.434)	4.352(0.595)	3.914(0.679)	**0.552**(1.166)
		10	1.884(4.740)	2.012(4.952)	6.821(2.224)	6.380(0.735)	**0.757**(2.312)
	0.2	5	1.532(3.783)	1.544(4.407)	3.760(2.121)	3.797(1.049)	**1.368**(1.068)
		10	2.686(5.453)	2.396(5.688)	6.228(2.203)	5.845(0.982)	**1.617**(2.619)
	0.3	5	2.381(2.027)	**2.306**(3.943)	3.105(1.202)	3.758(1.198)	2.350(1.283)
		10	3.903(4.442)	3.590(6.371)	4.730(2.724)	5.368(2.197)	**3.362**(2.762)
100	0.1	5	0.204(2.408)	0.214(2.432)	1.666(0.163)	1.717(0.128)	**0.170**(0.142)
		10	0.274(2.804)	0.297(2.871)	4.779(0.248)	4.677(0.265)	**0.203**(0.208)
	0.2	5	0.558(1.647)	**0.556**(1.438)	1.301(0.291)	1.144(0.334)	0.582(0.285)
		10	0.668(2.413)	0.682(2.327)	3.213(0.441)	2.624(0.396)	**0.606**(0.437)
	0.3	5	1.187(0.977)	**1.167**(1.058)	1.487(0.565)	1.542(0.638)	1.243(0.57)
		10	**1.196**(1.354)	1.241(1.186)	1.797(0.573)	1.702(0.649)	1.283(0.567)

5.2. Real data analysis

We investigated the performance of our model procedure for choosing the kernel parameters of the RRVR_1 model through an analysis of fossil data (Bralower *et al.*, 1997). The fossil dataset, which is available in the **SemiPar** R package (Ruppert *et al.*, 2003), has 106 observations on fossil shells. Each observation consists of the age in millions of years as x and a ratio of strontium isotopes as y. Because the range of values of variable y is very small, for the analysis, we normalized the values of y.

We compared our proposed procedure to other model selection criteria by using a subset of the data for testing and the remainder for training. More specifically, we selected a random subset of size m ($= 10$ or 20) from the fossil data for testing and added l ($= 5$ or 10) random outliers to the remaining data. The x values of the outliers were randomly distributed over the interval $[90, 125]$ and the y values were all set to -2. Figure 2 shows an example of the fossil data and outliers. We fitted the RRVR_1 model with Gaussian kernel (3) and selected the optimal value of kernel parameter h^2, i.e., the value that minimizes the model selection criterion, for each criterion. We set the candidates of the kernel parameter as $\{2.5^2, 3.0^2, \cdots, 7.5^2\}$.

For each situation, we performed 200 repetitions and calculated the mean predictive error for the test data, namely, $\sum_{i=1}^{m}\{u(x_i) - \hat{u}(x_i)\}^2/m$. Table 4 shows the results in terms of the mean value (MEAN) and standard deviation (SD) of the mean predictive error. The result presented to demonstrate that the PIC performs well in all situations.

6. Conclusion

We derived a model selection criterion for evaluating the Bayesian predictive distribution of RRVR models by Han and Zhao (2010) and Mitra *et al.* (2010). This criterion enabled us to objectively select an optimal value of the kernel parameter in RRVR models. Monte Carlo experiments and real data analysis showed that our proposed criterion performs well in various situations.

It is interesting to extend our methodology for other Bayesian procedures and

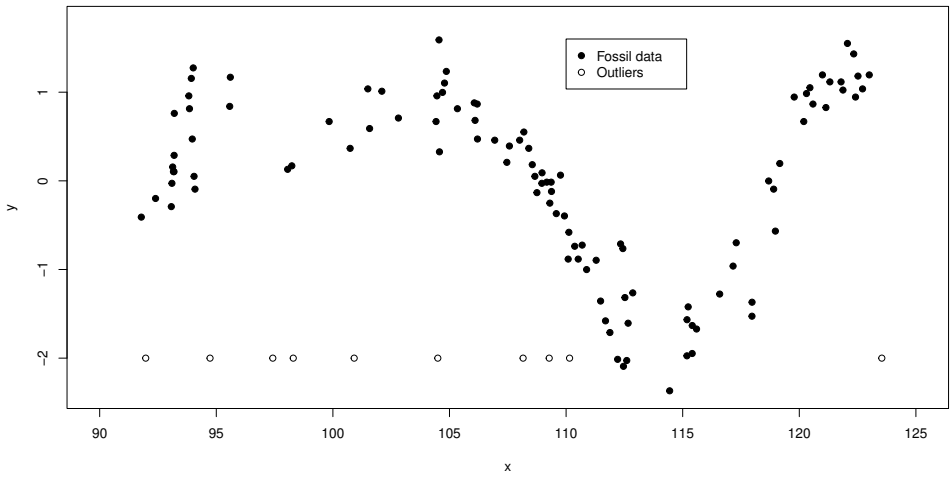

Figure 2: Fossil data with outliers.

Table 4: Results for the fossil data. The bold values correspond to the smallest means.

		Mean predictive error (MEAN(SD))			
m	l	GCV	mAIC	mBIC	PIC
10	5	0.150(0.075)	0.174(0.066)	0.176(0.067)	**0.148**(0.074)
	10	0.166(0.137)	0.168(0.090)	0.167(0.074)	**0.152**(0.093)
20	5	0.156(0.057)	0.180(0.055)	0.180(0.055)	**0.155**(0.054)
	10	0.164(0.084)	0.172(0.065)	0.170(0.044)	**0.159**(0.083)

the regularization methods: e.g., the nonconvex penalization methods (She and Owen, 2011). We consider these topics as future research.

Appendix

A. Derivation of Bayesian predictive distribution

The Bayesian predictive distribution of the RRVR_1 model is given by

$$
\begin{aligned}
&h(\mathbf{z}|\mathbf{y},\hat{\alpha},\hat{\beta},\hat{\sigma}^2)\\
&=\int p(\mathbf{z}|\hat{\mathbf{w}},\hat{\beta},\hat{\sigma}^2)p(\mathbf{w}|\mathbf{y},\hat{\alpha},\hat{\beta},\hat{\sigma}^2)\mathrm{d}\mathbf{w}\\
&=\int N(\mathbf{z};\Phi\mathbf{w},\hat{\sigma}^2\hat{B}^{-1})N(\mathbf{w};\hat{\mu},\hat{\Sigma})\mathrm{d}\mathbf{w}\\
&=c_1\int\exp\left\{-\frac{1}{2}\left(\hat{\sigma}^{-2}(\mathbf{z}-\Phi\mathbf{w})^T\hat{B}(\mathbf{z}-\Phi\mathbf{w})+(\mathbf{w}-\hat{\mu})^T\hat{\Sigma}^{-1}(\mathbf{w}-\hat{\mu})\right)\right\}\mathrm{d}\mathbf{w}\\
&=c_2\exp\left\{-\frac{1}{2}\hat{\sigma}^{-2}\mathbf{z}^T\hat{B}\mathbf{z}\right\}\int\exp\Bigg\{\\
&\quad-\frac{1}{2}\left(\mathbf{w}-(\hat{\Sigma}^{-1}+\hat{\sigma}^2\Phi^T\hat{B}\Phi)^{-1}(\hat{\Sigma}^{-1}\mu+\hat{\sigma}^{-2}\Phi^T\hat{B}\mathbf{z})\right)^T(\hat{\Sigma}^{-1}+\hat{\sigma}^2\Phi^TB\Phi)\\
&\quad\left(\mathbf{w}-(\hat{\Sigma}^{-1}+\hat{\sigma}^2\Phi^T\hat{B}\Phi)^{-1}(\hat{\Sigma}^{-1}\mu\hat{\sigma}^{-2}\Phi^T\hat{B}\mathbf{z})\right)\Bigg\}\\
&\quad+\frac{1}{2}(\hat{\Sigma}^{-1}\mu+\hat{\sigma}^{-2}\Phi^T\hat{B}\mathbf{z})^T(\hat{\Sigma}^{-1}+\hat{\sigma}^2\Phi^T\hat{B}\Phi)^{-1}(\hat{\Sigma}^{-1}\mu\hat{\sigma}^{-2}\Phi^T\hat{B}\mathbf{z})\mathrm{d}\mathbf{w}\\
&=c_3\exp\left\{-\frac{1}{2}\left(\mathbf{z}^T(\hat{\sigma}^{-2}\hat{B}-\hat{\sigma}^{-4}\hat{B}\Phi(\hat{\Sigma}^{-1}+\hat{\sigma}^{-2}\Phi^T\hat{B}\Phi)^{-1}\Phi^T\hat{B})\mathbf{z}\right.\right.\\
&\quad\left.\left.-2\hat{\sigma}^{-2}\mathbf{z}^T\hat{B}\Phi(\hat{\Sigma}^{-1}+\hat{\sigma}^{-2}\Phi^T\hat{B}\Phi)^{-1}\hat{\Sigma}^{-1}\hat{\mu}\right)\right\}\\
&=c_3\exp\left\{-\frac{1}{2}\left(\mathbf{z}^T(\hat{\sigma}^2\hat{B}^{-1}+\Phi\hat{\Sigma}\Phi^T)^{-1}\mathbf{z}\right.\right.\\
&\quad\left.\left.-2\hat{\sigma}^{-2}\mathbf{z}^T\hat{B}\Phi(\hat{\Sigma}^{-1}+\hat{\sigma}^{-2}\Phi^T\hat{B}\Phi)^{-1}\hat{\Sigma}^{-1}\hat{\mu}\right)\right\}\\
&=c_3\exp\Bigg\{-\frac{1}{2}\left(\mathbf{z}-\hat{\sigma}^{-2}(\hat{\sigma}^2\hat{B}^{-1}+\Phi\hat{\Sigma}\Phi^T)\hat{B}\Phi(\hat{\Sigma}^{-1}+\hat{\sigma}^{-2}\Phi^T\hat{B}\Phi)^{-1}\hat{\Sigma}^{-1}\hat{\mu}\right)^T(\hat{\sigma}^2\hat{B}^{-1}\\
&\quad+\Phi\hat{\Sigma}\Phi^T)^{-1}\left(\mathbf{z}-\hat{\sigma}^{-2}(\hat{\sigma}^2\hat{B}^{-1}+\Phi\hat{\Sigma}\Phi^T)\hat{B}\Phi(\hat{\Sigma}^{-1}+\hat{\sigma}^{-2}\Phi^T\hat{B}\Phi)^{-1}\hat{\Sigma}^{-1}\hat{\mu}\right)\Bigg\},
\end{aligned}
$$

where c_1, c_2, and c_3 are constant values. Here, the equation

$$(\hat{\Sigma}^{-1}+\hat{\sigma}^{-2}\Phi^T\hat{B}\Phi)^{-1}=\hat{\Sigma}-\hat{\Sigma}\Phi(\hat{\sigma}^2\hat{B}^{-1}+\Phi\hat{\Sigma}\Phi^T)^{-1}\Phi^T\hat{\Sigma}$$

is used. Also, we have

$$
\begin{aligned}
&\hat{\sigma}^{-2}(\hat{\sigma}^2\hat{B}^{-1}+\Phi\hat{\Sigma}\Phi^T)\hat{B}\Phi(\hat{\Sigma}^{-1}+\hat{\sigma}^{-2}\Phi^T\hat{B}\Phi)^{-1}\hat{\Sigma}^{-1}\hat{\mu}\\
&=\hat{\sigma}^{-2}(\hat{\sigma}^2\hat{B}^{-1}+\Phi\hat{\Sigma}\Phi^T)\hat{B}\Phi(\hat{\Sigma}-\hat{\Sigma}\Phi(\hat{\sigma}^2\hat{B}^{-1}+\Phi\hat{\Sigma}\Phi^T)^{-1}\Phi^T\hat{\Sigma})\hat{\Sigma}^{-1}\hat{\mu}
\end{aligned}
$$

$$
\begin{aligned}
&= \Phi\hat{\mu} - \Phi\hat{\Sigma}\Phi^T(\hat{\sigma}^2\hat{B}^{-1} + \Phi\hat{\Sigma}\Phi^T)^{-1}\Phi\hat{\mu} + \hat{\sigma}^{-2}\Phi\hat{\Sigma}\Phi^T\hat{B}\Phi\hat{\mu} \\
&\quad -\hat{\sigma}^{-2}\Phi\hat{\Sigma}\Phi^T\hat{B}\Phi\hat{\Sigma}\Phi^T(\hat{\sigma}^2\hat{B}^{-1} + \Phi\hat{\Sigma}\Phi^T)^{-1}\Phi\hat{\mu} \\
&= \Phi\hat{\mu} + \hat{\sigma}^{-2}\Phi\hat{\Sigma}\Phi^T\hat{B}\Phi\hat{\mu} - \hat{\sigma}^{-2}\Phi\hat{\Sigma}\Phi^T\hat{B}(\hat{\sigma}^2\hat{B}^{-1} + \Phi\hat{\Sigma}\Phi^T)(\hat{\sigma}^2\hat{B}^{-1} + \Phi\hat{\Sigma}\Phi^T)^{-1}\Phi\hat{\mu} \\
&= \Phi\hat{\mu}.
\end{aligned}
$$

Therefore, we obtain

$$h(\mathbf{z}|\mathbf{y}, \hat{\alpha}, \hat{\beta}, \hat{\sigma}^2) = N(\mathbf{z}; \mu_p, \Sigma_p),$$

where $\mu_p = \Phi\hat{\mu}$ and $\Sigma_p = \hat{\sigma}^2\hat{B}^{-1} + \Phi\hat{\Sigma}\Phi^T$.

B. Derivation of predictive information criterion

We assume that $q(\mathbf{z}) = N(\mathbf{z}|\mu_q, \Sigma_q)$, where $\mu_q = \Phi\mathbf{w}_q$ and $\Sigma_q = \sigma_q^2 B_q^{-1}$, and obtain

$$
\begin{aligned}
&E_{p(\mathbf{y}|\theta_q)}\left[(\mathbf{y} - \mu_q)(\mu_q - \mu_p)^T\right] \\
&= E_{p(\mathbf{y}|\theta_q)}\left[(\mathbf{y} - \mu_q)\left\{(\mu_q - \mathbf{y}) + (\mathbf{y} - H_p\mathbf{y})\right\}^T\right] \\
&= -\sigma_q^2 B_q^{-1} + E_{p(\mathbf{y}|\theta_q)}\left[(\mathbf{y} - \mu_q)\mathbf{y}^T\right](I - H_p)^T \\
&= -\sigma_q^2 B_q^{-1} + E_{p(\mathbf{y}|\theta_q)}\left[(\mathbf{y} - \mu_q)(\mathbf{y} - \mu_q)^T\right](I - H_p)^T \\
&= -\sigma_q^2 B_q^{-1} + \sigma_q^2 B_q^{-1}(I - H_p^T) \\
&= -\sigma_q^2 B_q^{-1} H_p^T.
\end{aligned}
$$

Acknowledgements

We thank the anonymous reviewer for his careful reading of the manuscript and his helpful comments. S. Kawano was supported by Grants-in-Aid for Scientific Research on Innovative Areas (16H06429, 16K21723, and 16H06430).

References

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

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

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.

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.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer.

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

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

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.

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**, 339–353.

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

de Boor, C. (2001). *A Practical Guide to Splines. Revised Edition*. Springer.

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

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. Society for Industrial and Applied Mathematics.

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

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

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

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

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

Han, M. and Zhao, Y. (2010). Robust relevance vector machine with noise variance coefficient. In: *Neural Networks (IJCNN), The 2010 International Joint Conference on* (pp. 1–6). IEEE.

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

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer.

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

Kawano, S., Hoshina, I., Shimamura, K. and Konishi, S. (2015). Predictive model selection criteria for Bayesian lasso regression. *Journal of the Japanese Society of Computational Statistics*, **28**, 67–82.

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

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

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

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

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.

Matsuda, K. (2015). Predictive model selection criteria for relevance vector regression models. *Josai Mathematical Monographs*, **8**, 97–113.

Mitra, K., Veeraraghavan, A. and Chellappa, R. (2010). Robust RVM regression using sparse outlier model. In *Computer Vision and Pattern Recognition (CVPR), The 2010 IEEE Conference on* (pp. 1887–1894). IEEE.

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

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

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

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

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

She, Y. and Owen, A. B. (2011). Outlier detection using nonconvex penalized regression. *Journal of the American Statistical Association*, **106**, 626–639.

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

Tipping, M. E. (2000). The relevance vector machine. *Advances in Neural Information Processing Systems*, **12**, 652–658.

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

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer.

Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley.

Received June 4, 2018
Revised December 29, 2018