Journal of the Korean Data & Information Science Society
2009, **20**(5), 949–954

한국데이터정보과학회지

Kernel method for autoregressive data[†]

Joo yong Shim[1] · Jang Taek Lee[2]

[1]Department of Applied Statistics, Catholic University of Daegu
[2]Department of Statistics, Dankook University

Received 15 July 2009, revised 9 September 2009, accepted 18 September 2009

Abstract

The autoregressive process is applied in this paper to kernel regression in order to infer nonlinear models for predicting responses. We propose a kernel method for the autoregressive data which estimates the mean function by kernel machines. We also present the model selection method which employs the cross validation techniques for choosing the hyper-parameters which affect the performance of kernel regression. Artificial and real examples are provided to indicate the usefulness of the proposed method for the estimation of mean function in the presence of autocorrelation between data.

Keywords: Autoregressive process, cross validation function, hyper-parameters, kernel regression.

1. Introduction

A modified version of SVM (Vapnik, 1995, 1998) in a least squares sense has been proposed for classification in Suykens and Vanderwalle (1999). In LS-SVM concerning classification problems, we have regression interpretations and direct links to work in classical statistics. In least squares support vector machine (LS-SVM) the solution is given by a linear system instead of a quadratic programming. The fact that LS-SVM has explicit primal-dual formulations has lots of advantages. For the application of LS-SVM the error terms are needed to be independently and identically distributed (error terms are iid).

Most nonparametric regression methods focus on estimating the mean function for various data types (Kim *et al.*, 2008; Shim and Seok, 2008). The estimation of mean function from a data set is usually performed under the assumption that the error terms are iid (Juditsky *et al.*, 1995). This assumption is not satisfied when the correlation is present in the given data (e.g. time series data), which leads to severe problems on the estimation of a model under the iid assumption.

[†] This research was supported by Korea SW Industry Promotion Agency (KIPA) under the program of Software Engineering Technologies Development and Experts Education.
[1] Adjunct Professor, Department of Applied Statistics, Catholic University of Daegu, Kyungbuk 702-701, Korea.
[2] Corresponding author: Professor, Department of Statistics, Dankook University, 126, Jukjeon-dong, Suji-gu, Yongin-si, Gyeonggi-do 448-701, Korea. E-mail: jtlee@dankook.ac.kr

In this paper, we consider the autoregressive model, where $y_t - \mu(\boldsymbol{x}_t)$ follows AR(p) process and $\boldsymbol{x}_t$ is the covariate vector including a constant 1. We propose a kernel method to take the autocorrelation into account and estimate the mean function under the AR model. The rest of this paper is organized as follows. The kernel method for autoregressive data is introduced in Section 2. In Section 3, the generalized cross validation function is given for the model selection. Also estimation method for AR(1) and AR(2) coefficient is presented. In Section 4 we perform the numerical studies through artificial and real examples. In Section 5 we give the conclusions.

2. Kernel method for autoregressive data

Let the given data set be denoted by $\{\boldsymbol{x}_t, y_t\}_{t=1}^n$, with $\boldsymbol{x}_t \in \boldsymbol{R}^d$ including a constant 1 and $y_t \in \boldsymbol{R}$, we consider the autoregressive model,

$$\Phi^p(B)(y_t - \mu(\boldsymbol{x}_t)) = e_t, \quad t = 1, 2, \cdots, n, \tag{2.1}$$

where $\Phi^p(B)$ is a polynomial in back-shift operator B with parameters ρ_i, $i = 1, \cdots, p$, such that $\Phi^p(B)y_t = y_t - \rho_1 y_{t-1} - \rho_2 y_{t-2} - \cdots - \rho_p y_{t-p}$, and e_t is assumed to follow independently normal distribution $(0, \sigma^2)$. Throughout this paper we assume that y_t's are known to follow AR(p) process such that $\Phi^0(B)(y_1 - \mu(\boldsymbol{x_1})) = e_1$, $\Phi^1(B)(y_2 - \mu(\boldsymbol{x}_2)) = e_2$, $\cdots$, $\Phi^{p-1}(B)(y_p - \mu(\boldsymbol{x}_p)) = e_p$, $\Phi^p(B)(y_t - \mu(\boldsymbol{x}_t)) = e_t$, $t = p+1, \ \ p+2, \cdots, n$. The mean function of y_t given $\boldsymbol{x}_t$ is given as $E(y_t|\boldsymbol{x}_t) = \mu(\boldsymbol{x}_t)$.

The negative log likelihood of the given data can be expressed as (constant terms are omitted)

$$L(\boldsymbol{\mu}) = \sum_{t=1}^{p} (\Phi^{t-1}(B)(y_t - \mu(\boldsymbol{x}_t)))^2 + \sum_{t=p+1}^{n} (\Phi^p(B)(y_t - \mu(\boldsymbol{x}_t)))^2 \tag{2.2}$$

The mean function can be estimated by a linear model, $\mu(\boldsymbol{x}) = \boldsymbol{\omega}'\phi(\boldsymbol{x})$, conducted in a high dimensional feature space. Here the feature mapping function $\phi(\cdot) : R^d -> R^{d_f}$ maps the input space to the higher dimensional feature space where the dimension d_f is defined in an implicit way. It is well known that $\phi(\boldsymbol{x_i})'\phi(\boldsymbol{x_j}) = K(\boldsymbol{x_i}, \boldsymbol{x_j})$, which are obtained from the application of Mercer's conditions (1909). Then the estimates of $\boldsymbol{\omega}$ is obtained by minimizing the regularized negative log likelihood,

$$L(\boldsymbol{\omega}) = \sum_{t=1}^{p} (\Phi^{t-1}(B)(y_t - \boldsymbol{\omega}'\phi(\boldsymbol{x}_t)))^2 + \sum_{t=p+1}^{n} (\Phi^p(B)(y_t - \boldsymbol{\omega}'\phi(\boldsymbol{x}_t)))^2 + \frac{\lambda}{2}||\boldsymbol{\omega}||^2 \tag{2.3}$$

where λ is a nonnegative constant which controls the tradeoff between the goodness-of-fit on the data and $||\boldsymbol{\omega}||^2$. The representation theorem (Kimeldorf and Wahba, 1971) guarantees that the minimizer of the regularized negative log likelihood to be $\mu(\boldsymbol{x}) = K\boldsymbol{\alpha}$ for some vector $\boldsymbol{\alpha}$. For AR(1) the likelihood (2.3) is written as

$$\begin{aligned} L(\boldsymbol{\alpha}) = &\sum_{t=2}^{n} (y_t - K(\boldsymbol{x}_t, \boldsymbol{x})\boldsymbol{\alpha} - \rho(y_{t-1} - K(\boldsymbol{x}_{t-1}, \boldsymbol{x})\boldsymbol{\alpha})^2 \\ &+ (y_1 - K(\boldsymbol{x}_1, \boldsymbol{x})\boldsymbol{\alpha})^2 + \frac{\lambda}{2}||\boldsymbol{\omega}||^2. \end{aligned} \tag{2.4}$$

For AR(2) the likelihood (2.3) is written as

$$L(\boldsymbol{\alpha}) = \sum_{t=3}^{n}(y_t - K(\boldsymbol{x}_t,\boldsymbol{x})\boldsymbol{\alpha} - \rho_1(y_{t-1} - K(\boldsymbol{x}_{t-1},\boldsymbol{x})\boldsymbol{\alpha} - \rho_2(y_{t-2} - K(\boldsymbol{x}_{t-2},\boldsymbol{x})\boldsymbol{\alpha})^2 + (y_1 - K(\boldsymbol{x}_1,\boldsymbol{x})\boldsymbol{\alpha})^2 + (y_2 - K(\boldsymbol{x}_2,\boldsymbol{x})\boldsymbol{\alpha} - \rho_1(y_1 - K(\boldsymbol{x}_1,\boldsymbol{x})\boldsymbol{\alpha})^2 + \frac{\lambda}{2}||\boldsymbol{\omega}||^2 \quad (2.5)$$

Now the problem (2.3) becomes obtaining $\boldsymbol{\alpha}$ to minimize

$$L(\boldsymbol{\alpha}) = (\boldsymbol{y}^* - K^*\boldsymbol{\alpha})'(\boldsymbol{y}^* - K^*\boldsymbol{\alpha}) + \frac{\lambda}{2}\boldsymbol{\alpha}' K\boldsymbol{\alpha}, \quad (2.6)$$

where

$$\boldsymbol{y}^* = \begin{pmatrix} y_1 \\ y_2 - \rho y_1 \\ \vdots \\ y_n - \rho y_{n-1} \end{pmatrix},\ K^* = \begin{pmatrix} K_1 \\ K_2 - \rho K_1 \\ \vdots \\ K_n - \rho K_{n-1} \end{pmatrix} for\ AR(1),$$

$$\boldsymbol{y}^* = \begin{pmatrix} y_1 \\ y_2 - \rho_1 y_1 \\ y_3 - \rho_1 y_2 - \rho_2 y_1 \\ \vdots \\ y_n - \rho_1 y_{n-1} - \rho_2 y_{n-2} \end{pmatrix},\ K^* = \begin{pmatrix} K_1 \\ K_2 - \rho_1 K_1 \\ K_3 - \rho_1 K_2 - \rho_2 K_1 \\ \vdots \\ K_n - \rho_1 K_{n-1} - \rho_2 K_{n-2} \end{pmatrix} for\ AR(2),$$

and K_i is the i -th row of K such that $K_i = K(\boldsymbol{x}_i,\boldsymbol{x})$.
The estimates of $\boldsymbol{\alpha}$ for the mean function can be found as

$$\widehat{\boldsymbol{\alpha}} = (K^*K^* + \lambda K)^{-1}K^*\boldsymbol{y}^*, \quad (2.7)$$

which leads $\widehat{\boldsymbol{\mu}} = K(K^*K^* + \lambda K)^{-1}K^*\boldsymbol{y^*} = A_\mu \boldsymbol{y^*}$.

$\boldsymbol{y}^*$ can be written as $\boldsymbol{y}^* = B_\rho \boldsymbol{y}$ with

$$B_\rho = \begin{pmatrix} 1 & 0 & 0 & & \cdots & 0 \\ -\rho & 1 & 0 & \cdots & \cdots & 0 \\ 0 & -\rho & 1 & 0 & \cdots & 0 \\ & & \vdots & & & \\ 0 & 0 & \cdots & \cdots & 0-\rho & 1 \end{pmatrix} for\ AR(1),$$

$$\begin{pmatrix} 1 & 0 & 0 & & \cdots & 0 \\ -\rho_1 & 1 & 0 & \cdots & \cdots & 0 \\ -\rho_2 & -\rho_1 & 1 & 0 & \cdots & 0 \\ & & \vdots & & & \\ 0 & 0 & \cdots & \cdots & -\rho_2-\rho_1 & 1 \end{pmatrix} for\ AR(2).$$

Given $\boldsymbol{x}_t$, the predicted value of $\mu(\boldsymbol{x}_t)$ is obtained as

$$\widehat{\mu}(\boldsymbol{x}_t) = K(\boldsymbol{x}_t,\boldsymbol{x})(K^*K^* + \lambda K)^{-1}K^*B_\rho \boldsymbol{y}. \quad (2.8)$$

3. AR coefficient estimation and model selection

In the mean function estimation, we should find the optimal values of hyper-parameters (λ and other tuning parameters included in the kernel K) and the estimate of ρ or (ρ_1 and ρ_2) for the estimation of $\widehat{\boldsymbol{\mu}} = K\widehat{\boldsymbol{\alpha}}$. We denote a set of hyper-parameters by $\boldsymbol{\theta}$.
Under the assumption that the estimate of ρ or (ρ_1 and ρ_2) is given, the optimal values of hyper-parameters can be chosen by minimizing the generalized cross validation function (Golub *et al.*, 1979):

$$GCV(\boldsymbol{\theta}) = \frac{n\boldsymbol{y}(I - A_\mu B_\rho)^2\boldsymbol{y}}{(n - tr(A_\mu B_\rho))^2}. \tag{3.1}$$

Under the assumption that the optimal values of hyper-parameters are given, the estimate of ρ for AR(1) is obtained by the conditional least squares method as follows,

$$\widehat{\rho} = \frac{\sum_{t=2}^{n}(y_t - \widehat{\mu}(\boldsymbol{x}_t))(y_{t-1} - \widehat{\mu}(\boldsymbol{x}_{t-1}))}{\sum_{t=2}^{n}(y_{t-1} - \widehat{\mu}(\boldsymbol{x}_{t-1}))^2}, \tag{3.2}$$

where $\widehat{\mu}(\boldsymbol{x}_t)$ is the estimate of $\mu(\boldsymbol{x}_t)$ given the previous estimate of ρ and the optimal values of hyper-parameters obtained from GCV function (3.1).
For AR(2) the estimates of ρ_1 and ρ_2 can be obtained as respectively,

$$\widehat{\rho}_1 = \frac{r_1 - r_1 r_2}{1 - r_1^2} \text{ and } \widehat{\rho}_2 = \frac{r_2 - r_1^2}{1 - r_1^2}, \tag{3.3}$$

where

$$r_1 = \frac{\sum_{t=2}^{n}(y_t - \widehat{\mu}(\boldsymbol{x}_t))(y_{t-1} - \widehat{\mu}(\boldsymbol{x}_{t-1}))}{\sum_{t=1}^{n}(y_t - \widehat{\mu}(\boldsymbol{x}_t))^2} \text{ and } r_2 = \frac{\sum_{t=3}^{n}(y_t - \widehat{\mu}(\boldsymbol{x}_t))(y_{t-2} - \widehat{\mu}(\boldsymbol{x}_{t-2}))}{\sum_{t=1}^{n}(y_t - \widehat{\mu}(\boldsymbol{x}_t))^2}.$$

Thus the optimal values of hyper-parameters for the mean estimation and the estimate of ρ or (ρ_1 and ρ_2) are obtained iteratively as follows:

1. Set the initial value of ρ or (ρ_1 and ρ_2).
2. Obtain the optimal values of the hyper-parameters from GCV function (3.1).
3. Obtain the estimate of ρ or (ρ_1 and ρ_2) from (3.2) or (3.3).
4. Reiterate 2.-3. until convergence.

4. Numerical studies

We illustrate the performance of the mean estimation method based on the kernel method for autoregressive data through a simulated data set and a real data set.

Example 1. For the first example, we consider the autoregressive model,

$$\begin{aligned} y_1 = \mu(x_1) + e_1, y_2 - \mu(x_2) &= \rho_1(y_1 - \mu(x_1)) + e_2, \\ y_t - \mu(x_t) &= \rho_1(y_{t-1} - \mu(x_{t-1})) + \rho_2(y_{t-2} - \mu(x_{t-2})) + e_t, t = 3, \cdots, 100 \end{aligned}$$

where $\rho_1 = 0.2$, $\rho_2 = -0.7$ $x_t = t/100$, $\mu(x_t) = 1+\sin(2\pi x_t)$, e_t follows a normal distribution $N(0, 0.5^2)$. The Gaussian kernel functions are utilized for the mean function estimation in this example. For the data set of Figure 4.1 (Left), λ and the kernel parameter are selected as 0.1 and 0.25, respectively, by GCV function (3.1) with the final estimates of $(\rho_1, \rho_2) =$ (0.1846, -0.7235). Figure 4.1 (Left) shows true mean functions (solid line) and estimated mean functions (dashed line) by proposed method, and estimated mean functions (dotted line) by LS-SVM which assumes iid errors, imposed on the scatter plots of 100 data points of y_t's in a data set. In Figure 4.1 (Left) we can see that the proposed method seems to represent the behavior of mean function of given data better than LS-SVM. We repeated the above procedure 100 times (we generated 100 data sets) to have the root mean squared error (RMSE)s for the true mean functions as follows,

$$RMSE_\mu = \sqrt{\frac{1}{100}\sum_{t=1}^{100}(\mu_t - \widehat{\mu}_t)^2}.$$

For the proposed method we obtained the average of 100 $RMSE_\mu$'s and their standard error as 0.0859 and 0.0021, respectively. For LS-SVM we obtained the average of 100 $RMSE_\mu$'s and their standard error as 0.1163 and 0.0039, respectively. The smaller values of $RMSE_\mu$ s indicate that the proposed method works better than LS-SVM on the mean function estimation in this example.

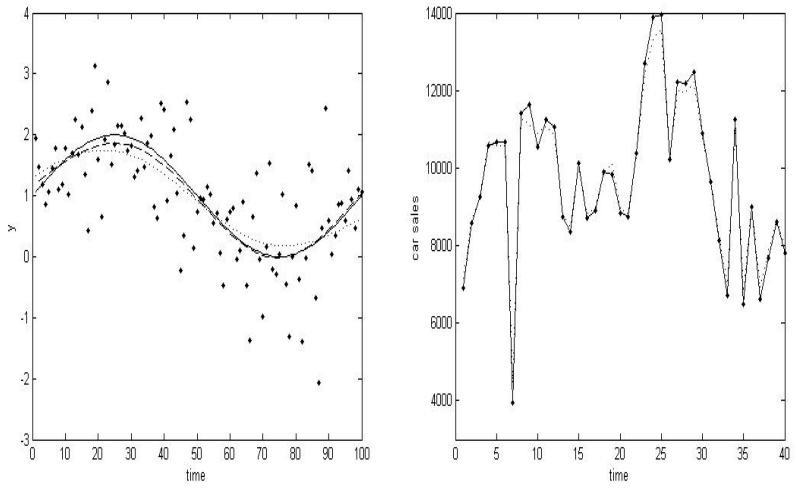

Figure 4.1 Mean function estimation for Example 1 (Left) and Example 2 (Right)

Example 2. For the second example, we use the car sales data set of the brand name Sonata which is available at www.autotimes.co.kr. The data set consists of the number of Sonatas sold in each month from January 2006 to April 2009. We assumed AR(1) model for this data set.

$$y_1 = \mu(\boldsymbol{x}_1) + e_1, \quad y_t - \mu(x_t) = \rho(y_{t-1} - \mu(\boldsymbol{x}_{t-1})) + e_t, \quad t = 2, \cdots, 40$$

where y_t is the standardized car sales (Y_t), x_{t1} is the Sonata's price at t -th month and x_{t2} is KOSPI (Korea composite stock price index) at t -th month which is assumed to affect the car sales. The Gaussian kernel functions are utilized for the mean function estimation in this example. λ and the kernel parameter are selected as 0.001 and 0.25, respectively. Figure 4.1 (Right) shows true mean functions (solid line) and estimated mean functions (dashed line) by proposed method, and estimated mean functions (dotted line) by LS-SVM which assumes iid errors, imposed on the scatter plots of 40 data points of Y_t's in a data set. The estimate of ρ is obtained as 0.1877, which implies that the previous car sales provides a small positive effect on the present car sales. We obtained the root mean squared error (RMSE) such as

$$RMSE_\mu = \sqrt{\frac{1}{40}\sum_{t=1}^{40}(Y_t - \widehat{Y}_t)^2}.$$

By the proposed method we obtained $RMSE_\mu$ as 4.9432 and 226.0812 by LS-SVM. The smaller value of $RMSE_\mu$ indicates that the proposed method works better than LS-SVM on the mean function estimation in this example.

5. Conclusions

In this paper, we dealt with estimating the mean functions for autoregressive model and obtained cross validation functions for the proposed method. Through the examples we showed that the proposed method yields the satisfying results. We also found that the proposed method has an advantage of an easy model selection method such as GCV function. Also an advantage of easy extension to the heteroscedastic autoregressive model by incorporating a doubly penalizing method.

References

Golub, G. H., Heath, M. and Wahba, G. (1979). Generalized cross validation as a method for choosing a good ridge parameter. *Technometrics,* **21**, 215-223.

Juditsky, A, Hjalmarsson, H., Benveniste, A., Deylon, B., Ljung, L., Sjöberg, J. and Zhang, Q. (1995). Nonlinear Black-box modelling in system identification: Mathematical foundations. *Automatica,* **31**, 1725-1750.

Kim, M. S., Park, H. J., Hwang, C., and Shim, J. (2008). Claims reserving via Kernel machine. *Journal of Korean Data & Information Science Society,* **19**, 1419-1417.

Kimeldorf, G. S. and Wahba, G. (1971). Some results on tchebycheffian spline functions. *Journal of Mathematical Analysis and its Applications,* **33**, 82-95.

Mercer, J. (1909). Functions of positive and negative type and their connection with theory of integral equations. *Philosophical Transactions of Royal Society,* **A**, 415-446.

Shim, J. and Seok, K. H. (2008). Kernel poisson regression for longitudinal data. *Journal of Korean Data & Information Science Society,* **19**, 1353-1360.

Suykens, J. A. K. and Vanderwalle, J. (1999). Least square support vector machine classifier. *Neural Processing Letters,* **9**, 293-300.

Vapnik, V. N. (1995). *The nature of statistical learning theory,* Springer, New York.

Vapnik, V. N. (1998). *Statistical learning theory,* John Wiley, New York.