*Journal of the Korean Data & Information Science Society*
*2010,* **21**(1), 155–162

한국데이터정보과학회지

# Doubly penalized kernel method for heteroscedastic autoregressive data[†]

Daehyeon Cho[1] · Jooyong Shim[2] · Kyung Ha Seok[3]

[13]Department of Data Science, Institute of Statistical Information, Inje University
[2]Department of Applied Statistics, Catholic University of Daegu

Received 15 October 2009, revised 2 January 2010, accepted 8 January 2010

**Abstract**

In this paper we propose a doubly penalized kernel method which estimates both the mean function and the variance function simultaneously by kernel machines for heteroscedastic autoregressive data. We also present the model selection method which employs the cross validation techniques for choosing the hyper-parameters which affect the performance of proposed method. Simulated examples are provided to indicate the usefulness of proposed method for the estimation of mean and variance functions.

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

## 1. Introduction

The estimation of a model from a data set is usually performed under the assumption that the error terms are independently and identically distributed (iid) (Juditsky *et al.*, 1995). Most nonparametric regression methods focus on estimating the mean function with iid assumption. This assumption is not satisfied when the correlation presents in the given data set, which leads to severe problems on the estimation of a model under the iid assumption.

It becomes an important issue in many fields including the estimation of the variance function (Anderson and Lund, 1997; Liu *et al.*, 2007; Shim *et al.*, 2009) and most of them are focused on heteroscedastic regression problems. The variance is estimated based on the regression residuals which are differences of responses and estimated means (Ruppert *et al.*, 1997; Fan and Yao, 1998). A penalized likelihood based the normal distribution to estimate both the mean and the variance simultaneously has been proposed by Yuan and Wahba (2004).

[†] This work was supported by the Korea Research Foundation (KRF) grant funded by the Korea government (MEST) (No.2009-0072211).
[1] Department of Data Science, Institute of Statistical Information, Inje University, Kimhae 621-749, Korea.
[2] Department of Applied Statistics, Catholic University of Daegu, Gyungbuk 712-702, Korea.
[3] Corresponding author: Department of Data Science, Institute of Statistical Information, Inje University, Kimhae 621-749, Korea. E-mail: statskh@inje.ac.kr.

In this paper, we consider the heteroscedastic autoregressive model, where $\boldsymbol{x}_t$ is the covariate vector including a constant 1, $y_t - \mu(\boldsymbol{x}_t)$ is assumed to follow AR(p) process, and the error term $e_t$ is assumed to follow a normal distribution $(0, \sigma^2(\boldsymbol{x}_t))$. We propose a doubly penalized kernel method (DPKM) for heteroscedastic autoregressive data to take the heteroscedasticity into account and estimate both the mean and the variance functions simultaneously under the AR model. The kernel trick is applied to DPKM, which has been applied to the regression problems of various data types since it was firstly introduced in Aizerman *et al.* (1964).

The rest of this paper is organized as follows. The DPKM is introduced in Section 2, the mean function is estimated from the linear system and the variance is obtained by Newton-Raphson method. In Section 3, cross validation functions are given for the model selections of the mean and the variance functions estimations. Also estimation method for autoregressive coefficient is presented. In Section 4 we perform the numerical studies through examples. In Section 5 we give the conclusions.

## 2. Mean and variance functions estimation

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

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

where $\Phi(B)$ is a polynomial in back-shift operator $B$ with parameters $\rho_i$, $i = 1, \cdots, p$, such that $\Phi(B)y_t = y_t - \rho_1 y_{t-1} - \rho_2 y_{t-2} - \cdots - \rho_p y_{t-p}$, and $e_t$ is assumed to follow independently normal distribution $(0, \sigma^2(\boldsymbol{x}_t))$. For the convenience we assume that $y_t$'s are known to follow AR(1) process throughout this paper, which is, $y_1 = \mu(\boldsymbol{x}_1) + e_1$ and $y_t = \mu(\boldsymbol{x}_t) + \rho(y_{t-1} - \mu(\boldsymbol{x}_{t-1})) + e_t$, $t = 2, 3, \cdots, n$. Given $\boldsymbol{x}_t$, the mean function and the variance function of $y_t$ are given as follows.

$$E(y_t|\boldsymbol{x}_t) = \mu(\boldsymbol{x}_t), Var(y_t|\boldsymbol{x}_t) = \rho^2 Var(y_{t-1}|\boldsymbol{x}_{t-1}) + \sigma^2(\boldsymbol{x}_t). \tag{2.2}$$

Here $\mu(\boldsymbol{x}_t)$ and $\sigma^2(\boldsymbol{x}_t)$ are functions to be estimated. $Var(y_t|\boldsymbol{x}_t)$ is estimated as $\widehat{Var}(y_t|\boldsymbol{x}_t) = \widehat{\rho}^2 \widehat{Var}(y_{t-1}|\boldsymbol{x}_{t-1}) + \widehat{\sigma}^2(\boldsymbol{x}_t)$ with $\widehat{V}ar(y_1|\boldsymbol{x}_1) = \widehat{\sigma}^2(\boldsymbol{x}_1)$. The negative log likelihood of the given data set can be expressed as (constant terms are omitted)

$$\begin{aligned} L(\boldsymbol{\mu}, \boldsymbol{\sigma}^2) = &\sum_{t=2}^{n} (y_t - \mu(\boldsymbol{x}_t) - \rho(y_{t-1} - \mu(\boldsymbol{x}_{t-1})))^2 \frac{1}{\sigma^2(\boldsymbol{x}_t)} \\ &+ (y_1 - \mu(\boldsymbol{x}_1))^2 \frac{1}{\sigma^2(\boldsymbol{x}_1)} + \sum_{t=1}^{n} \log \sigma^2(\boldsymbol{x}_t). \end{aligned} \tag{2.3}$$

Due to the positivity of the variance function we write the logarithm of $\sigma^2(\boldsymbol{x}_t)$ as $g(\boldsymbol{x}_t)$ , then the negative log likelihood (2.3) can reexpressed as

$$\begin{aligned} L(\boldsymbol{\mu}, \boldsymbol{g}) = &\sum_{t=2}^{n} (y_t - \mu(\boldsymbol{x}_t) - \rho(y_{t-1} - \mu(\boldsymbol{x}_{t-1})))^2 e^{-g(\boldsymbol{x}_t)} \\ &+ (y_1 - \mu(\boldsymbol{x}_1))^2 e^{-g(\boldsymbol{x}_1)} + \sum_{t=1}^{n} g(\boldsymbol{x}_t). \end{aligned} \tag{2.4}$$

Kernel methods are used widely for function estimation (Hwang, 2008; Shim and Seok, 2008). Among kernel methods, the mean function is estimated by a linear model, $\mu(\boldsymbol{x}) = \boldsymbol{\omega}_\mu'\phi_\mu(\boldsymbol{x})$, conducted in a high dimensional feature space, in this paper. Here the feature mapping function $\phi_\mu(\cdot) : R^d \to R^{d_f}$ 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 $\phi_\mu(\boldsymbol{x}_i)'\phi_\mu(\boldsymbol{x}_j) = K_\mu(\boldsymbol{x}_i, \boldsymbol{x}_j)$ which are obtained from the application of Mercer's conditions (1909). Also $g$ is estimated by a linear model, $g(\boldsymbol{x}) = \boldsymbol{\omega}_g'\phi_g(\boldsymbol{x})$.

Then the estimates of ( $\boldsymbol{\omega}_\mu, \boldsymbol{\omega}_g$ ) are obtained by minimizing the regularized negative log likelihood,

$$L(\boldsymbol{\omega}_\mu, \boldsymbol{\omega}_g) = \sum_{t=2}^{n}(y_t - \boldsymbol{\omega}_\mu'\phi_\mu(\boldsymbol{x}_t) - \rho(y_{t-1} - \boldsymbol{\omega}_\mu'\phi_\mu(\boldsymbol{x}_{t-1}))^2 e^{-\boldsymbol{\omega}_g'\phi_g(\boldsymbol{x}_t)} \tag{2.5}$$
$$+ (y_1 - \boldsymbol{\omega}_\mu'\phi_\mu(\boldsymbol{x}_1))^2 e^{-\boldsymbol{\omega}_g'\phi_g(\boldsymbol{x}_1)} + \sum_{t=1}^{n}\boldsymbol{\omega}_g'\phi_g(\boldsymbol{x}_t) + \lambda_\mu||\boldsymbol{\omega}_\mu||^2 + \lambda_g||\boldsymbol{\omega}_g||^2$$

where $\lambda_\mu$ ( $\lambda_g$ ) is a nonnegative constant which controls the tradeoff between the goodness-of-fit on the data and $||\boldsymbol{\omega}_\mu||^2$ ( $||\boldsymbol{\omega}_g||^2$ ). The representation theorem (Kimeldorf and Wahba, 1971) guarantees that the minimizer of the regularized negative log likelihood to be $\mu(\boldsymbol{x}) = K_\mu\boldsymbol{\alpha}_\mu$ and $g(\boldsymbol{x}) = K_g\boldsymbol{\alpha}_g$ for some vectors $\boldsymbol{\alpha}_\mu$ and $\boldsymbol{\alpha}_g$.

Now the problem (2.5) becomes that of obtaining ( $\boldsymbol{\alpha}_\mu, \boldsymbol{\alpha}_g$ ) to minimize

$$L(\boldsymbol{\alpha}_\mu, \boldsymbol{\alpha}_g) = \quad (\boldsymbol{y}^* - K_\mu^*\boldsymbol{\alpha}_\mu)'D_g^{-1}(\boldsymbol{y}^* - K_\mu^*\boldsymbol{\alpha}_\mu) + \mathbf{1}'\boldsymbol{g} + \lambda_\mu\boldsymbol{\alpha}_\mu{}'K_\mu\boldsymbol{\alpha}_\mu + \lambda_g\boldsymbol{\alpha}_g{}'K_g\boldsymbol{\alpha}_g, \tag{2.6}$$

where

$$\boldsymbol{y}^* = \begin{pmatrix} y_1 \\ y_2 - \rho y_1 \\ \vdots \\ y_n - \rho y_{n-1} \end{pmatrix}, K_\mu^* = \begin{pmatrix} K_{\mu,1} \\ K_{\mu,2} - \rho K_{\mu,1} \\ \vdots \\ K_{\mu,n} - \rho K_{\mu,n-1} \end{pmatrix},$$

$K_{\mu,t}$ is the $t$ -th row of $K_\mu$, $D_g$ is a diagonal matrix of $e^{g(\boldsymbol{x})} = e^{K_g\boldsymbol{\alpha}_g}$, and $\mathbf{1}$ is a $(n \times 1)$ vector with 1's.

The estimates of parameters $(\boldsymbol{\alpha}_\mu, \boldsymbol{\alpha}_g)$ for the mean and variance functions can be found via an iterative procedure, updating the mean function and the variance function alternatively.

Fixing $\boldsymbol{g} = \widehat{\boldsymbol{g}}$, the regularized negative log likelihood (2.6) reduces to

$$L(\boldsymbol{\alpha}_\mu) = \frac{1}{2}(\boldsymbol{y}^* - K_\mu^*\boldsymbol{\alpha}_\mu)'D_g^{-1}(\boldsymbol{y}^* - K_\mu^*\boldsymbol{\alpha}_\mu) + \frac{\lambda_\mu}{2}\boldsymbol{\alpha}_\mu'K_\mu\boldsymbol{\alpha}_\mu \tag{2.7}$$

The solution to (2.7) is

$$\widehat{\boldsymbol{\alpha}}_\mu = (K_\mu^*D_g^{-1}K_\mu^* + \lambda_\mu K_\mu)^{-1}K_\mu^*D_g^{-1}\boldsymbol{y}^*, \tag{2.8}$$

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

Here $\boldsymbol{y}^*$ can be written as

$$\boldsymbol{y}^* = B_\rho \boldsymbol{y} = \begin{pmatrix} 1 & 0 & 0 & & \cdots & 0 \\ -\rho & 1 & 0 & \cdots & \cdots & 0 \\ 0 & -\rho & 1 & 0 & \cdots & 0 \\ & & \vdots & & & \\ 0 & 0 & \cdots & \cdots & 0-\rho & 1 \end{pmatrix} \boldsymbol{y}, \tag{2.9}$$

so that $\widehat{\boldsymbol{\mu}}$ can be rewritten as $\widehat{\boldsymbol{\mu}} = A_\mu B_\rho \boldsymbol{y}$.

In case of homoscedastic regression model, the estimate of $\boldsymbol{\mu}$ is obtained as

$$\widehat{\boldsymbol{\mu}} = K_\mu (K_\mu^* K_\mu^* + \lambda_\mu K_\mu)^{-1} K_\mu^* \boldsymbol{y}^*, \tag{2.10}$$

which is equivalent to the results of Shim and Lee (2009) with AR(1).

To estimate $g(\boldsymbol{x}_t) = \log \sigma(\boldsymbol{x}_t)^2$ which is the logarithm of the variance of $e_t$, we use the current estimates of $\boldsymbol{\alpha}_\mu$ and $\rho$. The $\boldsymbol{\alpha}_g$ is estimated by minimizing the objective function (regularized negative log likelihood of Gamma distribution of independent $z_t$'s with shape parameter 1 and scale parameters $e^{g(\boldsymbol{x}_t)}$),

$$L(\boldsymbol{\alpha}_g) = \sum_{t=1}^{n} (z_t e^{-g(\boldsymbol{x}_t)} + g(\boldsymbol{x}_t)) + \lambda_g \boldsymbol{\alpha}_g{}' K_g \boldsymbol{\alpha}_g = \mathbf{1}'(D_g^{-1}\boldsymbol{z} + K_g \boldsymbol{\alpha}_g) + \lambda_g \boldsymbol{\alpha}_g{}' K_g \boldsymbol{\alpha}_g, \tag{2.11}$$

where $z_t = (y_t^* - K_{\mu,t}^* \widehat{\boldsymbol{\alpha}}_\mu)^2$. $\widehat{\boldsymbol{\alpha}}_g$ is obtained by Newton-Raphson method, $\widehat{\boldsymbol{\alpha}}_g^{new} = \widehat{\boldsymbol{\alpha}}_g^{old} - H^{-1}G$, where $G$ and $H$ are the gradient vector and Hessian matrix with respect to $\boldsymbol{\alpha}_g$, respectively. Then we have $\widehat{\sigma}^2(\boldsymbol{x}_t) = \exp(\widehat{g}(\boldsymbol{x}_t))$.

Summing up, we describe the algorithm for training and model selection of the DPKM as follows:

(a) With given values of $\widehat{\boldsymbol{g}} = K_g \widehat{\boldsymbol{\alpha}}_g$, find $\widehat{\boldsymbol{\alpha}}_\mu$ from (2.8).

(b) With $\widehat{\boldsymbol{\mu}} = K_\mu \widehat{\boldsymbol{\alpha}}_\mu$ , find $\widehat{\boldsymbol{\alpha}}_g$ from (2.11) using Newton-Raphson method.

(c) Iterate (a) and (b) until convergence.

With the final estimates of $\rho$, $\boldsymbol{\alpha}_\mu$ and $\boldsymbol{\alpha}_g$ we have the estimated mean and variance of $y_t$ as, $\widehat{E}(y_t|\boldsymbol{x}_t) = K_{\mu,t}\widehat{\boldsymbol{\alpha}}_\mu$ and $\widehat{V}ar(y_t|\boldsymbol{x}_t) = \widehat{\rho}^2 \widehat{V}ar(y_{t-1}|\boldsymbol{x}_{t-1}) + \widehat{\sigma}^2(\boldsymbol{x}_t)$ with $\widehat{V}ar(y_1|\boldsymbol{x}_1) = \widehat{\sigma}^2(\boldsymbol{x}_1)$.

## 3. Model selection

The functional structures of the estimation method of the mean and the variance functions are characterized by hyper-parameters, the regularization parameters $\lambda_\mu$, $\lambda_g$ and other tuning parameters included in the kernel.

In the mean function estimation, we should find the optimal values of hyper-parameters ($\lambda_\mu$ and tuning parameter $\gamma_\mu$ included in the kernel $K_\mu$) and the estimate of $\rho$ for the estimation of $\widehat{\boldsymbol{\mu}} = A_\mu B_\rho \boldsymbol{y}$. We denote a set of hyper-parameters by $\boldsymbol{\theta}_\mu = (\lambda_\mu, \gamma_\mu, \rho)$.

Under the assumption that the estimate of $\rho$ is given, the optimal values of hyper-parameters can be chosen by minimizing the generalized cross validation function (Golub *et al.*, 1979):

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

Under the assumption that the optimal values of hyper-parameters are given, the estimate of $\rho$ is obtained by the conditional least squares method as follows,

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

where $\widehat{\mu}(\boldsymbol{x}_t)$ is the estimate of $\mu(\boldsymbol{x}_t)$ given the previous estimate $\rho$ and the optimal values of hyper-parameters obtained from GCV function (3.1).

Thus the optimal values of hyper-parameters for the mean estimation and the estimate of $\rho$ are obtained iteratively as follows:

(a) Set the initial value of $\rho$.

(b) Obtain the optimal values of the hyper-parameters from GCV function (3.1).

(c) Obtain the estimate of $\rho$ from (3.2).

(d) Reiterate (b) and (c) until convergence.

For the model selection of the variance function estimation, the optimal values of hyper-parameters ($\lambda_g$ and other tuning parameters included in the kernel $K_g$) can be chosen by minimizing the generalized approximate cross validation function (Xiang and Wahba, 1996; Liu *et al.*, 2007):

$$GACV(\boldsymbol{\theta}_g) = \frac{1}{n}\mathbf{1}'(D_g^{-1}\boldsymbol{z} + K_g\boldsymbol{\alpha}_g) + \frac{1}{n}\frac{tr(D_g^{1/2}H_gD_g^{1/2})}{n - tr(D_g^{1/2}H_gD_g^{1/2})}(\boldsymbol{z} - \exp(\boldsymbol{g}))'D_g^{-2}\boldsymbol{z}, \tag{3.3}$$

where $H_g = (D_zD_g^{-1} + 2\lambda_gK_g^{-1})^{-1}$ is the inverse of Hessian matrix of (2.11) with respect to $\boldsymbol{g}$ and $D_z$ a diagonal matrix of $\boldsymbol{z}$.

## 4. Numerical studies

We illustrate the performance of the mean and variance estimations method based on the kernel method for autoregressive heteroscedastic data and autoregressive homoscedastic data through two simulated data sets.

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

$$y_1 = \mu(x_1) + e_1 y_t - \mu(x_t) = \rho(y_{t-1} - \mu(x_{t-1})) + e_t, t = 2, \cdots, 100,$$

where $\rho = 0.5$ , $x_t = t/100$ , $\mu(x_t) = 1+\sin(2\pi x_t)$, $e_t \sim N(0, 1.2+\sin(2\pi x_t))$. The Gaussian kernel functions are utilized for both the mean function estimation and the variance function estimation in this example. Figure 4.1 (Left) shows true mean functions (solid line) and estimated mean functions (dashed line) by DPKM, and estimated mean functions (dotted line) by LS-SVM (Least Squares Support Vector Machine, Suykens and Vanderwalle, 1999) which assumes iid errors, imposed on the scatter plots of 100 data points of $y_t$ 's in a data set. Figure 4.1 (Right) shows true variance functions (solid line) and estimated variance functions by DPKM (dashed line) of a data set. In Figure 4.1 (Right) we can see that the estimated variance function by DPKM seems to represent well the behavior of variance function of given data. We repeated the above procedure 100 times to have the root mean

squared errors (RMSE) for the true mean functions and variance functions as follows,

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

For the proposed method we obtained the average of 100 $RMSE_{\mu}$'s and their standard error as 0.4085 and 0.0112, respectively. For LS-SVM we obtained the average of 100 $RMSE_{\mu}$'s and their standard error as 0.4716 and 0.0139, respectively. The smaller values of $RMSE_{\mu}$s indicate that DPKM works better than LS-SVM on the mean function estimation in this example. And we obtained the average of 100 $RMSE_{\sigma^2}$'s and their standard error as 0.7422 and 0.0257, respectively. The average of 100 $\widehat{\rho}$'s and their standard error are obtained as 0.4521 and 0.0126, respectively.

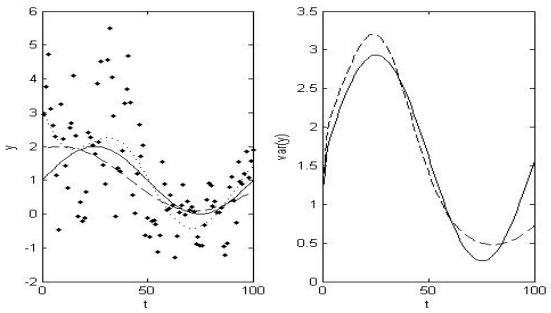


**Figure 4.1** Mean function estimation (Left) and variance function estimation (Right) of a data set in Example 1.

**Example 2.** For the second simulated example, we consider the homosecedastic autoregressive model,

$$y_1 = \mu(x_1) + e_1, y_t - \mu(x_t) = \rho(y_{t-1} - \mu(x_{t-1})) + e_t, t = 2, \cdots, 100$$

where $\rho = 0.5$ , $x_t = t/100$ , $\mu(x_t) = 1 + \sin(2\pi x_t)$ , $e_t \sim N(0, 2)$. The Gaussian kernel functions are utilized for both the mean function estimation and the variance function estimation in this example. Figure 4.2 (Left) shows true mean function (solid line) and estimated mean functions (dashed line) by DPKM, 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. Figure 4.2 (Right) shows true variance functions (solid line) and estimated variance functions (dashed line) of a data set. In Figure 4.2 (Right) we can see that DPKM seems to represent well the behavior of constant variance function of given data.

We repeated the above procedure 100 times to have the root mean squared errors (RMSE) for the true mean functions and variance functions. For DPKM we obtained the average of 100 $RMSE_{\mu}$'s and their standard error as 0.4725 and 0.0146, respectively. For LS-SVM we obtained the average of 100 $RMSE_{\mu}$'s and their standard error as 0.6479 and 0.0168, respectively. The smaller values of $RMSE_{\mu}$s indicate that DPKM works better than LS-SVM on the mean function estimation in this example. And we obtained the average of 100

$RMSE_{\sigma^2}$'s and their standard error as 0.4421 and 0.0302, respectively. The average of 100 $\widehat{\rho}$'s and their standard error are obtained as 0.4145 and 0.0103, respectively.

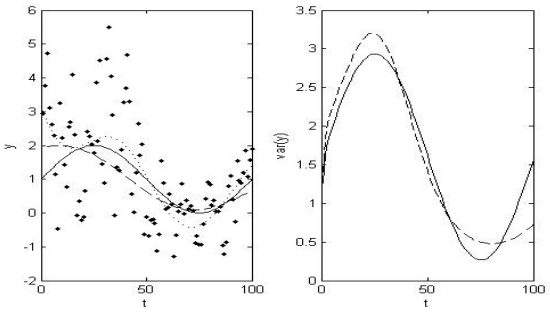


**Figure 4.2** Mean function estimation (Left) and Variance function estimation (Right) of a data set in Example 2.

## 5. Conclusions

In this paper, we dealt with estimating the mean and variance functions for heteroscedastic autoregressive model and obtained cross validation functions for the proposed method. DPKM can be applied even for the homoscedastic autoregressive model. Through the examples we showed that DPKM derives the satisfying results on estimating the mean and variance functions. We also found that DPKM has an advantage of using model selection methods such as GCV function and GACV function.

## References

Aizerman, M. A., Braverman, E. M. and Rozonoer, L. I. (1964). Theoretical foundation of potential function method in pattern recognition learning. *Automation and Remote Control*, **25**, 821-837.

Anderson, T. G. and Lund, J. (1997). Estimating continuous-time stochastic volatility models of short-term interest rate. *Journal of Econometrics*, **77**, 343-377.

Fan, J. Q. and Yao, Q. W. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika*, **85**, 645-660.

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 .

Hwang, C. (2008). Mixed effects kernel binomial regression. *Journal of Korean Data & Information Science Society*, **19**, 1327-1334.

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

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

Liu, A., Tong, T. and Wang, Y. (2007). Smoothing spline estimation of variance functions. *Journal of Computational and Graphical Statistics*, **16**, 312-329.

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.

Ruppert, D., Wand, M. P., Holst, U. and Hossjer, O. (1997). Local polynomial variance-function estimation. *Technometrics*, **39**, 262-73.

Shim, J. and Lee, J. T. (2009). Kernel method for autoregressive data. *Journal of Korean Data and Information Science Society*, **20**, 949 -954 .

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

Shim, J., Park, H. J. and Seok, K. H. (2009). Variance function estimation with LS-SVM for replicated data. *Journal of Korean Data and Information Science Society*, **20**, 925 -931.

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

Xiang, D. and Wahba, G. (1996). A generalized approximate cross validation for smoothing splines with non-gaussian data. *Statistian Sinica*, **6**, 675-692.

Yuan, M. and Wahba, G. (2004). Doubly penalized likelihood estimator in heteroscedastic regression. *Statistics & Probability Letters*, **69**, 11-20.