Communications of the Korean Statistical Society
2009, Vol. 16, No. 2, 383–388

Estimating Variance Function with Kernel Machine

Jong-Tae Kim[a], Changha Hwang[b], Hye Jung Park[c], Jooyong Shim[1,d]

[a]Dept. of Statistics, Daegu Univ., [b]Dept. of Statistics, Dankook Univ.,
[c]Computer Course Div., Daegu Univ., [d]Dept. of Applied Statistics, Catholic Univ. of Daegu

Abstract

In this paper we propose a variance function estimation method based on kernel trick for replicated data or data consisted of sample variances. Newton-Raphson method is used to obtain associated parameter vector. Furthermore, the generalized approximate cross validation function is introduced to select the hyper-parameters which affect the performance of the proposed variance function estimation method. Experimental results are then presented which illustrate thc performance of the proposed procedure.

Keywords: Heteroscedasticity, kernel trick, kernel function, hyper-parameters, generalized approximate cross validation function, variance function.

1. Introduction

It becomes an important issue in many fields modelling the volatility or the local variability, which are usually expressed in terms of variance functions. Researches on estimation of variance function can be found in Anderson and Lund (1997), Hall and Carroll (1989) and Liu *et al.* (2007) and most of them are focused on heteroscedastic error problems. In this paper we propose a variance function estimation method based on kernel trick (Vapnik, 1995) for the heteroscedastic regression problem under the assumption that sample variances follow independently gamma distribution. The kernel trick is a method for using a linear model to solve a nonlinear problem by mapping the input space into a higher-dimensional feature space. This is done by using Mercer's theorem (Mercer, 1909). The estimators are obtained by minimizing the penalized log-likelihood function of sample variances. Here sample variances are obtained from replicated data where errors are assumed to follow a normal distribution. The proposed method enables to select appropriate hyper-parameters easily from the generalized approximate cross validation(GACV) function, which is used to select hyperparameters for the achievement of high generalization performance. The rest of this paper is organized as follows. In Section 2 we propose a variance function estimation method using the principal idea of kernel machine. In Section 3 we present the model selection method using GACV function. In Section 4 we perform the numerical studies through examples. In Section 5 we give the conclusions.

2. Variance Function Estimation

Consider the heteroscedastic regression model with observations with $(k+1)$ replicates as follows,

$$z_{ij} = \mu_i + \epsilon_{ij}, \quad j = 1, \ldots, k+1, \ i = 1, \ldots, n,$$

where μ_i is a mean of z_{ij}'s for $j = 1, \ldots, k+1$, ϵ_{ij} follows a normal distribution $(0, e^{f(x_i)})$, the $\mathbf{x}_i \in \mathbf{R}^d$ is an input vector and $f(\mathbf{x}_i)$ is an unknown nonlinear function. We denote the sample variance obtained

[1] Corresponding author: Department of Applied Statistics, Catholic University of Daegu, Gyungbuk 712-702, Korea. E-mail: ds1631@hanmail.net

from z_{ij}'s for $j = 1, \ldots, k+1$ by y_i and assume that given data set consists of sample variances y_i's which are assumed to follow independently gamma distributions $(k/2,\ 2e^{f(x_i)}/k)$ for $i = 1, \ldots, n$. Here we define the variance function as $v(\mathbf{x}_i) = \mathrm{Var}(z_{ij} \mid \mathbf{x}_i) = E(y_i \mid \mathbf{x}_i) = e^{f(x_i)}$. The negative log-likelihood of the given data can be expressed as(a constant term is omitted)

$$L(\mathbf{f}) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i e^{-f(\mathbf{x}_i)} + f(\mathbf{x}_i)\right). \tag{2.1}$$

The nonlinear function $f(\mathbf{x}_i)$ can be estimated by a linear model, $f(\mathbf{x}_i) = \omega^t \phi(\mathbf{x}_i)$, conducted in a high dimensional feature space. Here the feature mapping function $\phi(\cdot) : R^d \rightarrow R^{d_f}$ maps the input space to the higher dimensional feature space where the dimension d_f is defined in an implicit way. Then the estimate of parameter vector satisfying $f(\mathbf{x}_i) = \omega^t \phi(\mathbf{x}_i)$ for $i = 1, \ldots, n$ is obtained by minimizing the penalized negative log-likelihood,

$$L(\omega) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i e^{-\omega^t \phi(\mathbf{x}_i)} + \omega^t \phi(\mathbf{x}_i)\right) + \frac{\lambda}{2}\|\omega\|^2, \tag{2.2}$$

where λ is a nonnegative regularization parameter which controls the trade-off between the goodness-of-fit on the data and $\|\omega\|^2$. It is known that $\phi(\mathbf{x}_i)\phi^t(\mathbf{x}_j) = K(\mathbf{x}_i, \mathbf{x}_j)$ which are obtained from the application of Mercer's conditions (Mercer, 1909). The representation theorem (Kimeldorf and Wahba, 1971) guarantees the minimizer of the penalized negative log-likelihood to be $f(\mathbf{x}_i) = \mathbf{K}_i\alpha$ for some $n \times 1$ vector α, where $\mathbf{K}_i$ is the i^{th} row of the $n \times n$ kernel matrix $\mathbf{K}$ with elements $K(\mathbf{x}_i, \mathbf{x}_j)$, $i, j = 1, \ldots, n$. Now the penalized negative log-likelihood (2.2) becomes

$$L(\alpha) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i e^{-\mathbf{K}_i\alpha} + \mathbf{K}_i\alpha\right) + \frac{\lambda}{2}\alpha^t \mathbf{K}\alpha. \tag{2.3}$$

The penalized negative log-likelihood (2.3) can be reexpressed as

$$L(\alpha) = \frac{1}{n}\left(\mathbf{y}^t e^{-\mathbf{K}\alpha} + \mathbf{1}_n^t \mathbf{K}\alpha\right) + \frac{\lambda}{2}\alpha^t \mathbf{K}\alpha, \tag{2.4}$$

where e is the componentwise exponential function and $\mathbf{1}_n$ is the $n \times 1$ vector of ones. By minimizing the penalized negative log-likelihood (2.4) we obtain the estimate of parameter vector α, but not in a explicit form, which leads to use the Newton-Raphson method. At each iteration the parameter vector α is updated as follows,

$$\alpha^{new} = \alpha - \mathbf{H}^{-1}\mathbf{G}, \tag{2.5}$$

where $\mathbf{G}$ is the gradient vector and $\mathbf{H}$ is the Hessian matrix of (2.4). With the estimate of parameter vector $\hat{\alpha}$, the predicted variance function given the input vector $\mathbf{x}_0$ is obtained as follows,

$$\hat{v}(\mathbf{x}_0) = e^{\hat{f}(\mathbf{x}_0)} = e^{\mathbf{K}_0\hat{\alpha}}, \tag{2.6}$$

where $\mathbf{K}_0$ is the $1 \times n$ row vector with elements $K(\mathbf{x}_0, \mathbf{x}_j)$, $j = 1, \ldots, n$.

3. Model Selection

The functional structure of the estimation method of variance function is characterized by hyper-parameters, the regularization parameter λ and the kernel parameter. For the model selection of the estimation method of variance function, we define the leave-one-out cross validation(CV) function (Xiang and Wahba, 1996) for a set of hyper-parameters $\boldsymbol{\theta}$, as follows,

$$\mathrm{CV}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i e^{-\hat{f}_\theta^{(-i)}(\mathbf{x}_i)} + \hat{f}_\theta(\mathbf{x}_i)\right), \tag{3.1}$$

where $\hat{f}_\theta(\mathbf{x}_i)$ is the estimate of $f_\theta(\mathbf{x}_i)$ from full data and $\hat{f}_\theta^{(-i)}(\mathbf{x}_i)$ is the estimate of $f_\theta(\mathbf{x}_i)$ from data without i^{th} observation. Since for each candidate of hyperparameter sets, n of $\hat{f}_\theta^{(-i)}(\mathbf{x}_i)'$s should be computed, selecting parameters using CV function is computationally formidable. Using Xiang and Wahba (1996) and Liu *et al.* (2007) we have the approximate cross validation(ACV) function as follows,

$$\mathrm{ACV}(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) + \frac{1}{n}\sum_{i=1}^{n}\frac{s_{ii}e^{-\hat{f}_\theta(\mathbf{x}_i)}y_i\left(y_i - e^{\hat{f}_\theta(\mathbf{x}_i)}\right)}{1 - s_{ii}e^{\hat{f}_\theta(\mathbf{x}_i)}}, \tag{3.2}$$

where $L(\boldsymbol{\theta}) = 1/n\sum_{i=1}^{n}(y_i e^{-\hat{f}_\theta(x_i)} + \hat{f}_\theta(\mathbf{x}_i))$, s_{ii} is the i^{th} diagonal element of $\mathbf{S}$, $\mathbf{S} = (\mathbf{W} + \lambda\mathbf{I})^{-1}\mathbf{V}$ with $\mathbf{W} = \mathbf{K} \times \mathrm{diag}\{y_i e^{-\hat{f}_\theta(x_i)}\}$ and $\mathbf{V} = \mathbf{K} \times \mathrm{diag}\{e^{-\hat{f}_\theta(x_i)}\}$. Replacing $s_{ii}e^{\hat{f}_\theta(x_i)}$ by their average h_θ, we have the generalized approximate cross validation(GACV) function as follows,

$$\mathrm{GACV}(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) + \frac{1}{n}\left(\frac{h_\theta}{1 - h_\theta}\right)\sum_{i=1}^{n} y_i\left(y_i - e^{\hat{f}_\theta(\mathbf{x}_i)}\right)e^{-2\hat{f}_\theta(\mathbf{x}_i)}. \tag{3.3}$$

4. Numerical Studies

We illustrate the performance of the variance estimation method based on the kernel method through two simulated data sets and one real data set from Wei *et al.* (2006).

Example 1. For the first simulated example, 200 $(x_i, y_i)'$s are generated to present the estimation performance of the proposed method such that $y_i = 1/2e^{f(x_i)}\epsilon_i^2$, $f(x_i) = 2\sin(2\pi x_i)$, $x_i = i/200$, $i = 1, \ldots, 200$, where ϵ_i^2 is generated from a chi-square distribution with 2 degrees of freedom. The Gaussian kernel function is utilized in this example, which is

$$K(x_k, x_l) = e^{-\frac{\|x_k - x_l\|^2}{\sigma^2}}.$$

From GACV function (3.3) (λ, σ^2) is obtained as $(1, 0.15)$. Figure 1(Left) shows true variance function(solid line) and estimated variance functions imposed on the scatter plots of 200 data points of y_i's, where the dashed and dotted lines represent the proposed method and smoothing spline method (Green and Silverman, 1994) with 6 degrees of freedom, respectively. Figure 1(Right) shows values of cross validation functions on various values of kernel parameter σ^2 for fixed $\lambda = 1$. In Figure 1(Right) CV function, ACV function and GACV function are depicted in solid line, dashed line, and dotted line, respectively, from which we can see that ACV function and GACV function are good approximates of CV function.

We repeated the above procedure 100 times to obtain 100 mean squared errors and their standard errors of the proposed method and smoothing spline method with 6 degrees of freedom. The averages of MSE's and standard errors were obtained as (0.2556, 0.0183) and (0.5877, 0.0307), respectively, which implies the proposed method provides a little bit better result than smoothing spline method in this example.

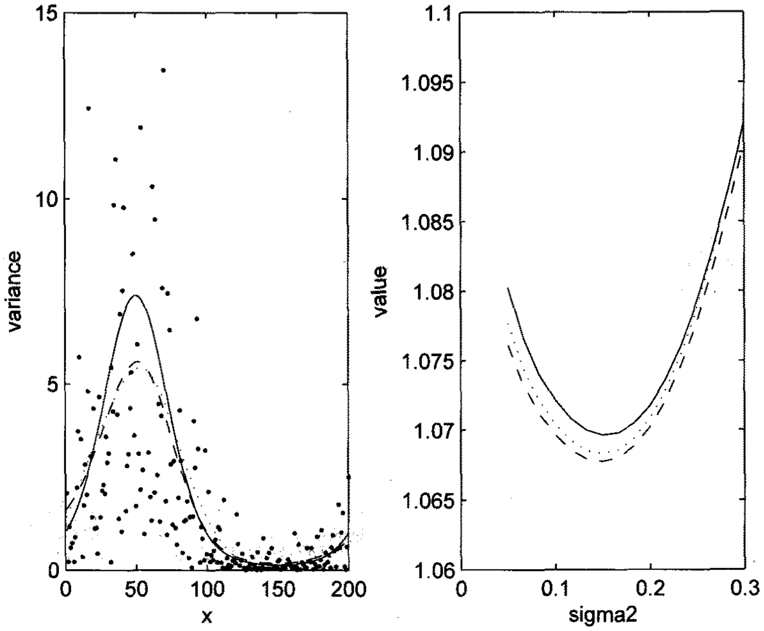

Figure 1: *Variance functions(Left), CV function, ACV function and GACV function(Right) for one of 100 data sets.*

Example 2. For the second simulated example, 100 $(\mathbf{x}_i, z_{ij})$'s with 10 replicates are generated such as $z_{ij} = \epsilon_{ij}$, $i = 1, \ldots, 100$, $j = 1, \ldots, 10$, where ϵ_{ij} is generated from a normal distribution $N(0,\ e^{f(\mathbf{x}_i)})$, $f(\mathbf{x}_i) = \cos(x_{1i} + x_{2i})$, $x_{1i} = (i-1)/99$, $x_{2i} = \pi(i-1)/99$. The Gaussian kernel function is utilized in this example and from GACV function (3.3) $(\lambda,\ \sigma^2)$ is obtained as (1, 0.5). Figure 2(Right) shows true variance function(solid line) and estimated variance functions imposed on the scatter plots of 200 data points of sample variances, where the dashed and dotted lines represent the proposed method and multivariate adaptive regression spline method (Friedman, 1991), respectively. In Figure 2 we can see that the estimated variances function by proposed method behave similarly as the true variance functions do.

We repeated the above procedure 100 times to obtain 100 mean squared errors and their standard errors of the proposed method and multivariate adaptive regression spline method. The averages of MSE's and standard errors were obtained as (0.0246, 0.0018) and (0.0344, 0.0026), respectively, which implies the proposed method provides a little bit better result than multivariate adaptive regression spline method in this example.

Example 3. From California Children Growth Data (Wei *et al.*, 2006), observations with more than 2 replicates on each girl's age are extracted. And from those, observations with 3 randomly chosen replicates on each girl's age are used for the study, $(x_i,\ z_{ij})$ for $i = 1, \ldots, 641$, $j = 1, 2, 3$. We utilize the polynomial kernel function with degree 2, which is empirically found to provides better result than Gaussian kernel function in this example,

$$K(x_k, x_l) = \left(1 + x_k x_l'\right)^2 .$$

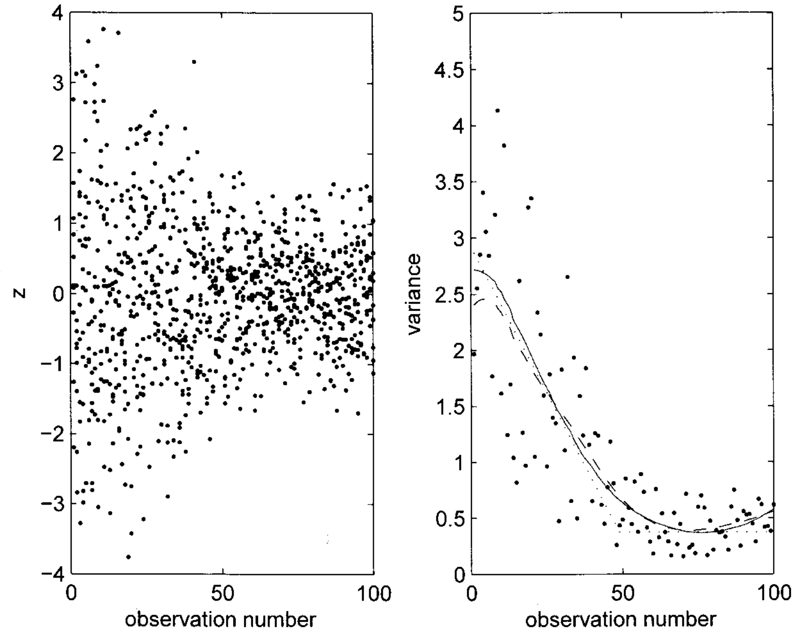

Figure 2: *Replicated data(Left), Sample variances and estimated variance functions(Right) for one of 100 data sets.*

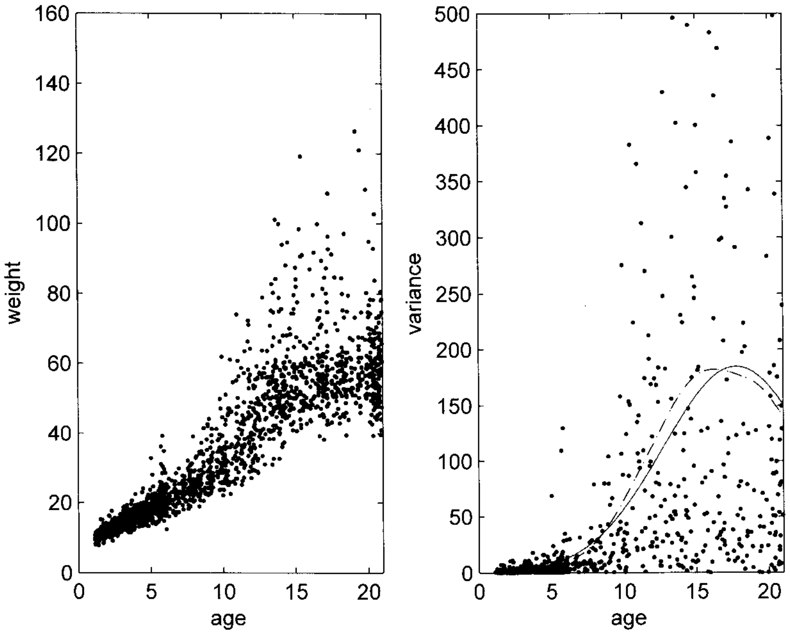

Figure 3: *Replicated data(Left), Sample variances and estimated variance functions(Right).*

From GACV function (3.3) λ is obtained as 4. In Figure 3(Right) estimated variance functions imposed on the scatter plots of data points of sample variances, where estimates by the proposed method are depicted in solid line and estimates by smoothing spline (Green and Silverman, 1994) with 6 degrees of freedom are depicted in dashed line. In Figure 3 we can see that the estimated variance function seems to represent well the behavior of variance of given data.

5. Conclusions

We dealt with estimating the variance function for the data set including observations with same replicates on each covariate or data consisted of sample variances by the kernel trick and obtained GACV function for the proposed method. Through the examples we showed that the proposed method derives the satisfying results. We also found that the proposed procedure has an advantage of an easy model selection method such as GACV function. The variance function estimation for the data set including observations with different replicates on each covariate will be the next study.

References

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

Friedman, J. H. (1991). Multivariate adaptive regression splines, *The Annals of Statistics*, **19**, 1–67.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, Chapman & Hall/CRC, London.

Hall, P. and Carroll, R. J. (1989). Variance function estimation in regression: The effect of estimating the mean, *Journal of the Royal Statistical Society, Series B*, **51**, 3–14.

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 the theory of integral equations, *Philosophical Transactions of Royal Society of London*, Serise A, **209**, 415–446.

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

Wei, Y., Pere, A., Koenker, R. and He, X. (2006). Quantile regression methods for reference growth charts, *Statistics in Medicine*, **25**, 1369–1382.

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

Received November 2008; Accepted December 2008