Journal of the Korean Data & Information Science Society
2009, **20**(5), 925–931

한국데이터정보과학회지

Variance function estimation with LS-SVM for replicated data

Jooyong Shim[1] · Hyejung Park[2] · Kyung Ha Seok[3]

[1]Department of Applied Statistics, Catholic University of Daegu
[2]Division of General Education and Teacher's Certification, Daegu University
[3]Department of Data Science, Institute of Statistical Information, Inje University

Received 12 July 2009, revised 6 September 2009, accepted 12 September 2009

Abstract

In this paper we propose a variance function estimation method for replicated data based on averages of squared residuals obtained from estimated mean function by the least squares support vector machine. Newton-Raphson method is used to obtain associated parameter vector for the variance function estimation. Furthermore, the cross validation functions are introduced to select the hyper-parameters which affect the performance of the proposed estimation method. Experimental results are then presented which illustrate the performance of the proposed procedure.

Keywords: Cross validation function, heteroscedasticity, hyper-parameters, kernel function, least squares support vector machine.

1. Introduction

It becomes an important issue in many fields to measure volatility or local variability, which are usually modelled in terms of variance functions. Researches on estimation of variance function can be found in Anderson and Lund (1997), 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, 1998; Suykens and Vanderwalle, 1999) for the heteroscedastic regression problem based on the average of squared residuals which are assumed to follow a 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 (1909). The kernel trick has been applied to the regression problems of various data types (Shim and Seok, 2008; Shim *et al.,* 2009). The estimated variance functions are

[1] Adjunct Professor, Department of Applied Statistics, Catholic University of Daegu, Gyungbuk 712-702, Korea.
[2] Visiting Professor, Computer Course Division of General Education and Teacher's Certification, Daegu University, Gyungbuk 712-714, Korea.
[3] Corresponding author: Professor, Department of Data Science, Institute of Statistical Information, Inje University, Kimhae 621-749, Korea. E-mail: statskh@inje.ac.kr

obtained by minimizing the penalized log-likelihood function of averages of squared residuals. Here averages of squared residuals are obtained from estimated mean function by applying LS-SVM (least squares support vector machine), (Suykens and Vanderwalle, 1999) with replicated data where errors assumed to follow a normal distribution. The basic idea and applications of LS-SVM can be found in Suykens *et al.* (2000, 2002). The proposed method enables to select appropriate hyper-parameters easily from the generalized cross validation (GCV) function and the generalized approximate cross validation (GACV) function for mean and variance function estimation, respectively, which are used to select hyper-parameters for the achievement of high generalization performance. The rest of this paper is organized as follows. In Section 2 we propose the mean and variance function estimations method using the principal idea of kernel machine. In Section 3 we present the model selection method using GCV function and GACV function. In Section 4 we perform the numerical studies through examples. In Section 5 we give the conclusions.

2. Mean and variance functions estimation

Consider the heteroscedastic regression model with n observations with m_i replicates as follows,

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

where ϵ_{ij} follows a normal distribution $(0, e^{f(\boldsymbol{x}_i)})$ with an unknown nonlinear function $f(\boldsymbol{x}_i)$, the $\boldsymbol{x}_i \in \boldsymbol{R}^d$ is an input vector including a constant 1 and $\mu_i = \mu(\boldsymbol{x}_i)$ is a mean function of z_{ij}'s for $i = 1, \cdots, n$ such that,

$$\mu(\boldsymbol{x}_i) = E(z_{ij}|\boldsymbol{x}_i).$$

Here $\frac{1}{m_i}\sum_{j=1}^{m_i} \epsilon_{ij}^2$ follows independently gamma distribution $(\frac{m_i}{2}, \frac{2}{m_i}e^{f(\boldsymbol{x}_i)})$ for $i = 1, \cdots, n$, which provides the justification that we can assume that the average of squared residuals, $y_i = \frac{1}{m_i}\sum_{j=1}^{m_i}(z_{ij} - \widehat{\mu}(\boldsymbol{x}_i))^2$ also follows independently gamma distribution $(\frac{m_i}{2}, \frac{2}{m_i}e^{f(\boldsymbol{x}_i)})$ for $i = 1, \cdots, n$. Here we define the variance function as

$$Var(z_{ij}|\boldsymbol{x}_i) = E(y_i|\boldsymbol{x}_i) = e^{f(\boldsymbol{x}_i)}.$$

To obtain estimator of mean function, $\widehat{\mu}(\boldsymbol{x}_i)$, we apply LS-SVM as follows,

$$min \frac{1}{2}\boldsymbol{w}_\mu'\boldsymbol{w}_\mu + \frac{C}{2}\sum_{i,j} e_{ij}^2$$

subject to $z_{ij} - \boldsymbol{w}_\mu'\phi_\mu(\boldsymbol{x}_i) = e_{ij}$, $j = 1, \cdots, m_i, i = 1, \cdots, n$, where C is a nonnegative regularization parameter which controls the trade-off between the goodness-of-fit on the data and $\| \boldsymbol{w}_\mu \|^2$, $\boldsymbol{w}_\mu$ is a weight vector which is used for $\widehat{\mu}(\boldsymbol{x}_i) = \boldsymbol{w}_\mu'\phi_\mu(\boldsymbol{x}_i)$. $\phi_\mu(\cdot)$ is the feature mapping function such that $\phi_\mu(\cdot) : R^d -> R^{d_f}$ maps the input space to the

higher dimensional feature space where the dimension d_f is defined in an implicit way. It is 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). Then $\widehat{\mu}(\boldsymbol{x}_i)$ can be expressed with the optimal values of Lagrange multipliers, β_{ij}'s, which are the solution of the linear system,

$$\widehat{\boldsymbol{\beta}} = (UK_\mu U' + I/C)^{-1}\boldsymbol{z},$$

where $K_\mu = K_\mu(\boldsymbol{x}, \boldsymbol{x})$, $\boldsymbol{x}$ is a $n \times d$ matrix of $\boldsymbol{x}_i$'s, U is a $N \times n$ block diagonal matrix consisted of $\mathbf{1}_{m_i \times 1}$ with $N = \sum_{i=1}^{n} m_i$, $\boldsymbol{z} = (z_{11}, z_{12}, \cdots, z_{n,m_n})'$ is a $N \times 1$ vector of z_{ij}'s. $\widehat{\mu}(\boldsymbol{x}_i)$ can be expressed with $\widehat{\boldsymbol{\beta}}$ as follows,

$$\widehat{\mu}(\boldsymbol{x_i}) = K_\mu(\boldsymbol{x_i}, \boldsymbol{x})U'\widehat{\boldsymbol{\beta}}.$$

For $y_i = \dfrac{1}{m_i}\sum_{j=1}^{m_i}(z_{ij} - \widehat{\mu}(\boldsymbol{x}_i))^2$, the negative log-likelihood of the given data can be expressed as (a constant term is omitted) under the assumption that y_i follows independently gamma distribution $(\dfrac{m_i}{2}, \dfrac{2}{m_i}e^{f(\boldsymbol{x}_i)})$ for $i = 1, \cdots, n$,

$$L(f) = \frac{1}{n}\sum_{i=1}^{n}(m_i y_i e^{-f(\boldsymbol{x}_i)} + m_i f(\boldsymbol{x}_i)).$$

The nonlinear function $f(\boldsymbol{x}_i)$ can be estimated by a linear model, $f(\boldsymbol{x}_i) = \boldsymbol{\omega}'\phi(\boldsymbol{x}_i)$, conducted in a high dimensional feature space. Then the estimate of parameter vector satisfying $f(\boldsymbol{x}_i) = \boldsymbol{\omega}'\phi(\boldsymbol{x}_i)$ for $i = 1, \cdots, n$ is obtained by minimizing the penalized negative log-likelihood,

$$L(\boldsymbol{\omega}) = \sum_{i=1}^{n}(m_i y_i e^{-\boldsymbol{\omega}'\phi(\boldsymbol{x}_i)} + m_i\boldsymbol{\omega}'\phi(\boldsymbol{x}_i)) + \frac{\lambda}{2} \parallel \boldsymbol{\omega} \parallel^2, \tag{2.1}$$

where λ is a nonnegative regularization parameter which controls the trade-off between the goodness-of-fit on the data and $\parallel \boldsymbol{\omega} \parallel^2$ and $\phi_\mu(\cdot)$ is the feature mapping function. The representer theorem (Kimeldorf and Wahba, 1971) guarantees the minimizer of the penalized negative log-likelihood to be $f(\boldsymbol{x}_i) = K_i\boldsymbol{\alpha}$ for some $n \times 1$ vector $\boldsymbol{\alpha}$, where K_i is the i -th row of the $n \times n$ kernel matrix K. Now the penalized negative log-likelihood (2.1) becomes

$$L(\boldsymbol{\alpha}) = (\boldsymbol{m}'\boldsymbol{Y}e^{-K\boldsymbol{\alpha}} + \boldsymbol{m}'K\boldsymbol{\alpha}) + \frac{\lambda}{2}\boldsymbol{\alpha}'K\boldsymbol{\alpha}, \tag{2.2}$$

where $\boldsymbol{m} = (m_1, \cdots, m_n)'$, $\boldsymbol{Y}$ is a diagonal matrix of $\boldsymbol{y}$, and e is the componentwise exponential function. By minimizing the penalized negative log-likelihood (2.2) we obtain the estimate of parameter vector $\boldsymbol{\alpha}$, but not in a explicit form, which leads to use the Newton-Raphson method. At each iteration the parameter vector $\boldsymbol{\alpha}$ is updated as follows,

$$\widehat{\boldsymbol{\alpha}}^{new} = \widehat{\boldsymbol{\alpha}} - \boldsymbol{H}^{-1}\boldsymbol{G},$$

where $\boldsymbol{G}$ is the gradient vector and $\boldsymbol{H}$ is the Hessian matrix of (2.2).

With the estimate of parameter vector $\widehat{\boldsymbol{\alpha}}$, the estimated variance function for the input vector $\boldsymbol{x}_i$ is obtained as,

$$\widehat{Var}(z_{ij}|\boldsymbol{x}_i) = e^{\widehat{f(\boldsymbol{x}_i)}} = e^{K_i\widehat{\boldsymbol{\alpha}}},$$

where K_i is the $1 \times n$ row vector with elements $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$, $j = 1, \cdots, n$.

3. Model selection

The functional structures of the estimation method of mean function and variance function are characterized by hyper-parameters, the regularization parameter C or λ and the kernel parameter.

Hyper-parameters θ_μ in the mean function estimation can be chosen by minimizing the cross validation (CV) function as follows:

$$CV(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{m_i}(z_{ij} - \widehat{\mu}_{\theta_\mu}^{(-ij)}(\boldsymbol{x}_i)),$$

where $\widehat{\mu}_{\theta_\mu}^{(-ij)}(\boldsymbol{x}_i)$ is the estimate of $\mu(\boldsymbol{x}_i)$ estimated data without (i,j) th observation. Since for each candidates of hyperparameters, $\widehat{\mu}_{\theta_\mu}^{(-ij)}(\boldsymbol{x}_i)$ for $i = 1, \cdots, n$, should be evaluated, selecting parameters using CV function is computationally formidable. The GCV function (Craven and Wahba, 1979) is given as follows.

$$GCV(\theta_\mu) = \frac{N||\boldsymbol{z} - U\widehat{\mu}_{\theta_\mu}(\boldsymbol{x})||^2}{N - tr(US_{\theta_\mu}))^2}, \tag{3.1}$$

where $S_{\theta_\mu} = K_\mu U'(UK_\mu U' + \mathrm{I}/C)^{-1}$ such $\widehat{\mu}_{\theta_\mu}(\boldsymbol{x}) = S_{\theta_\mu}\boldsymbol{z}$.
For the model selection of the variance function estimation, we consider the CV function as follows,

$$CV(\theta) = \frac{1}{n}\sum_{i=1}^{n}\{m_i y_i e^{-\widehat{f}_\theta^{(-i)}(\boldsymbol{x}_i)} + m_i\widehat{f}_\theta(\boldsymbol{x}_i)\},$$

where θ is the hyper-parameters in the variance function estimation and $\widehat{f}_\theta^{(-i)}(\boldsymbol{x}_i)$ is the estimate of $f(\boldsymbol{x}_i)$ estimated data without i th observation.
From Xiang and Wahba (1996) and Liu *et al.* (2007) we have the approximate cross validation (ACV) function as follows,

$$ACV(\theta) = L(\theta) + \frac{1}{n}\sum_{i=1}^{n}\frac{s_{ii}m_i e^{-\widehat{f}_\theta(\boldsymbol{x}_i)}y_i(y_i - e^{\widehat{f}_\theta(\boldsymbol{x}_i)})}{1 - s_{ii}e^{\widehat{f}_\theta(\boldsymbol{x}_i)}}.$$

where $L(\theta) = \frac{1}{n}\sum_{i=1}^{n}(m_i y_i e^{-\widehat{f}_\theta(\boldsymbol{x}_i)} + m_i\widehat{f}_\theta(\boldsymbol{x}_i))$, s_{ii} is the i th diagonal element of $S = (W + \lambda K^{-1})^{-1}$ with $W = diag\{m_i y_i e^{-\widehat{f}_\theta(\boldsymbol{x}_i)}\}$. Replacing $s_{ii}e^{\widehat{f}_\theta(\boldsymbol{x}_i)}$ by their average h_θ, the GACV function is obtained as follows,

$$GACV(\theta) = L(\theta) + \frac{1}{n}(\frac{h_\theta}{1 - h_\theta})\sum_{i=1}^{n} m_i y_i(y_i - e^{\widehat{f}_\theta(\boldsymbol{x}_i)})e^{-2\widehat{f}_\theta(\boldsymbol{x}_i)}. \tag{3.2}$$

4. Numerical studies

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

Example 1. For the simulated example, we consider the following model,

$$z_{ij} = \mu(x_i) + e_{ij}, \quad i = 1, \cdots, 100, \quad j = 1, \cdots, 10,$$

where $x_i = i/100$, $\mu(x_i) = \cos(2\pi x_i)$, e_{ij} follows a normal distribution $N(0, e^{f(x_i)})$ with $f(x_i) = 2sin(2\pi x_i)$. The Gaussian kernel function is utilized for both mean function and variance function estimation in this example.

For the mean function estimation (C, σ_μ^2) is chosen as (10, 0.5) from GCV function (3.1) and (λ, σ^2) for the variance function estimation is chosen as (1, 0.05) from GACV function (3.2).

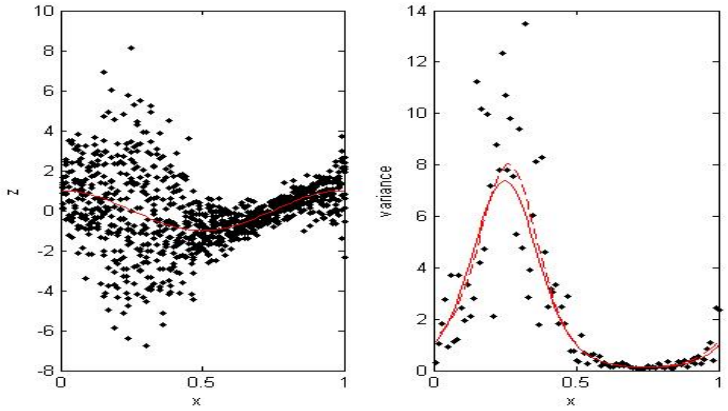

Figure 4.1 Replicated data (Left) and estimated variance functions (Right)

Figure 4.1 (Left) shows true mean functions (solid line) and estimated mean functions (dashed line) imposed on the scatter plots of 1000 data points of z_{ij}'s. True mean functions and estimated mean functions are too close to distinguish them from the figure (mse of their differences =0.0076). Figure 4.1 (Right) shows true variance functions (solid line) and estimated variance functions imposed on the scatter plots of 100 averages of squared residuals. In the figure we can see that the estimated mean and variance functions seem to represent well the behavior of mean and variance of given data.

We repeated the above procedure 100 times (generate 100 data sets) to get the root mean squared errors (RMSE) for the true mean functions and variance functions. We obtained the average of 100 RMSEs and their standard error for the mean function as (1.5019, 0.0049), the average of 100 RMSEs and their standard error for the variance function as (0.3855, 0.0161). The small values of RMSEs indicate that the proposed method works well.

Example 2. California Children Growth Data (Wei *et al.*, 2006) consist of girl's age and their weights, (x_i, z_{ij}) for $i = 1, \cdots, 1657$, and the number of whole data $N = \sum_{i=1}^{1657} m_i = 4011$. The Gaussian kernel functions are utilized for both mean function and variance function estimation in this example.

For the mean function estimation (C, σ_μ^2) is chosen as (1, 30) from GCV function (3.1) and (λ, σ^2) for the variance function estimation is chosen as (1, 50) from GACV function (3.2).

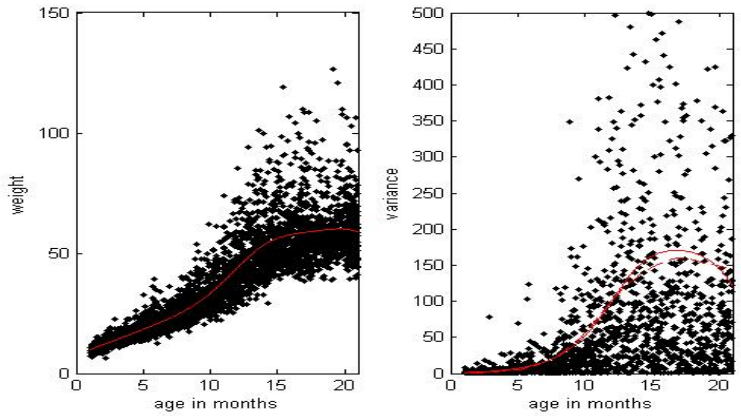

Figure 4.2 Replicated data (Left) and sample variances and estimated variance functions (Right).

Figure 4.2 (Left) shows estimated mean functions imposed on the scatter plots of 1923 data points. In Figure 4.2 (Right) estimated variance functions imposed on the scatter plots of data points of sample variances, where solid line=proposed, dashed line=cubic smoothing spline (Green and Silverman, 1994). In this figure we can see that the estimated variance function seems to represent well the behavior of variances of given data.

5. Conclusions

In this paper, we dealt with estimating the mean and variance functions for replicated data by the kernel trick and obtained cross validation functions for the proposed method. The proposed method has an advantage that it provide an accurate and simple estimation of the variance function and that it can be applied even for the non-replicated data. Through the examples we showed that the proposed method derives the satisfying results. We also found that the proposed method has an advantage of having an easy model selection methods such as GCV function and GACV function.

References

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

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerical Mathematics,* **31**, 377-403.

Green, P. J. and Silverman, B. W. (1995). *Nonparametric regression and generalized linear models,* Chapman & Hall, London.

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.

Shim, J. and Seok, K. H. (2008). Kernel ridge regression with randomly right censored data. *Journal of Korean Data & Information Science Society,* **19**, 1353 -1360.

Shim, J., Kim, T. Y., Lee, S. Y. and Hwang, C. (2009). Credibility estimation via Kernel mixed effects model. *Journal of Korean Data & Information Science Society,* **20**, 445 -452.

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

Suykens, J. A. K., De Brabanter, J., Lukas, L. and Vandewalle, J. (2002). Weighted least squares support vector machines: Robustness and sparse approximation. *Neurocomputing, Special issue on fundamental and information processing aspects of neurocomputing*, **48**, 85-105.

Suyken, J. A. K., Lukas, L. and Vandewalle, J. (2000). Sparse approximation using least squares support vector machines. *IEEE International Symposium on Circuits and Systems* (ISCAS 2000), 757-760, Geneva, Switzerland.

Vapnik, V. N. (1998). *Statistical learning theory,* John Wiley & Sons.

Wei, Y., Pere, A., Koenker, R. and He, X. (2006). Quantile regression 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.