Journal of the Korean Data & Information Science Society
2013, **24**(2), 391–399

http://dx.doi.org/10.7465/jkdi.2013.24.2.391
한국데이터정보과학회지

GACV for partially linear support vector regression[†]

Jooyong Shim[1] · Kyungha Seok[2]

[12]Department of Data Science, Inje University

Received 19 February 2013, revised 5 March 2013, accepted 10 March 2013

Abstract

Partially linear regression is capable of providing more complete description of the linear and nonlinear relationships among random variables. In support vector regression (SVR) the hyper-parameters are known to affect the performance of regression. In this paper we propose an iterative reweighted least squares (IRWLS) procedure to solve the quadratic problem of partially linear support vector regression with a modified loss function, which enables us to use the generalized approximate cross validation function to select the hyper-parameters. Experimental results are then presented which illustrate the performance of the partially linear SVR using IRWLS procedure.

Keywords: Generalized approximate cross validation function, iterative reweighted least squares procedure, partially linear regression, support vector regression.

1. Introduction

Support vector machine (SVM), firstly developed by Vapnik (1995, 1998), is being used as a new technique for regression and classification problems. SVM is based on the structural risk minimization (SRM) principle, which has been shown to be superior to traditional empirical risk minimization (ERM) principle. SRM minimizes an upper bound on the expected risk unlike ERM minimizing the error on the training data. By minimizing this bound, high generalization performance can be achieved. In particular, for the SVM regression case SRM results in the regularized ERM with the e-insensitive loss function. The introductions and overviews of recent developments of SVM regression can be found in Cho *et al.* (2010), Hwang (2010), Shim *et al.* (2011), Smola and Schölkopf (1998), Vapnik (1995, 1998), and Wang (2005).

Training an SVR requires the solution to a quadratic programming (QP) optimization problem. But QP problem presents some inherent limitations which results in computational difficulty especially for the large data sets. Platt (1998) developed the sequential minimal optimization algorithm which divides the QP problem into a series of small QP problems to avoid such computational difficulty. Perez-Cruz *et al.* (2000) proposed IRWLS algorithm

[†] This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012000646) and (2011-0009705).

[1] Adjunct professor, Institute of statistical Information, Department of Data Science, Inje University, Kimhae 621-749, Korea.

[2] Corresponding author: Professor, Institute of statistical Information, Department of Data Science, Inje University, Kimhae 621-749, Korea. E-mail: statskh@inje.ac.kr

for SVR by transforming the Lagrangian function into sum of quadratic terms by defining associated weights of predicted errors.

In this paper, we consider the partially linear regression case where the input vector included in the linear part of the regression function is assumed to be known to have the linear effect on the response variable and the input vector included in the nonlinear part of the regression function is assumed to be known to have the nonlinear effect on the response variable. We propose an IRWLS procedure to solve the QP problem of partially linear SVR (PLSVR) with a modified loss function of which original version is e-insensitive loss function used by Vapnik (1995, 1998). The modified loss function is attained by providing the differentiability at $\pm\, e$, which enables to solve QP problem by IRWLS procedure. To select appropriate hyper-parameters, a commonly used method is minimizing the cross validation (CV) function. Nychka *et al.* (1995) proposed the approximate cross validation (ACV) function for quantile spline estimation. This technique can be easily applied to PLSVR using IRWLS procedure. And by replacing each element of hat matrix by the average of trace of hat matix, the GACV function also can be obtained. GACV function is 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 give a review of PLSVR. In Section 3 we propose an IRWLS procedure for PLSVR and present the model selection method using GACV function which is a good approximate of the generalized comparative Kullback-Leibler distance (Wahba *et al.*, 1999). In Section 4 we perform the numerical studies through examples. In Section 5 we give the conclusions.

2. Partially linear SVR

Let the training data set denoted by $\{\boldsymbol{x}_i, \boldsymbol{z}_i, y_i\}_{i=1}^n$, with each input vector $\boldsymbol{x}_i \in \boldsymbol{R}^{d_1}$, $\boldsymbol{z}_i \in \boldsymbol{R}^{d_2}$ and the response $y_i \in \boldsymbol{R}$, where the output variable is assumed to be linearly related to the input vector $\boldsymbol{x}_i$ and nonlinearly related to the input vector $\boldsymbol{z}_i$. Here the feature mapping function $\phi(\cdot) : \boldsymbol{R}^{d_2} \to \boldsymbol{R}^{d_f}$ 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 $\phi(\boldsymbol{z}_i)'\phi(\boldsymbol{z}_j) = K(\boldsymbol{z}_i, \boldsymbol{z}_j)$ which are obtained from the application of Mercer's (1909) conditions. We consider the partially linear regression case, in which the regression function of the response given $\boldsymbol{x}$ and $\boldsymbol{z}$, $\mu(\boldsymbol{x}, \boldsymbol{z})$, can be regarded as a partially linear function of input vector $\boldsymbol{x}$ and $\boldsymbol{z}$ such that $\mu(\boldsymbol{x}, \boldsymbol{z}) = \boldsymbol{w}_1'\boldsymbol{x} + \boldsymbol{w}_2'\phi(\boldsymbol{z}) + b = \boldsymbol{w}'\phi(\boldsymbol{x}, \boldsymbol{z}) + b$, where $\boldsymbol{w} = \begin{pmatrix} \boldsymbol{w}_1 \\ \boldsymbol{w}_2 \end{pmatrix}$ and $\phi(\boldsymbol{x}, \boldsymbol{z}) = \begin{pmatrix} \boldsymbol{x} \\ \phi(\boldsymbol{z}) \end{pmatrix}$. In the nonlinear case, $\boldsymbol{w}_2$ is no longer explicitly given. However, it is uniquely defined in the weak sense by the dot products. Here the linear regression model can be regarded as the special case of the nonlinear regression model by using identity feature mapping function, that is, $\phi(\boldsymbol{z}) = \boldsymbol{z}$ which implies the linear kernel such that $K(\boldsymbol{z}, \boldsymbol{z}) = \boldsymbol{z}'\boldsymbol{z}$.

With e-insensitive loss function ℓ_e, the estimator of the regression function can be defined as any solution to the optimization problem,

$$\min \frac{1}{2}\boldsymbol{w}'\boldsymbol{w} + C\sum_{i=1}^{n} \ell_e(y_i - \mu(\boldsymbol{x}_i, \boldsymbol{z}_i)), \tag{2.1}$$

where $C > 0$ is a penalty parameter penalizing the training errors, $\ell_e(r) = 0$ if $|r| \le e$ and $\ell_e(r) = |r| - e$ if $|r| > e$. We can express the regression problem by formulation for SVM as

follows:

$$\min \frac{1}{2}\boldsymbol{w}'\boldsymbol{w} + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{2.2}$$

subject to

$$\begin{aligned} &y_i - \boldsymbol{w}'\phi(\boldsymbol{x}_i, \boldsymbol{z}_i) - b \le e + \xi_i \\ &\boldsymbol{w}'\phi(\boldsymbol{x}_i, \boldsymbol{z}_i) + b - y_i \le e + \xi_i^*,\ e, \xi_i, \xi_i^* \ge 0. \end{aligned} \tag{2.3}$$

We construct a Lagrange function as follows:

$$\begin{aligned} L =& \frac{1}{2}\boldsymbol{w}'\boldsymbol{w} + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) - \sum_{i=1}^{n}\alpha_i(e + \xi_i - y_i + \boldsymbol{w}'\phi(\boldsymbol{x}_i, \boldsymbol{z}_i) + b) \\ &- \sum_{i=1}^{n}\alpha_i^*(e + \xi_i^* + y_i - \boldsymbol{w}'\phi(\boldsymbol{x}_i, \boldsymbol{z}_i) - b) - \sum_{i=1}^{n}(\eta_i\xi_i + \eta_i^*\xi_i^*). \end{aligned} \tag{2.4}$$

We notice that the positivity constraints $\alpha_i, \alpha_i^*, \eta_i, \eta_i^* \ge 0$ should be satisfied. Taking partial derivatives of (2.4) with regard to the primal variables $(\boldsymbol{w}, b, \xi_i, \xi_i^*)$ we have,

$$\begin{aligned} \frac{\partial L}{\partial \boldsymbol{w}} =& \boldsymbol{0} \to \boldsymbol{w} = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\phi(\boldsymbol{x}_i, \boldsymbol{z}_i), \\ \frac{\partial L}{\partial b} =& 0 \to \sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0, \\ \frac{\partial L}{\partial \xi_i^{(*)}} =& 0 \to C - \alpha_i^{(*)} - \eta_i^{(*)} = 0. \end{aligned} \tag{2.5}$$

Plugging (2.5) into (2.4), we have the optimization problem as follows;

$$\max\ -\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)(\boldsymbol{x}_i'\boldsymbol{x}_j + K(\boldsymbol{x}_i, \boldsymbol{x}_j)) + \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)y_i - e\sum_{i}^{n}(\alpha_i + \alpha_i^*) \tag{2.6}$$

with constraints $0 \le \alpha_i^{(*)} \le C$, where the data points corresponding to positive values of α_i or α_i^* are called support vectors. Solving the above equation with the constraints determines the optimal Lagrange multipliers, α_i, α_i^*, the estimator of the regression function given the input vector $\boldsymbol{x}_t$ and $\boldsymbol{z}_t$ is obtained as follows;

$$\widehat{\mu}(\boldsymbol{x}_t, \boldsymbol{z}_t) = \sum_{i=1}^{n}(\widehat{\alpha}_i - \widehat{\alpha}_i{}^*)(\boldsymbol{x}_i'\boldsymbol{x}_t + K(\boldsymbol{z}_i, \boldsymbol{z}_t)) + \widehat{b}. \tag{2.7}$$

Here $\widehat{b}$ is obtained by KKT (Karush-Kuhn-Tucker, Kuhn and Tucker, 1951) conditions as follows;

$$\widehat{b} = \frac{1}{n_s}\sum_{i \in I_s}(y_i - (\boldsymbol{x}_i'\boldsymbol{x}' + K(\boldsymbol{x}_i, \boldsymbol{x}))(\widehat{\boldsymbol{\alpha}} - \widehat{\boldsymbol{\alpha}}^*)), \tag{2.8}$$

where n_s is the size of $I_s = \{i = 1, \cdots, n | 0 < \widehat{\alpha}_i < C, 0 < \widehat{\alpha}_i{}^* < C\}$.

3. Partially linear SVR using IRWLS procedure

Hyper-parameters are the penalty parameter and kernel parameter included in the kernel. The penalty parameter play an important role on determining the tradeoff between the goodness-of-fit on the data and $||\boldsymbol{\omega}||^2$. When it is too small, there is too much penalty placed on the estimate, which leads underfitting. Or when it is too large, we tend to interpolate the data more and this will lead to overfitting. The kernel parameter is also known to be related underfitting and overfitting. The main goal of model selection is to choose hyper-parameters such that the distance between the resulting estimate and the true regression function is minimized (the generalized comparative Kullback-Leibler distance, Wahba *et al.*, 1999). Since the true regression function and the error distribution are not known, the distance cannot be directly obtained. One popular proxy is the leave-one-out cross-validation defined as follows;

$$CV(\boldsymbol{\lambda}) = \frac{1}{n}\sum_{i=1}^{n} \ell_e(y_i - \widehat{\mu}^{(-i)}(\boldsymbol{x}_i, \boldsymbol{z}_i)) \tag{3.1}$$

where $\boldsymbol{\lambda}$ is the set of parameters and $\widehat{\mu}^{(-i)}(\boldsymbol{x}_i, \boldsymbol{z}_i)$ is the regression function estimated without i th observation. Since for each candidates of hyper-parameters, $\widehat{\mu}^{(-i)}(\boldsymbol{x}_i, \boldsymbol{z}_i)$ for $i = 1, \cdots, n$, should be evaluated, selecting hyper-parameters using CV function is computationally formidable. If $\widehat{\mu}(\boldsymbol{x}_i, \boldsymbol{z}_i)$ can be expressed as the linear product of the hat matrix and $\boldsymbol{y}$, the generalized approximate cross validation (GACV) function can be written as follows by Yuan (2006);

$$GACV(\boldsymbol{\lambda}) = \frac{1}{n - tr(H)}\sum_{i=1}^{n} \ell_e(y_i - \widehat{\mu}(\boldsymbol{x}_i, \boldsymbol{z}_i)), \tag{3.2}$$

where H is the hat matrix such that $\widehat{\mu}(\boldsymbol{x}, \boldsymbol{z}) = H\boldsymbol{y}$.

In fact we cannot use GACV function (3.2) with the estimator of the regression function (2.8). One reason is that the e-insensitive loss function $\ell_e(r)$ is not differentiable with respect to r at $\pm\, e$. We use the approximate of e-insensitive loss function, $\ell_{e,\delta}(r)$, which is attained by providing the differentiability at $\pm\, e$ by differing from the original e-insensitive loss function in the small intervals $(-e-\delta, -e+\delta)$ and $(e-\delta, e+\delta)$ as follows:

$$\ell_{e,\delta}(r) = \begin{cases} -r + e & \text{if } \ r \le -e - \delta, \\ \frac{r^2 - (e-\delta)^2}{4e} & \text{if } \ -e - \delta < r \le -e + \delta, \\ 0 & \text{if } \ -e + \delta < r \le e - \delta, \\ \frac{r^2 - (e-\delta)^2}{4e} & \text{if } \ e - \delta < r \le e + \delta, \\ r - e & \text{if } r > e + \delta. \end{cases} \tag{3.3}$$

Now the problem (2.1) becomes obtaining $(\boldsymbol{\beta}, b)$ to minimize

$$L(\boldsymbol{\beta}, b) = \frac{1}{2}\boldsymbol{\beta}'\widetilde{K}\boldsymbol{\beta} + C\sum_{i=1}^{n} \ell_{e,\delta}(y_i - \widetilde{K}_i\boldsymbol{\beta} - b), \tag{3.4}$$

where $\widetilde{K} = \boldsymbol{x}\boldsymbol{x}' + K(\boldsymbol{z}, \boldsymbol{z})$ and $\widetilde{K}_i$ is the i th row of $\widetilde{K}$.

Taking partial derivatives of (3.4) with regard to $(\boldsymbol{\beta}, b)$ leads to the optimal values of $(\boldsymbol{\beta}, b)$ to be the solution to

$$\begin{aligned} \mathbf{0} =& \widetilde{K}\boldsymbol{\beta} - C\widetilde{K}W(\boldsymbol{y} - \widetilde{K}\boldsymbol{\beta} - \mathbf{1}b) \\ 0 =& \mathbf{1}'W(\boldsymbol{y} - \widetilde{K}\boldsymbol{\beta} - \mathbf{1}b). \end{aligned} \tag{3.5}$$

Here W is a diagonal matrix composed of w_{ii} 's obtained from the derivative of the modified asymmetric e-insensitive loss function as follows;

$$w_{ii} = \begin{cases} -\frac{1}{r_i} & \text{if } \ r_i \le -e - \delta, \\ \frac{1}{2e} & \text{if } \ -e - \delta < r_i \le -e + \delta, \\ 0 & \text{if} -e + \delta < r_i \le e - \delta, \\ \frac{1}{2e} & \text{if } \ e - \delta < r_i \le e + \delta, \\ \frac{1}{r_i} & \text{if } \ r_i > e + \delta. \end{cases} \tag{3.6}$$

where $r_i = y_i - \widetilde{K_i}\boldsymbol{\beta} - b$. The e-insensitive loss function, the modified e-insensitive loss function and the derivative of the modified e-insensitive loss function are illustrated in Figure 3.1 with $e = 0.2$ and $\delta = 0.05$. The solution to (3.5) can be obtained with W which is composed of the values of $(\boldsymbol{\beta}, b)$ obtained in previous steps. Thus, $\widehat{\mu}(\boldsymbol{x}_i, \boldsymbol{z}_i) = \widetilde{K_i}\boldsymbol{\beta} + b, i = 1, \cdots, n$, can be estimated using IRWLS procedure as follows;

(0) Set $(\boldsymbol{\beta}^{(0)}, b^{(0)})$.

(i) Calculate W using prespecfied e, δ and $r_i = y_i - \widetilde{K_i}\boldsymbol{\beta}^{(t)} - b^{(t)}$.

(ii) Obtain $(\boldsymbol{\beta}^{(t+1)}, b^{(t+1)})$from $\begin{pmatrix} \boldsymbol{\beta}^{(t+1)} \\ b^{(t+1)} \end{pmatrix} = \begin{pmatrix} W\widetilde{K} + \mathrm{I}/C & W\mathbf{1} \\ \mathbf{1}'W\widetilde{K} & \mathbf{1}'W\mathbf{1} \end{pmatrix}^{-1} \begin{pmatrix} W \\ \mathbf{1}'W \end{pmatrix} \boldsymbol{y}.$ (3.7)

(iii) Iterate steps (i) and (ii) until convergence.

The hat matrix $\boldsymbol{H}$ which can be used in GACV function (2.8) is obtained as follows;

$$\boldsymbol{H} = (\widetilde{K}, \mathbf{1}) \begin{pmatrix} W\widetilde{K} + \mathrm{I}/C & W \\ \mathbf{1}'W\widetilde{K} & \mathbf{1}'W\mathbf{1} \end{pmatrix}^{-1} \begin{pmatrix} W \\ \mathbf{1}'W \end{pmatrix}, \tag{3.8}$$

where $\boldsymbol{W}$ is composed of the final estimated values of $(\boldsymbol{\beta}, b)$.

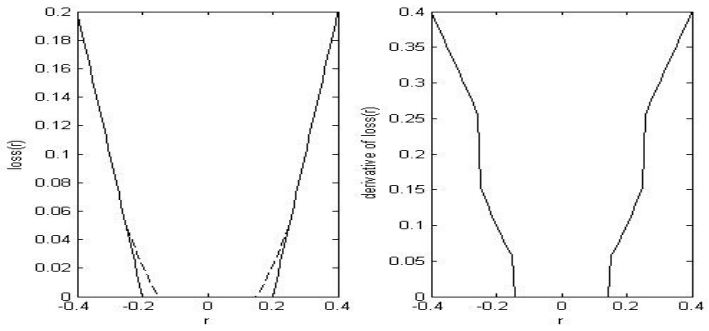

Figure 3.1 0.2-insensitive loss function (solid line), the modified 0.2-insensitive loss function (dashed line) (Left) and the derivative of the modified 0.2-insensitive loss function (Right) with δ =0.05

4. Numerical studies

We illustrate the performance of the regression estimation using PLSVR using IRWLS through the simulated example on the partially linear regression cases. 100 data sets are generated, where each data set consists of 200 (x, z)'s and 200 y's. Here x's are equally spaced ranging from 0 to 1, z's are generated from a uniform $U(0, 1)$ distribution and y's are generated from Laplace distribution $L(x + \sin(2\pi z), 1)$ and the normal distribution $N(x + \sin(2\pi z), 2)$. The regression function a given (x, z) can be modelled as $\mu(x, z) = w_1 x + w_2 \phi(z) + b$. True regression function is given as $f(x, z) = x + \sin(2\pi z)$. One of 100 datasets with Laplacian errors is shown in Figure 4.1, where true regression functions are superimposed on the scatter plots of y versus x and z.

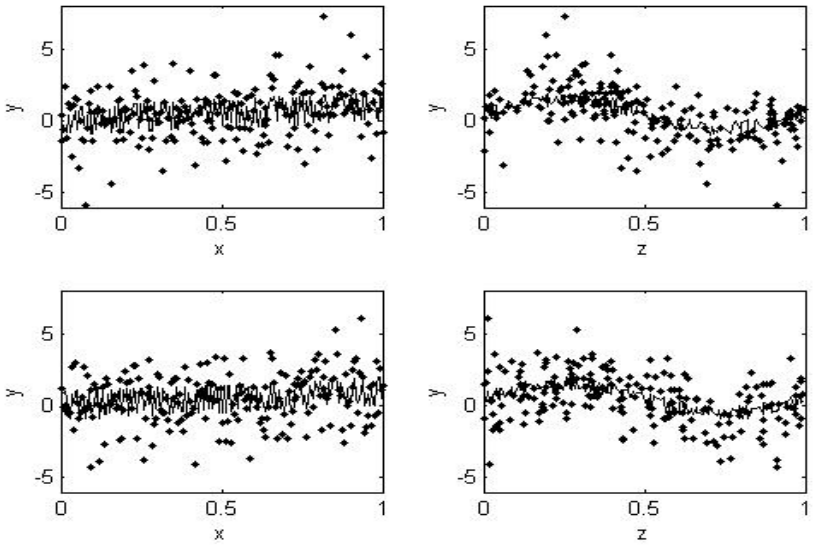

Figure 4.1 The true regression functions superimposed on the scatter plots of y versus x (Left) and z (Right) with Laplacian errors (Upper) and normal errors (Lower)

We set δ in the modified loss function (3.3) to 0.01 and e to 0.1. The radial basis kernel function is utilized in this example, which is,

$$K(z_1, z_2) = \exp(-\frac{1}{\sigma^2}(z_1 - z_2)^2).$$

We consider the generalized comparative Kullback-Leibler distance (GCKL distance; Wahba *et al.*, 1999) as follows;

$$GCKL(\boldsymbol{\lambda}) = \frac{1}{n}\sum_{i=1}^{n} E_y[\ell_e(y_i - \widehat{\mu}_i)],$$

where $\widehat{\mu}_i = \widehat{\mu}(x_i, z_i)$. The assumption of the Laplace distribution of errors provides a closed

form of GCKL distance as follows;

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} E_y[\ell_e(y_i-\widehat{\mu}_i)] = \frac{1}{n}\sum_{i=1}^{n} E_r[\ell_e(r_i+d_i)] \\
&= \frac{1}{n}\sum_{i=1}^{n} E_r[(r_i+d_i-e)1(r_i+d_i>e)+(-r_i-d_i-e)1(r_i+d_i<-e)] \\
&= \frac{1}{n}\sum_{i=1}^{n} E_r[(r_i+d_i-e)1(r_i>-d_i+e)] - \frac{1}{n}\sum_{i=1}^{n} E_r[(r_i+d_i+e)1(r_i<-d_i-e)] \\
&= \frac{1}{n}\sum_{i=1}^{n} E_r[r_i 1(r_i>-d_i+e)] + \frac{1}{n}\sum_{i=1}^{n}(d_i-e)P(r_i>-d_i+e) \\
&- \frac{1}{n}\sum_{i=1}^{n} E_r[r_i 1(r_i<-d_i-e)] - \frac{1}{n}\sum_{i=1}^{n}(d_i+e)P(r_i<-d_i-e) \\
&= \frac{1}{2n}\sum_{i=1}^{n}\exp(-|d_i-e|)[(1+d_i-e)1(-d_i+e<0)+(1-d_i+e)1(-d_i+e\geq 0)] \\
&+ \frac{1}{2n}\sum_{i=1}^{n}(d_i-e)(\exp(-|d_i-e|)-1(d_i-e<0)) \\
&+ \frac{1}{2n}\sum_{i=1}^{n}\exp(-|d_i+e|)[(1+d_i+e)1(d_i+e\geq 0)+(1-d_i-e)1(d_i+e<0)] \\
&+ \frac{1}{2n}\sum_{i=1}^{n}(d_i+e)(\exp(-|d_i+e|)-1(d_i+e<0))
\end{aligned}$$

where $r_i = y_i - \mu_i$ and $d_i = \mu_i - \widehat{\mu}_i$. The assumption of the standard normal distribution of errors provides a closed form of GCKL distance as follows;

$$\begin{aligned}
\frac{1}{n}E_y[\ell_e(y_i-\widehat{\mu}_i)] &= \frac{1}{n}\sum_{i=1}^{n}\frac{f(-d_i+e)}{1-F(-d_i+e)} + \frac{1}{n}\sum_{i=1}^{n}(d_i-e)(1-F(-d_i+e)) \\
&+ \frac{1}{n}\sum_{i=1}^{n}\frac{f(-d_i-e)}{F(-d_i-e)} - \frac{1}{n}\sum_{i=1}^{n}(d_i+e)F(-d_i-e),
\end{aligned}$$

where f is the probability density function and F is the cumulative distribution function of the standard normal distribution.

GCKL distances together with GACV functions were evaluated for different values of kernel parameter in the radial basis kernel function for fixed value of C=100. The 20 values of kernel parameter are equally spaced in the interval [0.1, 2]. We averaged the values over 100 simulated datasets. Figure 4.2 (Left) presents the average values of GCKL distances (solid lines) and GACV functions (dashed lines) versus the values of kernel parameter for Laplacian errors, and Figure 4.2 (Right) is for normal errors. From the figure, we find that the average GACV functions have similar pattern as the average GCKL distances, which implies GACV function is a good estimate of GCKL distance.

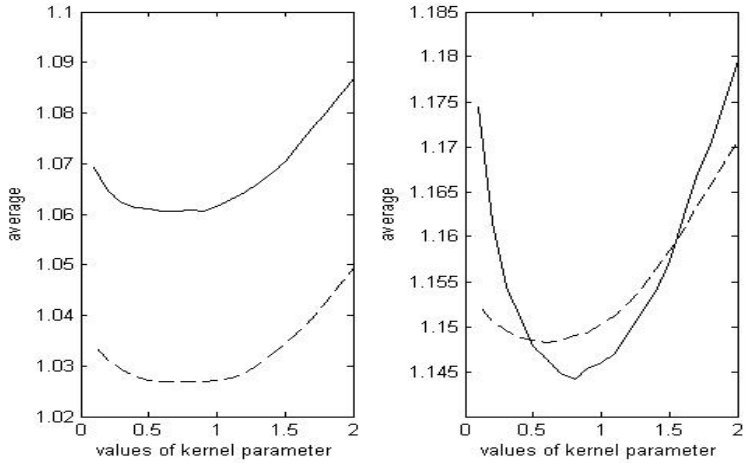

Figure 4.2 The average GCKL distances (solid lines) and GACV functions (dashed lines) from 100 simulated datasets with Laplacian errors (Left) and normal errors (Right)

For the comparison of prediction performance of PLSVR using IRWLS procedure and PLSVR using QP, we use one dataset for training and rest 99 datasets for test.

We select (C, σ^2) as (10,0.3) and (100,0.5) for Laplacian and normal error case, respectively, using GACV function (3.2). The predicted mean squared error (PMSE) is used as prediction performance measure defined by

$$PMSE = \frac{1}{n}\sum_{i=1}^{n}(\widehat{\mu}_i - \mu_i)^2.$$

The averages of 99 PMSEs from PLSVR using IRWLS procedure and QP are obtained as 0.0644 and 0.0666, respectively, (corresponding standard errors are 0.0039 and 0.0040) for Laplacian errors, 0.0586 and 0.0647, respectively, (corresponding standard errors are 0.0042 and 0.0046) for normal errors, which implies that the proposed procedure provides almost same prediction performance as PLSVR using QP. We can see that the proposed procedure works generally well for the partially linear regression in model selection and prediction.

5. Conclusions

In this paper, we dealt with estimating the regression function by PLSVR using IRWLS procedure and obtained GACV function for the proxy of GCKL distance. Through the examples we showed that the proposed procedure derives the satisfying solutions-easy model selection and good prediction performance. We also found that PLSVR using IRWLS procedure is much faster than PLSVR using QP, which implies that the proposed procedure is appropriate for the large training data sets. We showed the generalized approximate cross validation function from PLSVR using IRWLS procedure is good approximate of GCKL distance.

References

Cho, D. H., Shim, J. and Seok, K. H. (2010). Doubly penalized kernel method for heteroscedastic autoregressive data. *Journal of the Korean Data & Information Science Society*, **21**, 155-162.

Hwang, H. (2010). Fixed size LS-SVM for multiclassification problems of large datasets. *Journal of the Korean Data & Information Science Society*, **21**, 561-567.

Kuhn, H. W. and Tucker, A. W. (1951). Nonlinear programming. *Proceedings of 2nd Berkeley Symposium*, 481-492.

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.

Nychka, D., Gray, G., Haaland, P., Martin, D. and O'Connell, M. (1995). A nonparametric approach syringe grading for quality improvement. *Journal of American Statistical Association*, **432**, 1171-1178.

Perez-Cruz, F., Navia-Vazquez, A., Alarcon-Diana, P. L. and Artes-Rodriguez, A. (2000). An IRWLS procedure for SVR. In *Proceedings of European Association for Signal Processing*, EUSIPO 2000, Tampere, Finland.

Platt, J. (1998). *Sequential minimal optimization: A fast algorithm for training support vector machines*, Technical Report MSR-TR-98-14, Microsoft Research, California.

Shim, J., Kim, C. and Hwang, C. (2011). Semiparametric least squares support vector machine for accelerated failure time model. *Journal of the Korean Statistical Society*, **40**, 75-83.

Smola, A. J. and Schölkopf, B. (1998). On a kernel-based method for pattern recognition, regression, approximation and operator inversion. *Algorithmica*, **22**, 211-231.

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

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

Wahba, G., Lin, Y. and Zhang, H. (1999). *Generalized approximate cross validation for support vector machines, or another way to look at margin-like quantities*, Technical Report 1006, University of Wisconsin, Wisconsin.

Wang, L.(Ed.) (2005). *Support vector machines: Theory and application*, Springer, New York.

Yuan, M. (2006). GACV for quantile smoothing splines. *Computational Statistics and Data Analysis*, **50**, 813-829.