Journal of the Korean Data & Information Science Society
2010, **21**(2), 363–369

한국데이터정보과학회지

Mixed-effects LS-SVR for longitudinal data[†]

Daehyeon Cho[1]

Department of Data Science, Inje University

Received 12 February 2010, revised 18 March 2010, accepted 23 March 2010

Abstract

In this paper we propose a mixed-effects least squares support vector regression (LS-SVR) for longitudinal data. We add a random-effect term in the optimization function of LS-SVR to take random effects into LS-SVR for analyzing longitudinal data. We also present the model selection method that employs generalized cross validation function for choosing the hyper-parameters which affect the performance of the mixed-effects LS-SVR. A simulated example is provided to indicate the usefulness of mixed-effects method for analyzing longitudinal data.

Keywords: Generalized cross validation function, hyper-parameter, kernel function, least squares support vector machines, mixed-effects regression model.

1. Introduction

For data which are clustered and/or longitudinal, mixed-effects regression models are becoming increasingly popular (Hedeker and Gibbons, 2006). Mixed-effects models constitute both fixed and random effects. In clustered data, subjects are clustered within an organization such as a hospital, school, clinic or firm. In longitudinal data where individuals are repeatedly assessed, measurements are clustered within individuals. For clustered data the random effects represent cluster effects, while for longitudinal data the random effects represent subject effects. There has been much work done on mixed-effect models even for count data and binomial data (Hwang, 2008; Shim and Seok, 2008).

Support vector machine (SVM), a machine learning method developed by Vapnik (1998), has appealing ability to approximate complex, nonlinear decision boundaries and function estimations. Some studies (Liu *et al.*, 2003; Guler and Kocer, 2005) indicate that SVM is more accurate than artificial neural network (ANN; Allen and Murray, 1993). Because SVM does not seek to minimize generalization error, but rather the upper bound of generalization error, multivariate modeling of sparse data is not problematic (Christianini and Shawe-Taylor, 2000). SVM is less prone to overfitting than ANN. It is an additional strong point of SVM. However, SVM solves quadratic optimization problem which requires the use of optimization routines from numerical libraries. This step is computational intensive, can

[†] 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 (2009-0072211).

[1] Professor, Department of Data Science, Institute of Statistical Information, Inje University, Kimhae 621-749, Korea. E-mail: statcho@inje.ac.kr

be subject to stability problems and is nontrivial to implement. In recent years there have been many new and exciting developments in kernel-based learning, largely stimulated by work in statistical learning theory and SVM. Among kernel machines, least squares support vector machine (LS-SVM, Suykens and Vandewalle, 1999; Suykens *et al.*, 2001) has been proved to be a very appealing and promising method. Solving nonlinear modeling by convex optimization without suffering from many local minima like SVM is one of its strong points. In addition, LS-SVM uses the linear equation which is simple to solve and good for computational time saving. Many tests and comparisons showed great performance of LS-SVM on several benchmark data set problems and are applicable to various types of data (Shim and Lee, 2009). It can be used for variance estimation for replicated data (Shim *et al.*, 2009). Conceptually, the additional explicit primal-dual interpretations from the viewpoint of optimization theory turned out to be the essential of LS-SVM.

In this paper we propose a mixed-effects LS-SVR for analyzing longitudinal data. We also present a generalized cross validation (GCV) function in order to choose the hyper-parameters of the mixed-effects LS-SVR. The rest of this paper is organized as follows. In Section 2 the LS-SVR which can be considered as the fixed-effects LS-SVR in a view of the mixed-effects regression model is briefly reviewed. In Section 3 we propose a mixed-effects LS-SVR for analyzing longitudinal data and present GCV function for the model selection. In Section 4 we perform the numerical studies through the simulated data. In Section 5 we give the conclusions.

2. Fixed-effects LS-SVR

Let the training data set D be denoted by $\{\ \boldsymbol{x}_i, y_i\}\ _{i=1}^{n}$, with each input $\boldsymbol{x}_i \in R^d$ and the output $y_i\ \in\ R$. We consider the case of nonlinear regression. Then, we take the form

$$f(\boldsymbol{x}) = b_0 + \boldsymbol{w}'\phi(\boldsymbol{x}),$$

where the term b_0 is a bias term. 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.

The optimization problem is defined with a regularization parameter C as

$$\min \frac{1}{2}\boldsymbol{w}'\boldsymbol{w} + \ \frac{C}{2}\sum_{i=1}^{n} e_i^2 \tag{2.1}$$

over $\ \boldsymbol{w}, b_0, \boldsymbol{e}\ $ subject to equality constraints

$$y_i = b_0 + \boldsymbol{w}'\phi(\boldsymbol{x}_i) + e_i,\ i = 1, \cdots, n.$$

The Lagrangian function can be constructed as

$$L(\boldsymbol{w}, b_0, e : \alpha) = \frac{1}{2}\boldsymbol{w}'\boldsymbol{w} + \ \frac{C}{2}\sum_{i=1}^{n} e_i^2 - \sum_{i=1}^{n} \alpha_i\ (b_0 + \boldsymbol{w}'\phi(\boldsymbol{x}_i) + e_i - y_i), \tag{2.2}$$

where α_i's are the Lagrange multipliers. The conditions for optimality are given by

$$\begin{aligned}
&\frac{\delta L}{\delta \boldsymbol{w}} = 0 \to \boldsymbol{w} = \sum_{i=1}^{n} \alpha_i \phi(\boldsymbol{x}_i) \\
&\frac{\delta L}{\delta b_0} = 0 \to \sum_{i=1}^{n} \alpha_i = 0 \\
&\frac{\delta L}{\delta e_i} = 0 \to \alpha_i = Ce_i,\ i = 1, \cdots, n \\
&\frac{\delta L}{\delta e_i} = 0 \to b_0 + \boldsymbol{w}'\phi(\boldsymbol{x}_i) + e_i - y_i = 0,\ i = 1, \cdots, n,
\end{aligned}$$

with solution

$$\begin{bmatrix} 0 & \mathbf{1}' \\ \mathbf{1} & K + \boldsymbol{I}/C \end{bmatrix} \begin{bmatrix} b \\ \boldsymbol{\alpha} \end{bmatrix} = \begin{bmatrix} 0 \\ \boldsymbol{y} \end{bmatrix} \tag{2.3}$$

with $\boldsymbol{y} = (y_1, \cdots, y_n)'$, $\mathbf{1} = (1, \cdots, 1)'$, $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_n)'$, and $K = \{K_{kl}\}$, where $K_{kl} = \phi(\boldsymbol{x}_k)'\phi(\boldsymbol{x}_l) = K(\boldsymbol{x}_k, \boldsymbol{x}_l)$, $k, l = 1, \cdots, n$, which are obtained from the application of Mercer's conditions (Mercer, 1909). Several choices of the kernel $K(\cdot, \cdot)$ are possible.

Solving the linear equation (2.3) the optimal bias and Lagrange multipliers, $\widehat{b}_0$ and $\widehat{\alpha}_i$'s are obtained, then the optimal regression function for the given $\boldsymbol{x}_0$ is obtained as

$$\widehat{f}(\boldsymbol{x}_0) = \widehat{b}_0 + \sum_{i=1}^{n} \widehat{\alpha}_i K(\boldsymbol{x}_i, \boldsymbol{x}_0). \tag{2.4}$$

Note that in the nonlinear setting, the optimization problem corresponds to finding the flattest function in the feature space, not in the input space. In fact, LS-SVR has strong advantage that LS-SVR performs particularly well for the nonlinear regression model with high-dimensional covariates.

3. Mixed-effects LS-SVR

We now consider a mixed-effects LS-SVR for analyzing longitudinal data. Let y_{ij} be the j th response variable of the i th subject corresponding to $p \times 1$ fixed-effects covariates $\boldsymbol{x}_{ij}$, where $i = 1, \cdots, N$ and $j = 1, \cdots, n_i$. We assume that y_{ij} is related to covariates $\boldsymbol{x}_{ij}$ in a regression form as

$$y_{ij} = b_0 + \boldsymbol{w}'\phi(\boldsymbol{x}_{ij}) + \boldsymbol{b}_i'\boldsymbol{z}_{ij} + \epsilon_{ij}, fori = 1, 2, \cdots, N, j = 1, 2, \cdots, n_i, \tag{3.1}$$

where b_0 is the bias, $\phi(\boldsymbol{x}_{ij})$ is a nonlinear feature mapping function, $\boldsymbol{z}_{ij}$ is $q \times 1$ random-effects covariates, $\boldsymbol{b}_i$ is $q \times 1$ random-effects parameter vector from $N_q(\mathbf{0}, \boldsymbol{B})$, and $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \cdots, \epsilon_{in_i})'$ is $n_i \times 1$ error vector from $N_{n_i}(\mathbf{0}, \boldsymbol{R}_i)$. Here $\boldsymbol{B}$ and $\boldsymbol{R}_i$ are generally unknown, yet we are not particularly concerned with their estimation. We assume the effects of covariates $\boldsymbol{x}_{ij}$ in the nonparametric part on y_{ij} is not specified. For known $\boldsymbol{B}$ and $\boldsymbol{R}_i$ we can define the optimization problem,

$$\min \frac{1}{2}\boldsymbol{w}'\boldsymbol{w} + \frac{\lambda_1}{2}\sum_{i=1}^{N} \boldsymbol{b}_i'\boldsymbol{B}^{\mathbf{-1}}\boldsymbol{b}_i + \frac{\lambda_2}{2}\sum_{i=1}^{N}\sum_{j,k=1}^{n_i} \epsilon_{ij}\boldsymbol{R}_i^{-1}{}_{jk}\epsilon_{ik} \tag{3.2}$$

subject to equality constraints $y_{ij} = b_0 + \boldsymbol{w}'\boldsymbol{\phi}(\boldsymbol{x}_{ij}) + \boldsymbol{b}_i'\boldsymbol{z}_{ij} + \epsilon_{ij}$.

Here ${\boldsymbol{R}_i^{-1}}_{jk}$ is the (j,k) th element of the inverse matrix of $\boldsymbol{R}_i$, $j,k = 1,\cdots,n_i, i = 1,\cdots,N$, and λ_1 and λ_2 are regularization parameters.Then the Lagrangian function can be constructed as

$$\begin{aligned} L =& \frac{1}{2}\boldsymbol{w}'\boldsymbol{w} + \frac{\lambda_1}{2}\sum_{i=1}^{N}\boldsymbol{b}_i'\boldsymbol{B}^{-1}\boldsymbol{b}_i + \frac{\lambda_2}{2}\sum_{i=1}^{N}\sum_{j,k=1}^{n_i}\epsilon_{ij}{\boldsymbol{R}_i^{-1}}_{jk}\epsilon_{ik} \\ &+ \sum_{i=1}^{N}\sum_{j=1}^{n_i}\alpha_{ij}(y_{ij} - b_0 - \boldsymbol{w}'\boldsymbol{\phi}(\boldsymbol{x}_{ij}) - \boldsymbol{b}_i'\boldsymbol{z}_{ij} - \epsilon_{ij}). \end{aligned} \tag{3.3}$$

where α_{ij}'s are Lagrange multipliers. The conditions for optimality are given by

$$\begin{aligned} &\frac{\partial L}{\partial \boldsymbol{w}} = 0 \to \boldsymbol{w} = \sum_{i=1}^{N}\sum_{j=1}^{n_i}\alpha_{ij}\phi(\boldsymbol{x}_{ij}) \\ &\frac{\partial L}{\partial b_0} = 0 \to \sum_{i=1}^{N}\sum_{j=1}^{n_i}\alpha_{ij} = 0 \\ &\frac{\partial L}{\partial \boldsymbol{b}_i} = \boldsymbol{0} \to \lambda_1\boldsymbol{B^{-1}}\boldsymbol{b}_i - \sum_{j=1}^{n_i}\alpha_{ij}\boldsymbol{z}_{ij} = \boldsymbol{0}, \\ &\frac{\partial L}{\partial \alpha_{ij}} = 0 \to y_{ij} - b_0 - \boldsymbol{w}'\phi(\boldsymbol{x}_{ij}) - \boldsymbol{b}_i'\boldsymbol{z}_i - \epsilon_{ij} = 0, \\ &\frac{\partial L}{\partial \epsilon_{ij}} = 0 \to \alpha_{ij} - \lambda_2\sum_{k}{\boldsymbol{R}_i^{-1}}_{jk}\epsilon_{ik} = 0, i = 1,\cdots,N, j = 1,\cdots,n_i. \end{aligned} \tag{3.4}$$

The estimation of effects with these equations in (3.4) requires the computations of inner products $\phi(\boldsymbol{x}_{ij})'\phi(\boldsymbol{x}_{kl}), i,k = 1,\cdots,N, j,l = 1,\cdots,n_i$, in a potentially higher dimensional feature space. Under certain conditions these demanding computations can be reduced significantly by introducing a kernel function K such that $\phi(\boldsymbol{x}_{ij})'\phi(\boldsymbol{x}_{kl}) = K(\boldsymbol{x}_{ij}, \boldsymbol{x}_{kl})$.We see that the final regression estimation of new points only relies on the bias term b_0 and the Lagrange multipliers α_{ij}'s. These will be found by solving the set of linear equations (3.4). Therefore we will rewrite them into easier form by eliminating $\boldsymbol{w}$ and ϵ_{ij}'s:

$$\begin{bmatrix} 0 & \boldsymbol{1_{N_n}}' \\ \boldsymbol{1_{N_n}} & \boldsymbol{K} + \frac{1}{\lambda_1}\widetilde{\boldsymbol{Z}}\widetilde{\boldsymbol{B}}\widetilde{\boldsymbol{Z}}' + \frac{1}{\lambda_2}\widetilde{\boldsymbol{R}} \end{bmatrix}\begin{bmatrix} b_0 \\ \boldsymbol{\alpha} \end{bmatrix} = \begin{bmatrix} 0 \\ \boldsymbol{y} \end{bmatrix}, \tag{3.5}$$

where $N_n = \sum_{i=1}^{N} n_i$, $\boldsymbol{0}_k$ and $\boldsymbol{1}_k$ are the $k \times 1$ vectors of zeros and ones respectively, $\boldsymbol{K}$ is the $N_n \times N_n$ kernel matrix consisting of $K(\boldsymbol{x}_{ik}, \boldsymbol{x}_{jl})$, $i,j = 1,\cdots,N$, $k,l = 1,\cdots,n_i$, $\widetilde{\boldsymbol{Z}} = diag(\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_N)$ is $N_n \times Nq$ block diagonal matrix with $\boldsymbol{Z}_i = (\boldsymbol{z}_{i1}',\cdots,\boldsymbol{z}_{in_i}')'$, $\widetilde{\boldsymbol{B}} = diag(\boldsymbol{B},\cdots,\boldsymbol{B})$ is $Nq \times Nq$ block diagonal matrix, $\widetilde{\boldsymbol{R}} = diag(\boldsymbol{R}_1,\cdots,\boldsymbol{R}_N)$ is $N_n \times N_n$ block diagonal matrix, $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1',\cdots,\boldsymbol{\alpha}_N')'$ with $\boldsymbol{\alpha}_i = (\alpha_{i1},\cdots,\alpha_{in_i})'$, $\boldsymbol{y} = (\boldsymbol{y}_1',\cdots,\boldsymbol{y}_N')'$ with $\boldsymbol{y}_i = (y_{i1},\cdots,y_{in_i})'$.

The resulting mixed-effects LS-SVR for regression function estimation for a given $(\boldsymbol{x}_o, \boldsymbol{z}_o)$ becomes

$$\widehat{y}(\boldsymbol{x}_o, \boldsymbol{z}_o) = \widehat{b_0} + \sum_{i=1}^{N} \sum_{j=1}^{n_i} \widehat{\alpha_{ij}} K(\boldsymbol{x}_{ij}, \boldsymbol{x}_o) + \widehat{\boldsymbol{b}}_i' \boldsymbol{z}_o, \tag{3.6}$$

where $\hat{b}_0$ and $\hat{\alpha}_{ij}$ are the solution to the linear system and $\widehat{\boldsymbol{b}}_i = \frac{1}{\lambda_1} \boldsymbol{B} \boldsymbol{Z}_i' \widehat{\boldsymbol{\alpha}}_i$. Thus the estimated value $\widehat{\boldsymbol{y}}$ can be expressed in matrix notation as

$$\widehat{\boldsymbol{y}} = \widehat{b}_0 \mathbf{1}_{N_n} + \boldsymbol{K} \widehat{\boldsymbol{\alpha}} + \widetilde{\boldsymbol{Z}} \widehat{\boldsymbol{b}}, \tag{3.7}$$

where $\widehat{\boldsymbol{b}} = (\widehat{\boldsymbol{b}}_1', \cdots, \widehat{\boldsymbol{b}}_N')'$ is a $Nq \times 1$ vector.

The functional structures of the estimation method of the mean and variance functions are characterized by hyper-parameters, the regularization parameters λ_1, λ_2 and other tuning parameters included in the kernel. To choose optimal values of hyper-parameters of the model we use the generalized cross validation function. The inverse of the leftmost matrix in (3.5) can be divided into submatrices as follows,

$$LH^{-1} = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix},$$

where S_{11} is a scalar, S_{12} is a $1 \times N_n$ vector and S_{21} is a $N_n \times 1$ vector and S_{22} is a $N_n \times N_n$ matrix. Then $\widehat{\boldsymbol{y}}$ can be expressed as $\widehat{\boldsymbol{y}} = S\boldsymbol{y}$, where $S = \mathbf{1}S_{12} + (\boldsymbol{K} + \frac{1}{\lambda}{}_1 \widetilde{\boldsymbol{Z}} \widetilde{\boldsymbol{B}} \widetilde{\boldsymbol{Z}}')S_{22}$.The GCV function can be obtained by applying the leave-out-one lemma (Wahba, 1990) and the first order Taylor expansion,

$$GCV(\theta) = \frac{N_n \boldsymbol{y}'(\mathrm{I} - \mathrm{S})' \widetilde{R}^{-1} (\mathrm{I} - \mathrm{S}) \boldsymbol{y}}{(N_n - tr(S))^2}, \tag{3.8}$$

where θ is a set of hyper-parameters. Among the candidates of sets of hyper- parameters, we choose the optimal values of hyper-parameters which minimize the GCV function.

4. Numerical studies

We illustrate the performance of the mixed-effects LS-SVR through the simulated data by comparing with the fixed-effects LS-SVR which does not consider the random effect. A Monte Carlo simulation study is conducted to assess the performance of the mixed-effects LS-SVR. Let y_{ij} be the j th response variable of the i th subject corresponding to covariate t_{ij}, where $i = 1, 2, \cdots, 25$, $j = 1, \cdots, 10$. We shall fit the mixed-effects nonlinear regression model of the form,

$$y_{ij} = b_0 + \mu_{ij} + b_i + \epsilon_{ij} \text{ for } i = 1, 2, \cdots, 25, j = 1, \cdots, 10$$

where b_0 =2, $\mu_{ij} = \sin(\frac{\pi}{4} t_{ij})$, $t_{ij} = j$, ϵ_{ij} is independently generated from $N(0, 1)$ and the random effect of each subject b_i's are generated from a normal distribution $N(0, 9)$. The

Gaussian kernel is utilized to estimate the nonparametric component of regression function in this study. The regularization parameters $\lambda_1 = 1$, $\lambda_2 = 10$ and the kernel parameter $\sigma^2 = 20$ are obtained by GCV function (3.8). Figure 4.1 shows the scatter plots of data points and the results of the true and fitted regression functions for 4 subjects randomly selected. The data points are denoted by " · ". The results of the true and fitted regression functions are denoted by the solid lines, dotted lines (mixed-effects LS-SVR), and dashed lines (fixed-effects LS-SVR), respectively. In Figure 4.1, we can see that the mixed-effects LS-SVR works well for this simulated data since the results of both regression functions are close. For the given data in which covariates of each subject are the same we can see the fixed-effects LS-SVR does not work since it cannot take the random effect of each subject into account. Figure 4.2 shows the histogram of the random effect b_i. The left plot is the histogram for the true b_i, and the right plot is the histogram for the estimated b_i. These histograms show that the estimation of random effects performs well for given data.

5. Conclusions

This paper proposes a mixed-effects LS-SVR for the analysis of longitudinal data in nature with repeated measures of response taken over time. The mixed-effects LS-SVR is nonlinear and nonparametric model because it uses kernel trick and including random-effects terms. As we have shown here, the mixed-effects LS-SVR can be easily used without heavy computations under high-dimensional covariate settings or with large data set, since it takes over all advantages of LS-SVR. An important issue for LS-SVR is model selection. To this end, we provide the GCV method for choosing the hyper-parameters which affect the performance of the mixed-effects LS-SVR.

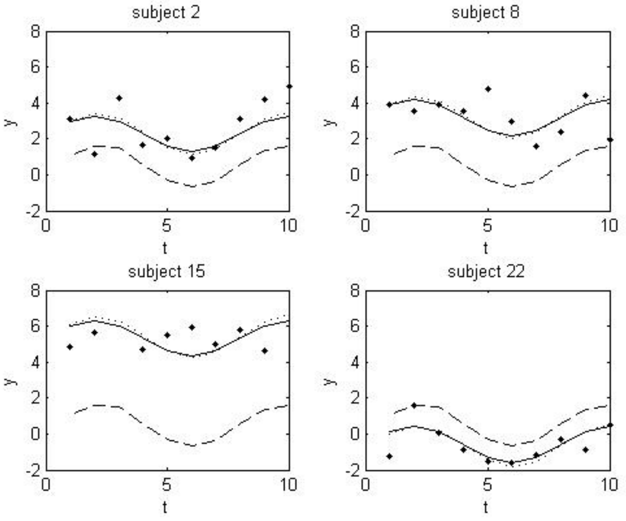

Figure 5.1 Results of regression functions for subjects in simulation study. Observation (dot), true regression function (solid line), fitted regression function by the mixed-effects LS-SVR (dotted line) and fitted regression function by the fixed- effects LS-SVR (dashed line).

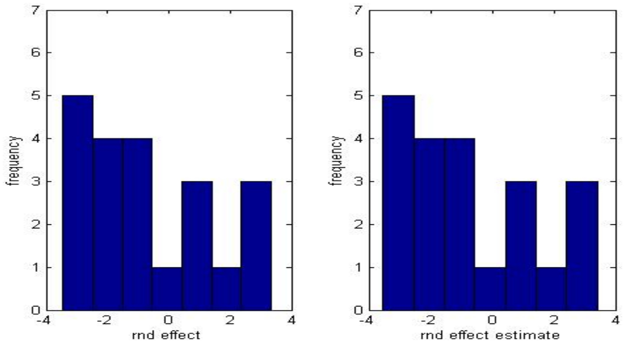

Figure 5.2 Histograms of random effects in simulation study. Left: Histogram of the true b_i. Right: Histogram of the estimated b_i.

References

Allen, J. and Murray, A. (1993). Development of a neural network screening aid for diagnosing lower limb peripheral vascular disease from photoelectric plethysmography pulse waveforms. *Physiological Measurement*, **14**, 13-22.

Christianini, N. and Shawe-Taylor, J. (2000). *An introduction to support vector machines*, Cambridge University Press, Cambridge.

Guler, N. F. and Kocer, S. (2005). Use of support vector machines and neural network in diagnosis of neuromuscular disorders. *Journal of Medical System*, **29**, 271-84.

Hedeker, D. and Gibbons, R. D. (2006). *Longitudinal data Analysis*, John Wiley & Sons, New York.

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

Liu, H. X., Zhang, R. S., Luan, F., Yao, X. J., Liu, M. C., Hu, Z. D. and Fan, B. T. (2003). Diagnosing breast cancer based on support vector machines. *Journal of Chemical Information and Computer Sciences*, **43**, 900-907.

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 Lee, J. T. (2009). Kernel method for autoregressive data. *Journal of Korean Data & Information Science Society*, **20**, 949-964 .

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. and Seok, K. H. (2009). Variance function estimation with LS-SVM for replicated data. *Journal of Korean Data & 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.

Suykens, J. A. K., Vanderwalle, J. and De Moor, B. (2001) Optimal control by least squares support vector machines. *Neural Networks*, **14**, 23-35.

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

Wahba, G. (1990). Spline models for observational data. SIAM, Philadelphia. *CMMS-NSF Regional Conference Series in Applied Mathematics*, **59**.