Journal of the Korean Data & Information Science Society
2017, **28**(2), 461–471

http://dx.doi.org/10.7465/jkdi.2017.28.2.461
한국데이터정보과학회지

Feature selection in the semivarying coefficient LS-SVR[†]

Changha Hwang[1] · Jooyong Shim[2]

[1]Department of Applied Statistics, Dankook University
[2]Department of Statistics, Inje University

Received 22 February 2017, revised 25 March 2017, accepted 27 March 2017

Abstract

In this paper we propose a feature selection method identifying important features in the semivarying coefficient model. One important issue in semivarying coefficient model is how to estimate the parametric and nonparametric components. Another issue is how to identify important features in the varying and the constant effects. We propose a feature selection method able to address this issue using generalized cross validation functions of the varying coefficient least squares support vector regression (LS-SVR) and the linear LS-SVR. Numerical studies indicate that the proposed method is quite effective in identifying important features in the varying and the constant effects in the semivarying coefficient model.

Keywords: Feature selection, generalized cross validation function, least squares support vector regression, semivarying coefficient model, varying coefficient model.

1. Introduction

Hastie and Tibshirani (1993) introduced the varying coefficient model, which is known as powerful and flexible for modeling the dynamic changes of regression coefficients. The varying coefficient model is a useful extension of the classical linear regression model. In the varying coefficient model, the regression coefficients are not set to be constant but are allowed to change with the value of other features called smoothing variables. The varying coefficient model inherits simplicity and easy interpretation of the classical linear regression models and is gaining its popularity in statistics literature in recent years. The introductions, various applications and current research areas of the varying coefficient model can be found in Hoover *et al.* (1998), and Fan and Zhang (2008). A great deal of attention has been focused on the problem of estimating the varying coefficients. Most of this attention has been paid to using kernel smoothing technique. Fan and Zhang (2008) give an excellent review of the varying coefficient models and discusses three approaches in estimating the coefficient function: kernel smoothing, polynomial splines and smoothing splines. Recently, some more flexible varying coefficient models have been developed and discussed. See, for example, Yang

[†] 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 (NRF-2015R1D1A1A01056582), (NRF-2015S1A3A2046715).

[1] Professor, Department of Applied Statistics, Dankook University, Yongin 16890, Korea.

[2] Corresponding author: Adjunct Professor, Department of Statistics, Institute of Statistical Information, Inje University, Kimhae 50834, Korea. E-mail: ds1631@hanmail.net

et al. (2006), Li and Racine (2010), Lee *et al.* (2012), Xue and Qu (2012), and Hwang *et al.* (2016).

Semivarying coefficient models with both nonparametric and parametric components have become increasingly useful in many scientific fields due to their appropriate representation of flexibility and interpretation. Zhang *et al.* (2002) proposed a two-step estimation procedures in the semivarying coefficient model. Fan and Huang (2005) proposed a profile least squares estimation for parametric coefficients. Suykens and Vandewalle (1999) proposed the least squares support vector machine (LS-SVM), which can be seen as the least squares version of SVM (Vapnik, 1995, 1998).By using LS-SVM the linear equations for solutions and the generalized cross validation (GCV) function for the model selection can be easily induced. Shim and Hwang (2015) proposed a method for fitting the semivarying coefficient regression model using least squares support vector regression (LS-SVR) technique, which analyzes the dynamic relation between a response and features. See for further details, Suykens and Vandewalle (1999), Suykens *et al.* (2001), and Hwang and Shim (2016).

The feature selection is used for the better understanding of underlying model and the better prediction performance (reduction of overfitting). Many feature selection methods for linear regression models have been widely used, including the best-subset selection, the stepwise selection, and Bootstrap procedures (Sauerbrei and Schumacher, 1992). LASSO (least absolute shrinkage and selection operator) has been proposed by Tibshirani (1997), which provides the selection of important features and the estimation of regression coefficients simultaneously by shrinking some regression coefficients to zero. Huang *et al.* (2005) proposed the regularization and feature selection approach using LASSO in the accelerated failure time model. Wu *et al.* (2015) proposed a semivarying coefficient model using a penalized rank-based loss function for the estimation and the identification of important feature in high dimensional genetic and genomic data.

In this paper we propose the feature selection method to identify important features in both varying and constant effects (nonparametric and parametric effects). The rest of this paper is organized as follows. In Section 2, we present the semivarying coefficient LS-SVR and a GCV technique in order to choose the optimal values of hyperparameters. In Section 3, we propose the feature selection method identifying important features by using GCV functions of the varying coefficient LS-SVR and the linear LS-SVR. In Section 4 and 5 we present numerical studies and conclusion, respectively.

2. Semivarying coefficient LS-SVR

In this section we first present the semivarying LS-SVR (Shim and Hwang, 2015) and the varying coefficient LS-SVR and a GCV function for choosing the hyperparameters.

Using the vector of features, $\boldsymbol{x}_i \in R^d$, the vector of smoothing variables, $\boldsymbol{u}_i \in R^{d_u}$ (Hastie and Tibshirani, 1993) and the response corresponding to $\boldsymbol{x}_i$ and $\boldsymbol{u}_i$, $y_i \in R^1$, we consider the semivayring coefficient LS-SVR as follows:

$$y_i = f(\boldsymbol{x}_i, \boldsymbol{u}_i) + e_i = a_0(u_i) + \sum_{k \in V_1} x_{ik} a_k(u_i) + b_0 + \sum_{k \in V_2} x_{ik} b_k + e_i, i = 1, \cdots, n, \tag{2.1}$$

where $|V_1| > 0$, $|V_2| > 0$, $V_1 \cup V_2 \subset D = (1, \cdots, d)$,

We assume that $a_k(u_i)$ for $k = 0, \cdots, d$ is nonlinearly related to the smoothing variable $\boldsymbol{u}_i$ such that $a_k(\boldsymbol{u}_i) = \boldsymbol{\omega}_k' \phi(\boldsymbol{u}_i)$ where $\boldsymbol{\omega}_k$ is a corresponding $d_f \times 1$ weight vector to $\phi(\boldsymbol{u}_i)$. Here the nonlinear feature mapping function $\phi : R^{d_u} \to R^{d_f}$ maps the input space to the higher dimensional feature space where the dimension d_f is defined in an implicit way. An inner product in the feature space has an equivalent kernel in the input space, $\phi(\boldsymbol{u}_1)'\phi(\boldsymbol{u}_2) = K(\boldsymbol{u}_1, \boldsymbol{u}_2)$ (Mercer, 1909). Several choices of the kernel $K(\cdot, \cdot)$ are possible. In this paper we utilize the Gaussian kernel as follows:

$$K(\boldsymbol{u}_i, \boldsymbol{u}_j) = \exp\left(-\frac{1}{\sigma^2}||\boldsymbol{u}_i - \boldsymbol{u}_j||^2\right),$$

where $\sigma^2 > 0$ is a kernel (bandwidth) parameter.

With the quadratic loss function the estimator of $(\boldsymbol{\omega}_0, \boldsymbol{\omega}_k, b_0, b_k)$ can be defined as any solution to the following optimization problem:

$$\min L = \frac{1}{2}||\boldsymbol{\omega}_0||^2 + \frac{1}{2}\sum_{k\in V_1}||\boldsymbol{\omega}_k||^2 + \frac{C}{2}\sum_{i=1}^{n}(y_i - f(\boldsymbol{x}_i, \boldsymbol{u}_i))^2, \tag{2.2}$$

where $C > 0$ is a penalty parameter which controls the balance between the smoothness and fitness of the estimator. We can express the above problem by formulation of LS-SVR as follows:

$$\min L = \frac{1}{2}||\boldsymbol{\omega}_0||^2 + \frac{1}{2}\sum_{k\in V_1}||\boldsymbol{\omega}_k||^2 + \frac{C}{2}\sum_{i=1}^{n}e_i^2 \tag{2.3}$$

subject to $e_i = y_i - a_0(\boldsymbol{u}_i) - \sum_{k\in V_1} x_{ik}a_k(\boldsymbol{u}_i) - b_0 - \sum_{k\in V_2} x_{ik}b_k,\ i = 1, \cdots, n.$

We construct a Lagrange function as follows:

$$L = \frac{1}{2}\|\boldsymbol{\omega}_0\|^2 + \frac{1}{2}\sum_{k\in V_1}\|\boldsymbol{\omega}_k\|^2 + \frac{C}{2}\sum_{i=1}^{n}e_i^2 - \sum_{i=1}^{n}\alpha_i\left(e_i - y_i + a_0(\boldsymbol{u}_i) + \sum_{k\in V_1}x_{ik}a_k(\boldsymbol{u}_i) + b_0 + \sum_{k\in V_2}x_{ik}b_k\right),$$

where α_i's are the Lagrange multipliers. Then, the optimality conditions are given by

$$\begin{aligned}
\frac{\partial L}{\partial \boldsymbol{\omega}_0} &= \mathbf{0} \to \boldsymbol{\omega}_0 = \sum_{i=1}^{n}\phi(\boldsymbol{u}_i)\alpha_i, \\
\frac{\partial L}{\partial \boldsymbol{\omega}_k} &= \mathbf{0} \to \boldsymbol{\omega}_k = \sum_{i=1}^{n}x_{ik}\phi(\boldsymbol{u}_i)\alpha_i,\ k \in V_1, \\
\frac{\partial L}{\partial b_0} &= 0 \to \sum_{i=1}^{n}\alpha_i = 0, \\
\frac{\partial L}{\partial b_k} &= 0 \to \sum_{i=1}^{n}x_{ik}\alpha_i = 0,\ k \in V_2 \\
\frac{\partial L}{\partial e_i} &= 0 \to Ce_i - \alpha_i = 0, \\
\frac{\partial L}{\partial \alpha_i} &= 0 \to e_i - y_i + a_0(\boldsymbol{u}_i) + \sum_{k\in V_1}x_{ik}a_k(\boldsymbol{u}_i) + b_0 + \sum_{k\in V_2}x_{ik}b_k = 0,\ i = 1, \cdots, n.
\end{aligned}$$

After eliminating e_i's and $\boldsymbol{\omega}_k$'s, we have the optimal values of (α_i, b_0, b_k)'s from the linear equation as follows:

$$\begin{pmatrix} \boldsymbol{X}(V_1)\boldsymbol{X}(V_1)' \odot K(\boldsymbol{u},\boldsymbol{u}) + \frac{1}{C}\boldsymbol{I} & \boldsymbol{X}(V_2) \\ \boldsymbol{X}(V_2)' & \boldsymbol{O}_{22} \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{b} \end{pmatrix} = \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{pmatrix} \tag{2.4}$$

where $\boldsymbol{X}(V_1) = [\boldsymbol{1}_{n\times 1}, \boldsymbol{x}(V_1)], \boldsymbol{x}(V_1) = \{x_{ik}\}_{i=1,k\in V_1}^n, \boldsymbol{X}(V_2) = [\boldsymbol{1}_{n\times 1}, \boldsymbol{x}(V_2)]$, d_2 =length(V_2), $\boldsymbol{x}(V_2) = \{x_{ik}\}_{i=1,k\in V_2}^n$, $\boldsymbol{b} = b_0 \cup \{b_k\}_{k\in V_2}$, $(d_2+1)\times 1$ vector, $\boldsymbol{O}_{22}$ is the zero matrix of size $(d_2+1)\times(d_2+1)$, and $\odot$ denotes a component-wise product.

Thus, the estimators of $a_k(u_t)$'s are obtained as follows:

$$\widehat{a}_0(\boldsymbol{u}_t) = \sum_{i=1}^n K(\boldsymbol{u}_t, \boldsymbol{u}_i)\widehat{\alpha}_i \text{ and } \widehat{a}_k(\boldsymbol{u}_t) = \sum_{i=1}^n x_{ik}K(\boldsymbol{u}_t, \boldsymbol{u}_i)\widehat{\alpha}_i, k \in V_1, \tag{2.5}$$

The estimated regression function given $(\boldsymbol{x}_t, \boldsymbol{u}_t)$ is obtained as

$$\widehat{f}(\boldsymbol{x}_t, \boldsymbol{u}_t) = \sum_{i=1}^n K(\boldsymbol{u}_t, \boldsymbol{u}_i)\widehat{\alpha}_i + \sum_{i=1}^n \sum_{k\in V_1} x_{tk}x_{ik}K(\boldsymbol{u}_t, \boldsymbol{u}_i)\widehat{\alpha}_i + \widehat{b}_0 + \sum_{k\in V_2} x_{tk}\widehat{b}_k, \tag{2.6}$$

The functional structures of the semivarying coefficient LS-SVR is characterized by hyperparameters (penalty parameter and kernel parameter). To choose the optimal values of hyperparameters of the semivarying coefficient LS-SVR we consider the cross validation(CV) function as follows:

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^n \left(y_i - \widehat{f}(\boldsymbol{x}_i, \boldsymbol{u}_i)^{(-i)}\right)^2, \tag{2.7}$$

where λ is the set of hyperparameters and $\widehat{f}(\boldsymbol{x}_i, \boldsymbol{u}_i)^{(-i)}$ is the regression function estimated without the i th observation. Since for each set of candidates of hyperparameters, $\widehat{f}(\boldsymbol{x}_i, \boldsymbol{u}_i)^{(-i)}$ for $i = 1, \cdots, n$, should be calculated, choosing the optimal hyperparameters using CV function is computationally formidable. By using the leaving-out-one lemma (Craven and Wahba, 1979) the ordinary cross validation function can be obtained as

$$OCV(\boldsymbol{\lambda}) = \frac{1}{n}\sum_{i=1}^n \left(\frac{y_i - \widehat{f}(\boldsymbol{x}_i, \boldsymbol{u}_i)}{1 - \frac{\partial \widehat{f}(\boldsymbol{x}_i, \boldsymbol{u}_i)}{\partial y_i}}\right)^2 = \frac{1}{n}\sum_{i=1}^n \left(\frac{y_i - \widehat{f}(\boldsymbol{x}_i, \boldsymbol{u}_i)}{1 - h_{ii}}\right)^2. \tag{2.8}$$

Here H is the hat matrix such that $\widehat{f}(\boldsymbol{x}, \boldsymbol{u}) = H\boldsymbol{y}$, h_{ii} is the i th diagonal element of $H = (A_1^*, \boldsymbol{X}(V_2))B$, where A_1^* is a block diagonal matrix of $\boldsymbol{X}(V_1)\boldsymbol{X}(V_1)' \odot K(\boldsymbol{u}, \boldsymbol{u})$, $\boldsymbol{X}(V_2)$ is obtained in (2.4), B is a $(n + d_2 + 1)\times n$ leftmost submatrix of the inverse matrix of $\begin{pmatrix} \boldsymbol{X}(V_1)\boldsymbol{X}(V_1)' \odot K(\boldsymbol{u},\boldsymbol{u}) + \frac{1}{C}\boldsymbol{I} & \boldsymbol{X}(V_2) \\ \boldsymbol{X}(V_2)' & \boldsymbol{O}_{22} \end{pmatrix}$ in (2.4).

Replacing h_{ii} by their average $trace(H)/n$, the generalized cross validation (GCV) function can be obtained as

$$GCV(\lambda) = \frac{n\sum_{i=1}^n \left(y_i - \widehat{f}(x_i, u_i)\right)^2}{(n - trace(H))^2}. \tag{2.9}$$

3. Feature selection

In this section we first briefly introduce the varying coefficient LS-SVR and the linear LS-SVR, which are used for the selection of important features. Here the varying coefficient LS-SVR can be considered as the reduced model of semivarying coefficient LS-SVR (Shim and Hwang, 2015). Next we propose the feature selection method identifying important features by using GCV functions of the varying coefficient LS-SVR and the linear LS-SVR.

3.1. Varying coefficient LS-SVR

Using the vector of features, $\boldsymbol{x}_i \in R^d$, the vector of smoothing variables, $\boldsymbol{u}_i \in R^{d_u}$ and the response corresponding to $\boldsymbol{x}_i$ and $\boldsymbol{u}_i$, $y_i \in R^1$, we consider the varying coefficient LS-SVR as follows:

$$y_i = f(\boldsymbol{x}_i, \boldsymbol{u}_i) + e_i = a_0(\boldsymbol{u}_i) + \sum_{k\in V_1} x_{ik} a_k(\boldsymbol{u}_i) + e_i,\ i = 1, \cdots, n, \tag{3.1}$$

where $f(\boldsymbol{x}_i, u_i) = a_0(u_i) + \sum_{k\in V_1} x_{ik} a_k(u_i)$, $i = 1, \cdots, n$, and $|V_1| > 0$, $V_1 \subset V = (1, \cdots, d)$.

We assume that $a_k(\boldsymbol{u}_i)$ for $k = 0, \cdots, d$ is nonlinearly related to the smoothing variables $\boldsymbol{u}_i$ such that $a_k(\boldsymbol{u}_i) = \boldsymbol{\omega}_k' \phi(\boldsymbol{u}_i)$ where $\boldsymbol{\omega}_k$ is a corresponding $d_f \times 1$ weight vector to $\phi(\boldsymbol{u}_i)$.

With the quadratic loss function the estimator of $(\boldsymbol{\omega}_0, \boldsymbol{\omega}_k)$ can be defined as any solution to the following optimization problem:

$$\min L = \frac{1}{2}||\boldsymbol{\omega}_0||^2 + \frac{1}{2}\sum_{k\in V_1} ||\boldsymbol{\omega}_k||^2 + \frac{C}{2}\sum_{i=1}^{n}(y_{ij} - f(\boldsymbol{x}_i, \boldsymbol{u}_i))^2, \tag{3.2}$$

where $C > 0$ is a penalty parameter which controls the balance between the smoothness and fitness of the estimator. We can express the above problem by formulation of LS-SVR as follows:

$$\min L = \frac{1}{2}||\boldsymbol{\omega}_0||^2 + \frac{1}{2}\sum_{k\in V_1} ||\boldsymbol{\omega}_k||^2 + \frac{C}{2}\sum_{i=1}^{n} e_i^2 \tag{3.3}$$

subject to $e_i = y_i - a_0(\boldsymbol{u}_i) - \sum_{k\in V_1} x_{ik} a_k(\boldsymbol{u}_i)$, $i = 1, \cdots, n$.

We construct a Lagrange function as follows:

$$L = \frac{1}{2}||\boldsymbol{\omega}_0||^2 + \frac{1}{2}\sum_{k\in V_1} ||\boldsymbol{\omega}_k||^2 + \frac{C}{2}\sum_{i=1}^{n} e_i^2 - \sum_{i=1}^{n} \alpha_{ij}(e_{ij} - y_i + a_0(\boldsymbol{u}_i) + \sum_{k\in V_1} x_{ik} a_k(\boldsymbol{u}_i)),$$

where α_{ij}'s are the Lagrange multipliers. Then, the optimality conditions are given by

$$\begin{aligned}
\frac{\partial L}{\partial \boldsymbol{\omega}_0} =& \mathbf{0} \rightarrow \boldsymbol{\omega}_0 = \sum_{i=1}^{n} \phi(\boldsymbol{u}_i)\alpha_i, \\
\frac{\partial L}{\partial \boldsymbol{\omega}_k} =& \mathbf{0} \rightarrow \boldsymbol{\omega}_k = \sum_{i=1}^{n} x_{ik}\phi(\boldsymbol{u}_i)\alpha_i, k \in V_1, \\
\frac{\partial L}{\partial e_i} =& 0 \rightarrow Ce_i - \alpha_i = 0, \\
\frac{\partial L}{\partial \alpha_i} =& 0 \rightarrow e_i - y_i + a_0(\boldsymbol{u}_i) + \sum_{k\in V_1} x_{ik} a_k(\boldsymbol{u}_i) = 0, i = 1, \cdots, n.
\end{aligned}$$

After eliminating e_i's and $\boldsymbol{\omega}_k$'s, we have the optimal values of the estimates of α_j's are obtained from the linear equation as follows:

$$\widehat{\boldsymbol{\alpha}} = (\boldsymbol{X}(V_1)\boldsymbol{X}(V_1)' \odot K(\boldsymbol{u},\boldsymbol{u}) + \frac{1}{C}\boldsymbol{I})\boldsymbol{y} \tag{3.4}$$

where $\boldsymbol{X}(V_1) = [\mathbf{1}_{n\times 1}, \boldsymbol{x}(V_1)], \boldsymbol{x}(V_1) = \{x_{ik}\}_{i=1,k\in V_1}^{n}$, $\widehat{\boldsymbol{\alpha}} = (\widehat{\alpha}_1, \cdots, \widehat{\alpha}_n)'$ is $n \times 1$ vector, $\odot$ denotes a component-wise product.

Thus, the estimators of $a_k(\boldsymbol{u}_t)$'s are obtained as follows:

$$\widehat{a}_0(\boldsymbol{u}_t) = \sum_{i=1}^{n} K(\boldsymbol{u}_t, \boldsymbol{u}_i)\widehat{\alpha}_i \text{ and } \widehat{a}_k(\boldsymbol{u}_t) = \sum_{i=1}^{n} x_{ik}K(\boldsymbol{u}_t, \boldsymbol{u}_i)\widehat{\alpha}_i,\ k \in V_1, \tag{3.5}$$

The estimated regression function given $(\boldsymbol{x}_t, \boldsymbol{u}_t)$ is obtained as

$$\widehat{f}(\boldsymbol{x}_t, \boldsymbol{u}_t) = \sum_{i=1}^{n} K(\boldsymbol{u}_t, \boldsymbol{u}_i)\hat{\alpha}_i + \sum_{i=1}^{n}\sum_{k\in V_1} x_{tk}x_{ik}K(\boldsymbol{u}_t, \boldsymbol{u}_i)\hat{\alpha}_i. \tag{3.6}$$

The functional structures of the varying coefficient LS-SVR is characterized by hyperparameters (penalty parameter and kernel parameter). To choose the optimal values of hyperparameters of the varying coefficient LS-SVR we use the generalized cross validation (GCV) function as follows:

$$GCV(\lambda) = \frac{n\sum_{i=1}^{n}\left(y_i - \widehat{f}(\boldsymbol{x}_i, \boldsymbol{u}_i)\right)^2}{(n - trace(H))^2}. \tag{3.7}$$

3.2. Linear LS-SVR

Using the vector of features, $\boldsymbol{x}_i \in R^d$ and the response corresponding to $\boldsymbol{x}_i$, $y_i \in R^1$, we consider the linear LS-SVR as follows:

$$y_i = f(\boldsymbol{x}_i) + e_i = b_0 + \sum_{k\in V_2} x_{ik}b_k + e_i,\ i = 1, \cdots, n, \tag{3.8}$$

where $|V_2| > 0$, $V_2 \subset D = (1, \cdots, d)$.

With the quadratic loss function the estimator of (b_0, b_k) can be defined as any solution to the following optimization problem:

$$\min L = \frac{1}{2}\sum_{k\in V_2} b_k^2 + \frac{C}{2}\sum_{i=1}^{n}(y_i - f(\boldsymbol{x}_i))^2, \tag{3.9}$$

where $C > 0$ is a penalty parameter which controls the balance between the smoothness and fitness of the estimator. We can express the above problem by formulation of LS-SVR as follows:

$$\min L = \frac{1}{2}\sum_{k\in V_2} b_k^2 + \frac{C}{2}\sum_{i=1}^{n} e_i^2 \tag{3.10}$$

subject to $e_i = y_i - b_0 + \sum_{k \in V_2} x_{ik} b_k$, $i = 1, \cdots, n$.

We construct a Lagrange function as follows:

$$L = \frac{1}{2} \sum_{k \in V_2} b_k^2 + \frac{C}{2} \sum_{i=1}^{n} e_i^2 - \sum_{i=1}^{n} \alpha_i \left(e_i - y_i + b_0 + \sum_{k \in V_2} x_{ik} b_k \right),$$

where α_i's are the Lagrange multipliers. Then, the optimality conditions are given by

$$\begin{aligned}
\frac{\partial L}{\partial b_0} =& 0 \rightarrow \sum_{i=1}^{n} \alpha_i = 0, \\
\frac{\partial L}{\partial b_k} =& 0 \rightarrow b_k = \sum_{i=1}^{n} x_{ik} \alpha_i, k \in V_2, \\
\frac{\partial L}{\partial e_i} =& 0 \rightarrow C e_i - \alpha_i = 0, \\
\frac{\partial L}{\partial \alpha_i} =& 0 \rightarrow e_i - y_i + b_0 + \sum_{k \in V_2} x_{ik} b_k = 0, \; i = 1, \cdots, n.
\end{aligned}$$

After eliminating e_{ij}'s and $\omega_{k(j)}$'s, we have the optimal values of (α_i, b_0)'s from the linear equation as follows:

$$\begin{pmatrix} \boldsymbol{x}(V_2)\boldsymbol{x}(V_2)' + \frac{1}{C}\mathrm{I} & \mathbf{1}_{n \times 1} \\ \mathbf{1}'_{n \times 1} & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha} \\ b_0 \end{pmatrix} = \begin{pmatrix} \boldsymbol{y} \\ 0 \end{pmatrix} \tag{3.11}$$

where $\boldsymbol{x}(V_2) = \{x_{ik}\}_{i=1, k \in V_2}^{n}$ and $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_n)'$ is a $n \times 1$ vector.

Thus, the estimators of b_k's are obtained as follows:

$$\widehat{b}_k = \sum_{i=1}^{n} x_{ik} \widehat{\alpha}_i, \; k \in V_2. \tag{3.12}$$

The estimated regression function given (x_t) is obtained as

$$\widehat{f}(x_t) = \widehat{b}_0 + \sum_{k \in V_2} x_{tk} \widehat{b}_k. \tag{3.13}$$

The functional structures of the linear LS-SVR is characterized by the penalty parameter. To choose the optimal value of the penalty parameter of the linear LS-SVR we use the generalized cross validation (GCV) function as follows:

$$GCV(\lambda) = \frac{n \sum_{i=1}^{n} \left(y_i - \widehat{f}(\boldsymbol{x}_i) \right)^2}{(n - trace(H))^2}. \tag{3.14}$$

3.3. Feature selection

To select important features for the varying coefficient effect we use GCV functions of the varying coefficient LS-SVR, and we use GCV functions of the linear LS-SVR for the constant effect. We first define the function, gCV(k) for $k = 1, \cdots, d$, which is obtained as follows:

1) Find $k_1 \in (1, \cdots, d)$ and gCV(1) such that gCV(1) is the minimum of GCV functions computed with one feature and candidate sets of hyperparameters.

For example, when candidate sets of hyperparameters (penalty parameter, kernel parameter) = (100,1), (100,2), (200,1), (200,2), gCV(1) is the minimum of GCV function with $(\boldsymbol{y}, \boldsymbol{x}_{.1}, \boldsymbol{u}, (100, 1))$, GCV function with $(\boldsymbol{y}, \boldsymbol{x}_{.2}, \boldsymbol{u}, (100, 1))$, $\cdots$, GCV function with $(\boldsymbol{y}, \boldsymbol{x}_{.d}, \boldsymbol{u}, (200, 2))$, where $\boldsymbol{x}_{.k} = \{x_{ik}\}_{i=1}^{n}$ is a $n \times 1$ vector.

2) Find $k_2 \in (1, \cdots, d)(k_1)$ and gCV(2) such that gCV(2) is the minimum of GCV functions computed with two features including k_1 and candidate sets of hyperparameters.

$$\vdots$$

d-1) Find $k_{d-1} \in (1, \cdots, d)(k_1, \cdots k_{d-2})$ and gCV($d-1$) such that gCV($d-1$) is the minimum of GCV functions computed with $d-1$ features including $(k_1, \cdots, k_{d-2})$ and candidate sets of hyperparameters.

d) Find gCV(d) such that gCV(d) is the minimum of GCV functions computed with d features and candidate sets of hyperparameters.

The second derivative of gCV represents the rate at which the first derivative of gCV is changing. If the second derivative is negative, it means that gCV is slowly changing direction, that is, adding another features to compute the next gCV does not reduce sufficiently the value of gCV. We want to isolate the point at which the second derivative is negative over the domain of the gCV. Features corresponding to this point is the candidate of the important features. Actually we use the differences instead of derivatives.

4. Numerical studies

In this section we illustrate the performance of the proposed feature selection method through the synthetic and real examples. We use the penalized robust semiparametric approach (PRSA) of Wu *et al.* (2015) for the feature selection and prediction.

Example 4.1 We generate 50 training and test data sets, we identify the important features using training data set and we obtain the prediction performance in terms of the predicted mean squared error using test data set. Each data set consists of 50 features as the following semivarying coefficient models:

$$y_i = a_0(u_i) + a_2(u_i)x_{i4} + a_3(u_i)x_{i10} - 0.8x_{i3} + 1.5x_{i10} + e_i, i = 1, \cdots, 100,$$

where u_i and x_{ik}'s are independently generated from the uniform distribution U(0,1), $a_0(u_i) = u_i^2$ $a_1(u_i) = 2sin(2\pi u_i)$ and $a_2(u_i) = 2cos(2\pi u_i)$.

Table 4.1 shows average numbers of selected features by the proposed method and PRSA, average numbers of selected important features for two effects - varying coefficient effect (x_4, x_{10}), constant effect (x_3, x_{10}) - and recovery rates by the proposed method and PRSA. Recovery rate is defined as the ratio of average number of selected important features ((x_4, x_{10}) and (x_3, x_{10})) to the average number of selected features by the proposed method and PRSA. For both the varying coefficient effect and the constant effect, the average number of selected important features and the recovery rates of the proposed method are higher than those of PRSA.

Table 4.1 Average number of selected features and average number of selected important features in 50 synthetic training data sets, and recovery rate on two methods. The numbers in parentheses are standard errors.

	effect	Avg selected features	Avg selected important features	Recovery rate
proposed	VC	2.28 (0.1071)	1.52 (0.0769)	0.6667
	Constant	2.18 (0.1057)	0.98 (0.0782)	0.4495
PRSA	VC	2.36 (0.1975)	0.14 (0.0572)	0.0593
	Constant	1.26 (0.2153)	0.32 (0.0725)	0.2540

We obtained the average predicted mean squared errors and their standard errors by the semivarying coefficient LS-SVR for 50 test data sets with selected features by the proposed method as (0.4725, 0.0394), and by PRSA (1.3485, 0.0552), which implies that the semivarying LS-SVR has better prediction performance than PRSA. For reference the average predicted mean squared errors and their standard errors by the semivarying LS-SVR without feature selection were obtained as (5.6926, 0.2631)

Example 4.2 For real example we consider a subset of the wage data set studied in Wooldridge (2012), which consists of 5 variables collected on each of 526 working individuals for the year 1976. The response variable is the logarithm of the wage (in dollars per hour), and features are u (years of education), x_1 (years of potential labor force experience), x_2 (an indicator for female), and x_3 (marital status). The last two features are binary (zero-one) in nature and serve to indicate qualitative features of the individual (the person is female or not; the person is married or not). Taking these features into account, we consider the semivarying coefficient model.

To test the feature selection performance, we standardize x_1 and add $(x_4, \cdots, x_{50})$ generated from the standard normal distribution. We randomly divide the whole data into 263 training data and 263 test data. and We repeat this procedure 50 times to identify the important features using training data set and we obtain the prediction performance predicted mean squared error using test data set with selected features. Table 4.2 shows average numbers of selected features and average numbers of selected important features for either effects - varying coefficient effect or constant effect - by the proposed method and PRSA. In Figure 4.1 (Left) (x_1, x_3) are identified as the important features for the varying coefficient effect, and (x_1, x_2) as the important features for the constant effect by the proposed method. In Figure 4.1 (Right) (x_1, x_2) are identified as the important features for the varying coefficient effect, and x_3 as the important feature for the constant effect by PRSA. We can see that the proposed feature selection method provides the larger average number of selected important features than PRSA.

Table 4.2 Average number of selected features and average number of selected important features in 50 training data sets generated from the wage data set, and recovery rate on two methods. The numbers in parentheses are standard errors.

	Avg selected features	Avg selected important features	Recovery rate
proposed	2.96 (0.1399)	2.38 (0.0693)	0.8041
PRSA	1.44 (0.1595)	1.06 (0.0339)	0.7361

We obtained the average predicted mean squared errors and their standard errors by the semivarying coefficient LS-SVR for 50 test data sets with selected features by the proposed method as (0.2909, 0.0616), and by PRSA (0.3126, 0.0769). For reference the average predicted mean squared errors and their standard errors by the semivarying LS-SVR without feature selection were obtained as (4.733, 0.0082).

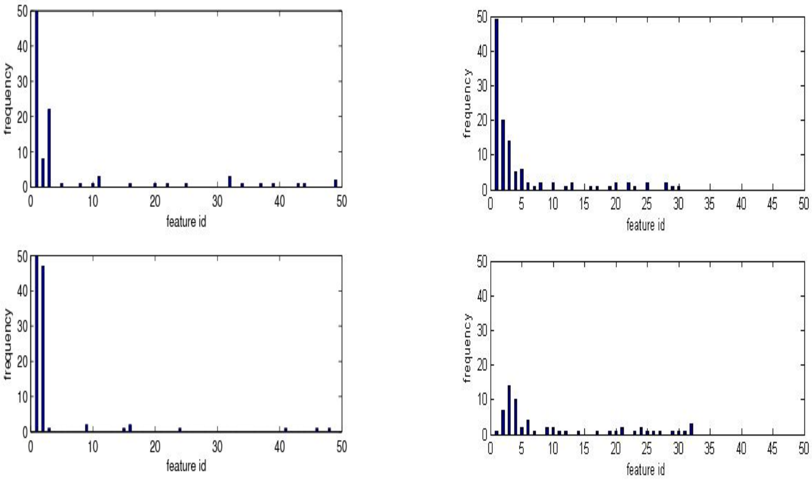

Figure 4.1 requency of selected features for the varying coefficient effect (Upper) and the constant effect (Lower) by the proposed method (Left) and PRSA (Right)

5. Conclusions

We have developed a feature selection method for semivarying coefficient LS-SVR for identifying important features. Important issues in such semivarying coefficient model are how to estimate the parametric and nonparametric components and how to identify important features. In this paper we have proposed a feature selection method which is able to attack these issues. Numerical studies indicate that the proposed feature selection method is quite effective in identifying constant and varying coefficients in a semivarying coefficient model.

Through the examples we showed that the proposed feature selection method derives the satisfying solutions. Also we showed that the proposed method is simple and reliable.

References

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.

Fan, J. and Huang, T. (2005). Profile likelihood inferences on semiparametric varying coefficient partially linear models. *Bernoulli*, **11**, 1031-1057.

Fan, J. and Zhang, W. (2008). Statistical methods with varying coefficient models. *Statistics and Its Interface*, **1**, 179-195.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society B*, **55**, 757-796.

Hoover, D. R., Rice, J. A., Wu, C. O. and Yang, L. P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, **85**, 809-822.

Huang, J., Ma, S. and Xie, H. (2005). *Regularized estimation in the accelerated failure time model with high dimensional covariates*, Technical Report No. 349, Department of Statistics and Actuarial Science, The University of Iowa, IA, USA.

Hwang, C. and Shim, J. (2016). Deep LS-SVM for regression. *Journal of the Korean Data & Information Science Society*, **27**, 827-833.

Hwang, C., Bae, J. and Shim, J. (2016). Robust varying coefficient model using L1 penalized locally weighted regression. *Journal of the Korean Data & Information Science Society*, **27**, 1059-1066.

Lee, Y. K., Mammen, E. and Park, B. U. (2012). Flexible generalized varying coefficient regression models. *Annals of Statistics*, **40**, 1906-1933.

Li, Q. and Racine, J. S. (2010). Smooth varying-coefficient estimation and inference for qualitative and quantitative data. *Econometric Theory*, **26**, 1607-1637.

Mercer. J. (1909) Function of positive and negative type and their connection with theory of integral equations. *Philosophical Transactions of Royal Society A*, 415-446.

Sauerbrei, W. and Schumacher, M. (1992). A bootstrap resampling procedure for model building: Application to the Cox regression model. *Statistical Medicine*, **11**, 2093-2099.

Shim, J. and Hwang, C. (2015). Varying coefficient modeling via least squares support vector regression. *Neurocomputing*, **161**, 254-259.

Suykens, J. A. K. and Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Processing Letters*, **9**, 293-300.

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

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine*, **16**, 385-395.

Vapnik, V. (1995). *The nature of statistical learning theory*, Springer, Berlin.

Vapnik, V. (1998). *Statistical Learning Theory*, Wiley, New York.

Wooldridge, J. M. (2012). *Introductory econometrics: A modern approach*, South-Western Cengage Learning, Mason.

Wu, C., Shi. X., Cui, Y. and Ma, S. (2015) A penalized robust semiparametric approach for gene-environment interactions. *Statistics in Medicine*, **34**, 4016-4030.

Xue, L. and Qu, A. (2012). Variable selection in high-dimensional varying-coefficient models with global optimality. *Journal of Machine Learning Research*, **13**, 1973-1998.

Yang, L., Park, B. U., Xue, L. and H ä rdle, W. (2006). Estimation and testing for varying coefficients in additive models with marginal integration. *Journal of the American Statistical Association*, **101**, 1212-1227.

Zhang, W., Lee, S. and Song, X. (2002). Local polynomial fitting in semivarying coefficient models. *Journal of Multivariate Analysis*, **82**, 166-188.