Scientific Research Publishing

Open Journal of Statistics, 2019, 9, 555-570
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Variance Estimation for High-Dimensional Varying Index Coefficient Models

Miao Wang*, Hao Lv, Yicun Wang

Department of Statistics, School of Economics, Jinan University, Guangzhou, China
Email: *williamjnu@163.com

How to cite this paper: Wang, M., Lv, H. and Wang, Y.C. (2019) Variance Estimation for High-Dimensional Varying Index Coefficient Models. *Open Journal of Statistics*, **9**, 555-570.
https://doi.org/10.4236/ojs.2019.95037

Received: September 11, 2019
Accepted: October 5, 2019
Published: October 8, 2019

Abstract

This paper studies the re-adjusted cross-validation method and a semiparametric regression model called the varying index coefficient model. We use the profile spline modal estimator method to estimate the coefficients of the parameter part of the Varying Index Coefficient Model (VICM), while the unknown function part uses the B-spline to expand. Moreover, we combine the above two estimation methods under the assumption of high-dimensional data. The results of data simulation and empirical analysis show that for the varying index coefficient model, the re-adjusted cross-validation method is better in terms of accuracy and stability than traditional methods based on ordinary least squares.

Keywords

High-Dimensional Data, Refitted Cross-Validation,
Varying Index Coefficient Models, Variance Estimation

1. Introduction

The variance estimate, in this paper, is the residual variance of the model. In the process of statistical modeling, the variance estimation of the model has been extensively studied. Most of the research methods are simple two-stage method, in the first stage, the important variables in the model are selected by the method of variable selection; in the second stage, the variance is estimated by the ordinary least squares method. In the first phase, the traditional variable selection method has two criteria, namely the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). These two traditional methods use the empirical likelihood method to select the model with the smallest AIC and BIC values. At the same time, the variables contained in the model are the selected optimal variables. However, this variable selection method is neither con-

tinuous nor ordered. Therefore, the variance of the model estimated by the traditional method will be large. Moreover, with the development of technology, high-dimensional data is applied to all aspects of life. The number of variables increases exponentially, and the calculation of the above two criteria also shows an exponential increasing trend, so the above method cannot be applied to high-dimensional data.

In the past research, many important variable selection methods such as LASSO (Least Absolute Shrinkage Selection Operator) and SCAD (Smoothly Clipped Absolute Deviation) have been proposed. LASSO was proposed by Tibshirani (1996) [1]. For more details, see Fan and Peng (2004), Zhao and Yu (2006), Bunea (2007), Zhang and Huang (2008), Lv and Fan (2009), Fan and Lv (2011), and Kim (2008) [2]-[8]. In this method, a penalty term is added on the basis of the ordinary least squares method, and the coefficient value is reduced to 0, so that the corresponding variable is excluded from the model. Another type of variable selection tool is DS (Dantzig Selector). This method was first proposed by Candes and Tao (2005) [9] and can be easily reshaped into a linear model. Fan and Lv (2008) [10] sort the covariance matrix between covariate and response variables, and then select the first few variables with the largest correlation coefficient to complete the variable selection. This method is called SIS (Sure Independence Screening). In later studies, some scholars extended the SIS, namely the iterative SIS (ISIS) method: the regression analysis was performed using the variables and dependent variables selected by SIS, and the regression residuals were replaced with response variables. Then continue to use the SIS method for a new round of variable selection. And repeat the above steps until all the important variables. For details, see Fan *et al*. (2009) [11]. After screening out the important variables, the second step of the simple two-stage method is generally calculated by least squares method. However, in order to overcome the root cause of the dimension, many scholars study the variance estimation in the case of high-dimensional data. Fan *et al*. proposed a re-adjusted cross-validation method (RCV) in 2012 to improve the simple two-stage approach. It is proved that the variance estimated by this method is stable and accurate. Zhao *et al*. (2014) [12] studied the variance estimation of linear models under certain assumptions. Reid, Tibshirani, Friedman (2016) [13] studied the model residual estimation in LASSO regression and performed a large number of simulations. They considered that the variance estimation of the residual sum of squares based on adaptive regularization parameter selection has the properties of finite samples.

A well-behaved variance estimation method can improve the prediction accuracy of the model and better explain the socio-economic phenomena. However, it is more important to choose a suitable regression model. There is also a large amount of literature on the study of regression models. When the data dimension is low, the parametric model and the nonparametric model are sufficient to solve the problem. But as the dimension increases, a more flexible semi-parametric

model is more suitable. The literature research on semi-parametric models is mostly focused on the introduction of new models, such as linear models, add-on models, and so on. Hastie and Tibishirani (1993) [14] proposed the Variable Coefficient Model (VCM), which has been widely used in practical applications. In addition, some scholars studied the single index coefficient model (SICM). The Variable Coefficient Single Index Model (VICSIM) was proposed by Wong *et al*. (2008) [15]. Ma and Song (2014) [16] proposed the varying index coefficient model for the first time, which has overcome the problems that the variable coefficient model cannot solve. Most scholars apply variable selection methods such as SIS, LASSO, and SCAD to the parametric model, while the method used in nonparametric estimation are kernel estimation, local linear kernel estimation, and spline functions. For the estimation of semi-parametric regression models, such as partial linear regression model, variable coefficient model, single-index model, etc., the parameter part is estimated by Profile Least Square Estimation (PLSE), and its non-parametric part is still using the previous non-parametric method. For example, Xue and Liang (2010) [17] used the PLSE method of kernel estimation when estimating the non-parametric part of the single-index model. However, there are few literatures on varying index coefficients proposed in 2015, and the related estimation algorithms mainly use the profile least squares estimation method with B-spline to estimate the variable coefficient index model. Lv *et al*. (2016) [18] improved the PLSE, proposed a robust estimation procedure combining the logarithmic regression and the B-spline, and established the large sample property of the parameter estimation. The estimation of the unknown coefficient β is estimated by the profile spline modal estimator method (PSME). Moreover, in order to obtain the progressive distribution of the unknown function $m_l(Z)$, they also proposed a two-stage method of local linear kernel estimation.

The rest of the paper is organized as follows. In Section 2, we briefly introduce the varying index coefficient model, including the estimation method, the statistical inference of the coefficients, and the RCV estimation of the model. In Section 3, simulation studies are conducted to evaluate the finite sample performance of the proposed methods. In Section 4, a real data set is analyzed to compare the proposed methods with the existing methods. A discussion is given in Section 5.

2. Methodology

2.1. Varying Index Coefficient Models

The semiparametric model is widely used in regression models, especially the varying coefficient model (VCM) proposed by Hastie and Tibishirani in 1993, which has been widely used in real data. An important feature of the varying coefficient model is that the coefficients of its covariates are controlled by smooth functions, which can show nonlinear reactions. The form of the variable coefficient model is as follows:

$$Y = \sum_{l=1}^{d} m_l(Z) X_l + \varepsilon \tag{2.1}$$

where Y is a response variable, $X = (X_1, \cdots, X_p)^{\mathrm{T}}$ and $Z \in [0,1]$ (for simplicity) are explanatory covariates, $m(\cdot) = (m_1(\cdot), \cdots, m_p(\cdot))^{\mathrm{T}}$ is a p-dimensional vector of the unknown coefficient functions, and model error ε is independent of (X, Z) with mean zero and finite variance σ^2. The variable coefficient model of Equation (2.1) faces two challenges in the case of today's complex data. First, the variable Z has little effect relative to Y, so the interaction between the variables Z and X is difficult to detect; second, in many complex situations, Z is multi-dimensional, for example, studying the effects between chemical constituents. Thus, the coefficient function $m_l(Z)$ in the VCM model will fall into the dimension curse. To overcome these two problems, Ma and Song proposed the Varying index Coefficient Model (VICM) in 2015. The varying index coefficient model is as follows:

$$Y = m(Z, X, \beta) + \varepsilon = \sum_{l=1}^{d} m\left(Z^{\mathrm{T}} \beta_l\right) X_l + \varepsilon \tag{2.2}$$

where $\beta_l = (\beta_{l1}, \cdots, \beta_{lp})^{\mathrm{T}}$ is the coefficient of the variable Z and β_{lk} is the coefficient of Z_k in Z. The introduction of the varying index coeffcient model was based on Ma and Song's study of this biomedical project that affects children's growth rates.

2.2. Estimation Procedure for the VICM

The estimation of the varying index coefficient models has two main aspects: one is the estimation of the parameter part β, and the other is the estimation of the function coefficient $m_l(u_l)$ of the non-parametric part. In this paper, the estimation of the unknown coefficient β is estimated by the profile spline modal estimator method (PSME). Once β is fixed, the unknown function coefficient $m_l(u_l)$ is estimated using B-spline. The specific estimation process of the varying index coefficient models is as follows.

Let $\{(X_i, Z_i, Y_i), 1 \le i \le n\}$ be the independent and identically distributed samples from model (2.2). Our main interest is to estimate the coefficient vectors β_l and the non-parametric functions $m_l(\cdot)$ for $l = 1, \cdots, d$. The estimation of β_l and $m_l(\cdot)$ in VICM is equivalent to maximizing

$$\frac{1}{n} \sum_{i=1}^{d} \phi_{h_1} \left\{ Y_i - \sum_{l=1}^{d} m_l\left(Z_i^{\mathrm{T}} \beta_l\right) X_{il} \right\} \tag{2.3}$$

subject to the constraint $\|\beta_l\| = 1$ and $\beta_{l1} > 0$, where $\phi_{h_1}(t) = h_1^{-1} \phi(t/h_1)$, ϕ_t is a kernel density function symmetric about 0 and h_1 is a bandwidth which determines the degree of robustness of the estimate. We use the standard normal density for ϕ_t throughout this paper to simplify the calculation. We use a basic approximation to estimate nonparametric functions. That is, we approximate $m_l(\cdot)$ by the B-spline basis function because they have bounded support and are numerically stable. More specially, let $B_q(u) = (B_{1q}(u), \cdots, B_{J_n q}(u))^{\mathrm{T}}$ be the

B-spline basis functions of order $q(q \geq 2)$, where $J_n = N_n + q$ and N_n is the number of interior knots for a knot sequence

$$\xi_1 = \cdots = 0 = \xi_q < \xi_{q+1} < \cdots < \xi_{N_n+q} < 1 = \xi_{N_n+q+1} = \cdots = \xi_{N_n+2q},$$

where N_n increases along with the sample size *n*. Consider the distance between two neighboring knots $H_i = \xi_i - \xi_{i-1}$ and $H = \max_{1 \leq i \leq N_n+1}\{H_i\}$. Then, there exists constants C_0 such that $\frac{H}{\min_{1 \leq i \leq N_n+1}\{H_i\}} < C_0$, $\max_{1 \leq i \leq N_n}\{H_{i+1} - H_i\} = o\left(N_n^{-1}\right)$. Let $U_l(\beta_l) = Z^{\mathrm{T}}\beta_l$, without loss of generality, we assume that $U_l(\beta_l)$ is confined in a compact set $[0,1]$. Then, nonparametric functions $m_l(u_l)$ can be approximated by

$$m_l(u_l) \approx B_q(u_l)^{\mathrm{T}} \lambda_l(\beta), \quad l = 1, \cdots, d \tag{2.4}$$

where $\lambda_l(\beta) = \left(\lambda_{s,l}(\beta) : 1 \leq s \leq J_n\right)^{\mathrm{T}}$. Let $\lambda(\beta) = \left(\lambda_1(\beta)^{\mathrm{T}}, \cdots, \lambda_d(\beta)^{\mathrm{T}}\right)^{\mathrm{T}}$. Based on the above approximation, the objective function (2.3) becomes

$$\frac{1}{n}\sum_{i=1}^{n}\phi_{h2}\left\{Y_i - \sum_{l=1}^{d}\sum_{s=1}^{J_n} B_{s,q}\left(U_{il}(\beta_l)\right)\lambda_{s,l}X_{il}\right\}. \tag{2.5}$$

Subsequently, we estimate the parameter vectors β_l and the nonparametric functions $m_l(\cdot)$ in two steps below.

Step 1. Given β, we obtain estimate $\hat{\lambda}(\beta)$ of $\lambda(\beta)$ by maximizing the objective function (2.5). Then, the estimator of $m_l(u_l)$ can be obtained by

$$\hat{m}_l(u_l, \beta) = \sum_{s=1}^{J_n} B_{s,q}(u_l)\hat{\lambda}_{s,l}(\beta) = B_q(u_l)^{\mathrm{T}}\hat{\lambda}_l(\beta). \tag{2.6}$$

In order to obtain efficient estimators of β, the "remove-one-component" method is employed. Specifically, for $\beta_l = \left(\beta_{l1}, \cdots, \beta_{lp}\right)^{\mathrm{T}}$, let $\beta_{l,-1} = \left(\beta_{l2}, \cdots, \beta_{lp}\right)^{\mathrm{T}}$ be a $p-1$ dimensional vector by removing the 1st component β_{l1} in β_l for all $1 \leq l \leq d$. Then β_l can be rewritten as

$$\beta_l = \beta_l\left(\beta_{l,-1}\right) = \left(\sqrt{1 - \left\|\beta_{l,-1}\right\|^2}, \beta_{l,-1}^{\mathrm{T}}\right)^{\mathrm{T}}, \quad \left\|\beta_{l,-1}\right\|^2 < 1. \tag{2.7}$$

Thus, β_l is infinitely differentiable with respect to $\beta_{l,-1}$ and the Jacobian matrix is

$$J_l = \frac{\partial \beta_l}{\partial \beta_{l,-1}} = \begin{pmatrix} -\beta_{l,-1}^{T} \Big/ \sqrt{1 - \left\|\beta_{l,-1}\right\|^2} \\ I_{p-1} \end{pmatrix}, \tag{2.8}$$

where I_p is the $p \times p$ identity matrix. We denote $\beta_{-1} = \left(\beta_{l,-1}^{\mathrm{T}}, \cdots, \beta_{d,-1}^{\mathrm{T}}\right)^{\mathrm{T}}$ and reformulate the parameter space of β_{-1} as follows:

$$\Theta_{-1} = \left\{\beta_{-1} = \left(\beta_{l,-1}^{\mathrm{T}} : 1 \leq l \leq d\right)^{\mathrm{T}} : \left\|\beta_{l,-1}\right\|^2 < 1, \beta_{l,-1} \in R^{p-1}\right\}. \tag{2.9}$$

Let $\beta = \beta(\beta_{-1})$ with $\beta_l = \beta_l\left(\beta_{l,-1}\right)$ for $1 \leq l \leq d$. Since the estimation procedure of β requires estimates of both m_l and its first order derivative $\dot{m}_l$. We can adopt the spline functions of one order lower than that of m_l to approximate the $\dot{m}_l$. Following Ma and Song (2014), a spline estimator of $\dot{m}_l$

is given by

$$\hat{\dot{m}}_l(u_l,\beta)=\sum_{s=1}^{J_n}\dot{B}_{s,q}(u_l)\hat{\lambda}_{s,l}(\beta)=\sum_{s=2}^{J_n}B_{s,q-1}(u_l)\hat{\omega}_{s,l}(\beta) \tag{2.10}$$

where $\hat{\omega}_{s,l}(\beta)=(q-1)\{\hat{\lambda}_{s,l}(\beta)-\hat{\lambda}_{s-1,l}(\beta)\}/(\xi_{s+q-1}-\xi_s)$ for $2\le s\le J_n$. Thus, one has

$$\hat{\dot{m}}_{s,l}(u_l,\beta)=B_{q-1}(u_l)^{\mathrm{T}}D_1\hat{\lambda}_l(\beta),$$

where $B_{q-1}(u_l)=(B_{s,q-1}(u_l):2\le s\le J_n)^{\mathrm{T}}$ and

$$D_1=(q-1)\begin{bmatrix}\frac{-1}{\xi_{q+1}-\xi_2} & \frac{1}{\xi_{q+1}-\xi_2} & 0 & \cdots & 0\\ 0 & \frac{-1}{\xi_{q+2}-\xi_3} & \frac{1}{\xi_{q+2}-\xi_3} & \cdots & 0\\ \vdots & \vdots & \ddots & \ddots & \vdots\\ 0 & 0 & \cdots & \frac{-1}{\xi_{N+2q-1}-\xi_{N+q}} & \frac{1}{\xi_{N+2q-1}-\xi_{N+q}}\end{bmatrix}_{(J_n-1)\times J_n}.$$

Step 2. After this re-parametrization, combine with the estimators $\dot{m}_l$ and $\hat{\dot{m}}_l$ for $l=1,\cdots,d$, we can construct the profile spline modal objective function for the parametric components. Then, we can obtain the estimator $\hat{\beta}_{-1}$ of β_{-1} by maximizing $L_n(\beta(\beta_{-1}))$ over $\beta_{-1}\in\Theta_{-1}$, where

$$L_n(\beta(\beta_{-1}))=\frac{1}{n}\sum_{i=1}^{n}\phi_{h2}\left\{Y_i-\sum_{l=1}^{d}\sum_{s=1}^{J_n}B_{s,q}(U_{il}(\beta_l))\lambda_{s,l}(\beta)X_{il}\right\}, \tag{2.11}$$

which is equivalent to solve the following estimating equations:

$$\begin{aligned}&\partial L_n(\beta(\beta_{-1}))/\partial\beta_{-1}\\&=-\frac{1}{n}\sum_{i=1}^{n}\dot{\phi}_{h2}\left\{Y_i-\sum_{l=1}^{d}\sum_{s=1}^{J_n}B_{s,q}(U_{il}(\beta_l))\hat{\lambda}_{s,l}(\beta)X_{il}\right\}\\&\quad\times\begin{Bmatrix}\left\{\hat{\dot{m}}_1(U_{i1}(\beta_1),\beta)X_{i1}J_1^{\mathrm{T}}Z_i+\left(\partial\hat{\lambda}(\beta)^{\mathrm{T}}/\partial\beta_{1,-1}\right)D_i(\beta)\right\}\\ \vdots\\ \left\{\hat{\dot{m}}_d(U_{id}(\beta_d),\beta)X_{id}J_d^{\mathrm{T}}Z_i+\left(\partial\hat{\lambda}(\beta)^{\mathrm{T}}/\partial\beta_{d,-1}\right)D_i(\beta)\right\}\end{Bmatrix}\\&=0\end{aligned} \tag{2.12}$$

where $D_i(\beta)=(D_{i,sl}(\beta_l),1\le s\le J_n,1\le l\le d)^{\mathrm{T}}$ with $D_{i,sl}(\beta_l)=B_{s,q}(U_{il}(\beta_l))X_{il}$, $\hat{\dot{m}}_l(\cdot,\beta)$ is given in (2.10) and $\dot{\phi}_{h2}$ is the first derivative of ϕ_{h2}. We obtain the estimate of β_{-1}, say, $\hat{\beta}_{-1}$ and then obtain $\hat{\beta}$ via the transformation (2.7). Thus, we call the estimator $\hat{\beta}$ as the profile spline modal estimator (PSME).

3. Simulation Studies

3.1. Results in Finite Sample

In this section, we conduct simulation studies to evaluate the finite sample per-

formance of the proposed methodology. We generate data from the following VICM:

$$Y_i = m\left(Z_i, X_i, \beta\right) + \varepsilon_i = m_1\left(Z_i^{\mathrm{T}}\beta_1\right)X_{i1} + m_2\left(Z_i^{\mathrm{T}}\beta_2\right)X_{i2} + m_3\left(Z_i^{\mathrm{T}}\beta_3\right)X_{i3} + \varepsilon_i \quad (3.1)$$

with $X_i = \left(X_{i1}, X_{i2}, X_{i3}\right)^{\mathrm{T}}$, where X_i is generated from Bernoulli (p = 0.5), and $\left(X_{i2}, X_{i3}\right)^{\mathrm{T}}$ is drawn from a bivariate normal distribution with mean 0, variance 1, and covariance 0.2. To generate $Z_i = \left(Z_{i1}, Z_{i2}, Z_{i3}\right)^{\mathrm{T}}$, we first sample $\left(Z_{i1}^*, Z_{i2}^*, Z_{i3}^*\right)^{\mathrm{T}}$ from a multivariate normal with mean 0, variance 1, and covariance 0.2, and then let $Z_{ik} = \Phi\left(Z_{ik}^*\right) - 0.5, k = 1,2,3$, where $\Phi(\cdot)$ is the CDF of the standard normal. The true loading parameters are set as $\beta_1 = \frac{1}{\sqrt{14}}(2,1,3)^{\mathrm{T}}$, $\beta_1 = \frac{1}{\sqrt{14}}(3,2,1)^{\mathrm{T}}$, $\beta_1 = \frac{1}{\sqrt{14}}(2,3,1)^{\mathrm{T}}$. Set

$$m_l\left(u_l\right) = m_l^*\left(u_l\right) - E\left\{m_l^*\left(u_l\right)\right\}, \quad l = 1,2,3$$

where $m_1^*\left(u_1\right) = 10\exp\left(5u_1\right)/\left\{1+\exp\left(5u_1\right)\right\}$, $m_2^*\left(u_2\right) = 5\sin\left(\pi u_2\right)$, and $m_3^*\left(u_3\right) = 3\left\{\sin\left(\pi u_3\right) + \cos\left(2\pi u_3 - 4\pi/3\right)\right\}$. Finally, Y_i, $1 \le i \le n$, are generated from the VICM (3-1), where $\beta = \left(\beta_1^{\mathrm{T}}, \beta_2^{\mathrm{T}}, \beta_3^{\mathrm{T}}\right)^{\mathrm{T}}$, and errors ε_i follow $N\left(0, \sigma^2\left(Z_i, X_i\right)\right)$ with $\sigma^2\left(Z_i, X_i\right) = \left\{100 - m\left(Z_i, X_i, \beta\right)\right\}/\left\{100 + m\left(Z_i, X_i, \beta\right)\right\}$.

Although the estimation process of the varying index coefficient model is introduced in Section 2.2, it is still difficult to directly estimate (2.12). Therefore, an iterative calculation algorithm is needed to estimate the unknown parameters and the unknown function coefficients. The specific algorithm is divided into the following two steps:

Step 1. The initial value $\left(\beta_1, \beta_2, \beta_3\right)$ of β is obtained in the following four steps:

1) Assuming that the unknown function m_l is a linear function, then $m\left(Z_i, X_i, \beta\right) = \sum_{i=1}^{d} a_l + b_l\left(Z_i^{\mathrm{T}}\beta_l\right)X_{il}$.

2) The estimated value $\left(\hat{a}_l, \hat{v}_l\right)$ of $\left(a_l, v_l\right)$ is estimated by minimizing $\sum_{i=1}^{n}\left\{Y_i - \sum_{l=1}^{d}\left(a_l + v_l^{\mathrm{T}}Z_i X_{il}\right)\right\}^2$, and thus the expression $\hat{\beta}_l^0 = \left(\hat{v}_l/\left\|\hat{v}_l\right\|_2\right)\mathrm{sgn}\left(\hat{v}_{1l}\right)$ is obtained, where $\hat{v}_{1l}$ is a part of $\hat{v}_l$.

3) Let $\hat{U}_l = Z_i^{\mathrm{T}}\hat{\beta}_l^0$, then obtain the initial unknown function $\hat{m}_l^{ini}(\cdot)$ from the varying coefficient model $Y = \sum_{l=1}^{d} m_l\left(\hat{U}_l\right)X_l + \varepsilon$.

4) Obtain β_1^{ini} by minimizing $2^{-1}\sum_{i=1}^{n}\left\{Y_i - \sum_{l=1}^{n}\hat{m}_l^{ini}\left(Z_i^{\mathrm{T}}\beta_l\right)X_{il}\right\}^2$, *i.e.* the initial value.

Step 2. Iterative calculations are performed by the asymptotic properties of the large sample parameter estimates and the theorems given by Ma and Song (2015)

[16]. Under certain assumptions, the estimated parameters satisfy the following asymptotic properties:

$$\sqrt{n}\left(\hat{\beta}_{-1}-\beta_{-1}^{0}\right)=\left\{n^{-1}\sum_{i=1}^{n}\Phi\left(X_i,Z_i,\beta^{0}\right)^{\otimes 2}\right\}^{-1}\times\left\{n^{-1/2}\sum_{i=1}^{n}\left(Y_i-m\left(Z_i,X_i\right)\right)\Phi\left(X_i,Z_i,\beta^{0}\right)\right\}+o_p\left(1\right) \tag{3.2}$$

where $\Phi\left(X_i,Z_i,\beta^{0}\right)=\left[\left\{\dot{m}_l\left(U_l\left(\beta_l^{0}\right),\beta_l^{0}\right)X_l J_l^{\mathrm{T}}\tilde{Z}\right\}^{\mathrm{T}},1\le l\le d\right]$, and $\tilde{Z}=Z-P\left(Z\right)$ in the above expression. Here $\tilde{Z}$ can be estimated by $\tilde{Z}=Z-P_n\left(Z\right)$, where

$$P_n\left(Z_k\right)=\sum_{l=1}^{d}\hat{g}_{1J}^{0}\left(U_l\left(\hat{\beta}\right),\hat{\beta}\right)X_l. \tag{3.3}$$

The estimation procedure of $\hat{g}_{1J}^{0}\left(\cdot,\hat{\beta}\right)$ in Equation (3.3) is similar to the estimation of the unknown function $\hat{m}_l\left(\cdot,\hat{\beta}\right)$, except that the response variable Y is replaced by Z_k in the iterative estimation process. According to the asymptotic properties (3.2) we can get an equation and use this equation for iterative calculations. The iteration stops when the absolute difference (dif) from the last calculated unknown parameter is less than 10^{-4} or the iteration number (iter) is greater than or equal to 100.

According to the idea of the above specific algorithm, we use R (64-bit) to write four functions such as Design matrix, transform, Jac, vicmest. Among them, vicmest is the main program for estimating unknown parameters, and the other three functions are intermediate conversion functions. First, we calculate the initial value β_1,β_2,β_3 of β through the first step. The results are shown in **Table 1**. It can be seen from **Table 1** that the initial value calculated by Step 1 is consistent with the trend of the actual value, but the deviation from the actual value is still large. Therefore, it is necessary to further calculate the estimated value by the second step. At this point, by running the four programs such as vicmest, the result of stopping the main program after 64 iterations is finally obtained, and $dif=8.45267\times10^{-6}$ at this time. The specific calculation results of the estimated values $\hat{\beta}$ of β and their deviations are shown in **Table 2**. It can be seen from **Table 2** that the value of a estimated by the profile spline modal estimator (PSME) is better, the deviation from the actual value (Bias) is smaller, and the mean deviation is less than 5%.

Table 1. The initial values of β calculated by step 1.

Initial value	β_1	β_2	β_3
1	0.715	0.951	0.722
2	0.345	0.287	0.897
3	0.933	0.747	0.311

Table 2. The estimated value of β and its deviation from the true value.

$n=200$	β_{11}	β_{12}	β_{13}	β_{21}	β_{22}	β_{23}	β_{31}	β_{32}	β_{33}
True	0.534	0.267	0.801	0.801	0.534	0.267	0.534	0.801	0.267
$\hat{\beta}$	0.537	0.316	0.782	0.798	0.538	0.271	0.461	0.833	0.305
Bias	0.003	0.049	−0.019	−0.003	0.003	0.004	−0.074	0.032	0.038

From the main program vicmest, not only can the estimated value $\hat{\beta}$ be obtained, but also can we obtain gamm0, which is the coefficient after the expansion of the B-spline basis function. Bring the calculated $\hat{\beta}$ and the coefficient gamm0 into the Formula (2.10), get the value of the unknown function $\hat{m}_l(\cdot,\beta)$ and the predicted value of the response variable Y. The results of the gamm0 coefficient are shown in **Table 3**. An estimate of $\hat{m}_l(\cdot,\beta)$ can be seen from **Figure 1**, where the red curve represents the estimate and the black curve represents the actual value. It can be seen intuitively from **Figure 1** that the fitting effect of the B-spline expansion is very good, not only the general trend of the unknown non-parametric function is well maintained, but also the accuracy of the estimation is relatively high. As shown in **Table 4**, we calculate the root mean square error of the coefficient of $\hat{m}_l(\cdot,\beta)$ by further calculation. It can be seen from **Table 4** that the unknown function has a small deviation, and the RMSE is less than 0.28, which indicates that the estimation effect is better. Moreover, Y can be calculated after obtaining the estimated values $\hat{\beta}$ and $\hat{m}_l(\cdot,\beta)$. Finally, the variance of error of the model (3.1) is calculated to be 5.777.

3.2. Results in High-Dimensional Case

In this section, we numerically simulate the variance estimation of the varying index coefficient model in high-dimensional conditions.

The profile spline modal estimator (PSME) shows good estimation variance under low-dimensional data settings. However, in the case of high-dimensional data, it will fall into the dimension curse, and the deviation of the estimated variance will increase as the dimension increases. The re-adjustment cross-validation method proposed by Fan *et al.* (2012) [19] can be considered as an effective way to overcome the dimension curse in high-dimensional problems through theoretical proof and data simulation test. Naturally, this paper applies the re-adjusted cross-validation method (RCV) to the high-dimensional varying index coefficient model for the first time. There are two types of covariates in the varying index coefficient model (2.2). The first type is Z. If Z is from a single variable, its relationship with the covariate X is more difficult to detect. The second type of variable is the covariate X. What we are concerned about is the estimation of the variance of the varying index coefficient model when the first type of covariate Z is high-dimensional.

We first perform data simulation. The setting of the real model is the same as model (3.1), with only the dimension of the covariant Z changed, that is, the first

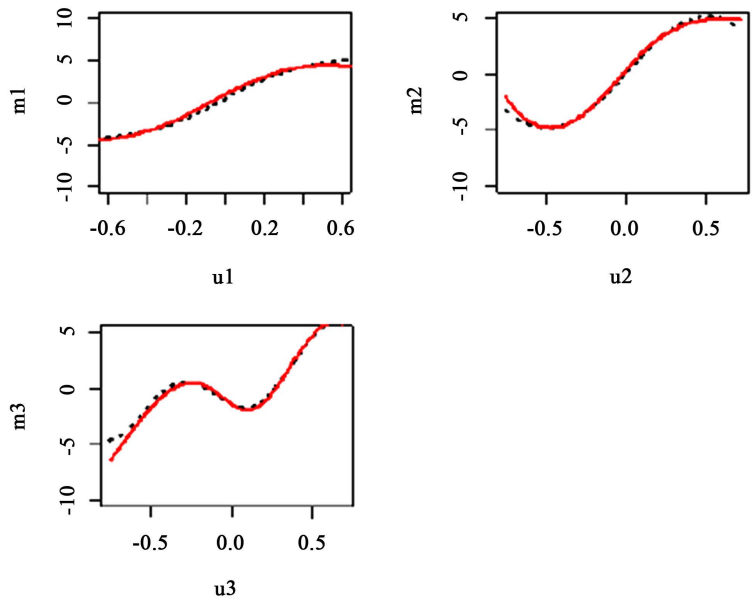

Figure 1. Plots of the estimated nonparametric curves.

Table 3. The coefficient of the B-spline basis function.

	B_{i1}	B_{i2}	B_{i3}	B_{i4}	B_{i5}	B_{i6}
B_{1i}	−4.444	−4.808	−3.0887	3.912	5.004	3.973
B_{2i}	−2.083	−6.179	−4.543	4.683	5.185	4.876
B_{3i}	−6.294	−2.710	3.652	−5.970	7.624	5.766

Table 4. Root mean square error (RMSE) of $\hat{m}_l(\cdot,\beta)$.

	$\hat{m}_1(\cdot,\beta)$	$\hat{m}_2(\cdot,\beta)$	$\hat{m}_3(\cdot,\beta)$
RMSE	0.349	0.222	0.258

type of covariate Z is set to a high dimensional variable. Where $Z_{ik}=\Phi\left(Z_{ik}^{*}\right)-0.5$, $k=1,2,\cdots,d$, that is, Z is a d-dimensional covariate. The setting of $m_l(u_l)$ and the error term is also the same as model (3.1).

In the case of high dimensional data, the independent variables are often highly correlated. However, not all independent variables are related to the dependent variable Y. In fact, only a small number of covariates are associated with the dependent variable Y. For the selection of such high-dimensional variables, Fan *et al.* (2008) [10] proposed SIS method with Sure Screening properties based on the relevant criteria, which can first reduce the dimension d to a relatively small number. Therefore, all important variables could be filtered into the model. So that lower-dimensional model selection methods such as SCAD, Dangit selector, LASSO, or adaptive LASSO could be used. With lower-dimensional model selection method, some smaller coefficients can be compressed to zero, thereby removing the extraneous variables that are filtered by the SIS method.

The idea of SIS makes high-dimensional model selection possible, greatly speeding up the selection of variables, and making model selection problems efficient and modular. The SIS variable selection method can be used in conjunction with any model selection technique. Fan *et al*. (2010) [20] apply the SIS method to the Cox proportional hazard model. The Cox proportional hazard model is similar to the varying index coefficient model mentioned in this paper. They are all nonlinear models and both have the need to estimate the coefficients of the nonparametric function and its parameter parts. Therefore, we believe that in the case of high-dimensional data, it is feasible to use the SIS method to make the first variable selection of the varying index coefficient model.

We use the SIS method proposed by Fan *et al*. (2008) [10] to select variables. The number of variables selected is tentatively 20. The calculation process is simulated using R software. We have written VicmRCV and the vicmest function for the estimation of the RCV process. The data simulation process was repeated 100 times, and a box plot of the variance as shown in **Figure 2** was obtained. In the figure, naïve represents a simple two-stage approach, while rcv represents a re-adjusted cross-validation method.

It can be seen from **Figure 2** that in the d dimension (the dimension of the variable Z, is as high as 100), the sample size is only 200, the variance of the re-adjusted cross-validation (RCV) two-stage method is better than the simple two-stage method. However, the calculated error variance value is large. To some extent, the estimation method of the estimated varying index coefficient model mentioned in Section 3.1 of this paper is not accurate enough and not robust enough as the estimated error variance value is large.

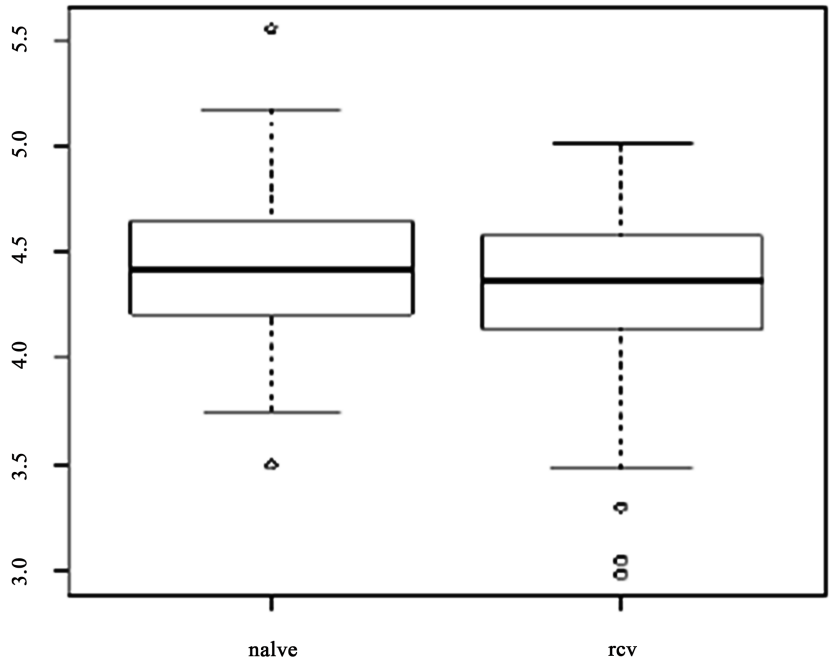

Figure 2. Box plot of the variance calculated by using the simple two-stage method and the RCV method.

As shown in Table 5, changing the values of p and n gives more simulation results. Table 5 compares the normal two-stage method (Naive-SIS) with the RCV two-stage method (RCV-SIS) at $n=100$, $d=50,100,500$. By comparing the root mean square error estimated by the two estimation methods, we find that the mean square error (MSE) of the RCV two-stage estimation is smaller in each dimension than the MSE estimated by the ordinary two-stage method. That is, the model estimated by the RCV method is more accurate. But from Table 5, we can also find other laws. Conventionally, as the dimension p increases, the estimated accuracy decreases, which results in the root mean square error becomes larger. However, from the results of Table 5, this law is completely inapplicable in the ordinary two-stage method. When the dimension comes to maximum ($d=500$), the root mean square error is the smallest, and its value is 6.692. When $d=100$, the MSE is the largest with a value of 8.273. In conclusion, the order is disorganized, and the mean square error does not become larger as the dimension becomes larger in general cases.

In fact, it is not difficult to explain this phenomenon because in the variable selection phase, for the SIS method, we select the variables with the co-correlation coefficients ranked in the top twenty (descending order). Since the fixed value 20 is small relative to the covariate, the probability of selecting all important variables is relatively low. From the data in the RCV-SIS column in Table 5, it can be seen that the SIS method is much more stable after combining RCV. At d = 50, the estimated MSE is the smallest with value of 4.838. In the case of three different dimensions, the error estimated by the RCV method is smaller than the mean square error estimated by the ordinary two-stage method.

4. Real Data Analysis

In this section, we will use the data collected by the Mayo Clinic. These data were obtained from trials conducted by the Mayo Clinic in primary biliary cirrhosis (PBC) from 1974 to 1984. Specific data can be found in the R language Survival package. The dataset included 424 PBC patients who were referred to the Mayo Clinic during the decade between 1974 and 1984. The data met the randomized placebo-based eligibility criteria.

In the data set, the first 312 patients participated in the randomized trial while the other 112 patients did not participate in the clinical trial, but agreed to record the basic measurements and follow the medical recommendations. Six of the above samples lost follow-up shortly after diagnosis. Thus there are 106 cases and 312 random participants. We preprocessed the data set via R software.

Table 5. Mean Square Error (MSE) for two different estimation methods at $n=100$.

	Naïve-SIS	RCV-SIS
$d=50$	8.171	4.838
$d=100$	8.273	5.199
$d=500$	6.692	5.043

We first remove some samples with missing values. The number of samples that were eventually brought into the calculation after deletion was 276. The specific variables are described in Table 6.

As can be seen from Table 6, the response variable Y is the survival time of the patient. Since the difference among the response variables Y is large, we logarithmically convert the time Y to reduce the error. There are three covariates of X, which are the patient's state X_1 (Status), the patient's age X_2 (Age), and the patient's gender X_3 (Sex). Here we need to explain why we want to add gender variables. The gender variable was added because Huang Siyu (1985) [21] found that the incidence of men with primary biliary cirrhosis (PBC) was much lower than that of women. Therefore, we can know that gender has a great relationship with PBC. Another type of covariate Z has a total of 15 variables including albumin (albumin), alkaline phosphatase (alk.phos), triglyceride (Trig), and platelet count (Platelet). Therefore, the varying index coefficient model constructed in this section is as follows:

$$Ln\left(Y_i\right) = m\left(Z, X, \beta\right) + \varepsilon_I = \sum_{l}^{3} m_l \left(\sum_{d}^{15} Z_{ld} * \beta_{ld} \right) X_{il} + \varepsilon_i \,. \tag{4.1}$$

Table 6. Interpretation of experimental variables in primary biliary cirrhosis (PBC).

Serial number	Variables	Description
1	Time (Y)	number of days between registration and the earlier of death
2	Status ($X1$)	status at end point, 0/1/2 for censored, transplant, dead
3	Age ($X2$)	in years
4	Sex ($X3$)	m/f
5	Trt ($Z1$)	1/2/NA for D-penicillmain, placebo, not randomised
6	Ascites ($Z2$)	presence of ascites
7	Hepato ($Z3$)	presence of hepatomegaly or enlarged liver
8	Spiders ($Z4$)	blood vessel malformations in the skin
9	Edema ($Z5$)	0 no edema, 0.5 untreated or successfully treated 1 edema despite diuretic therapy
10	Bili ($Z6$)	serum bilirubin (mg/dl)
11	Chol ($Z7$)	serum cholesterol (mg/dl)
12	Albumin ($Z8$)	serum albumin (g/dl)
13	Copper ($Z9$)	urine copper (ug/day)
14	alk.phos ($Z10$)	alkaline phosphotase (U/liter)
15	Ast ($Z11$)	aspartate aminotransferase, once called SGOT (U/ml)
16	Trig ($Z12$)	triglycerides (mg/dl)
17	Platelet ($Z13$)	platelet count
18	Protime ($Z14$)	standardised blood clotting time
19	Stage ($Z15$)	histologic stage of disease (needs biopsy)

Since the covariate Z has different physical meanings and different dimensions, it is meaningless to simulate the model at this time. So we need to eliminate the effects of different dimensions of the data through transformation. Therefore, these 15 variables should be standardized before the specific calculation, which is Z-Score standardization.

Through previous studies, we have roughly learned that variables such as serum bilirubin content ($Z6$), albumin content ($Z8$), urinary copper content ($Z9$), alkaline phosphatase content ($Z10$), prothrombin time ($Z14$) have a strong relationship with the response variable Y. We first use the SIS method to select the variables with the first 8 covariate correlations, and then use the simple two-stage method and the re-adjusted cross-validation (RCV) two-stage method to estimate the coefficient β of the covariate Z and the model variance. The results are shown in Table 7.

As can be seen from Table 7, when using SIS for variable selection, the important variables such as $Z6$ and $Z9$ are selected three times. Important variables have strong correlations in theoretical analysis. From this aspect, it can be seen that the SIS variable selection method can select all important variables with a high probability to a certain extent. The RCV can repeatedly select variables by selecting the first missing variable or deleting the extra variable that was selected for the first time. The model variance estimated by the RCV-SIS two-stage method is significantly better than the N-SIS simple two-stage method. In the high-dimensional case, the re-adjusted cross-validation method (RCV) has a better performance in the varying index coefficient model. The root mean square error and the resulting variance are smaller than the simple two-stage estimate. Therefore, the RCV-SIS two-stage method is more accurate in predicting the survival time of patients, and can provide more reasonable guidance and advice for follow-up medical treatments.

5. Discussion

In this paper, we study a new class of semiparametric regression models: varying index coefficient models. The estimation of the unknown coefficient β is estimated by the profile spline modal estimator method (PSME), while the unknown non-parametric function part is expanded with the B-spline. After studying the gradual nature of the coefficients, we estimate the coefficient β

Table 7. PBC dataset estimation results.

	N-SIS	RCV-SIS
AMS (variables selected)	$Z2$, $Z4$, $Z5$, $Z6$, $Z8$, $Z9$, $Z12$, $Z15$	The first stage: $Z4$, $Z5$, $Z6$, $Z7$, $Z9$, $Z11$, $Z12$, $Z15$ The second stage: $Z2$, $Z5$, $Z6$, $Z8$, $Z9$, $Z12$, $Z14$, $Z15$
MSE	2.184	1.661
VARIANCE	2.674	1.495

using an iterative method. With data simulation, we found that the estimated β of this method has a small deviation, and the unknown function part of the B-spline estimation has a good fitting effect as well. Finally, under the setting conditions of high-dimensional data, we carried out a two-stage RCV estimation of the varying index coefficient model. We find that the variance and mean square error estimated by the RCV method are superior to the simple two-stage method. In the final empirical phase, it was originally intended to model the PBC data using a survival model (semi-parametric varying coefficient additive risk model). However, through research literature, it is known that gender variables and state variables are closely related to the survival time of patients with primary biliary cirrhosis. The variable Z has a certain relationship with the three variables X (status, gender and age). Therefore, we used the varying index coefficient model to model the PBC data, and found that the variance and mean square error of the RCV method are better than the simple two-stage method.

Further researches for the proposed method are needed. Firstly, further effort to investigate the asymptotic properties of the proposed method needs to be done. Secondly, this paper only estimates the variance and mean square error of the varying index coefficient model, but lacks the research on the coefficient β and the estimation of the nonparametric function of the parameter part of the model. Therefore, we can study more robust estimation methods in the future. In addition, we can focus more on the asymptotic properties of the non-parametric part of the varying index coefficient model.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society* (*Series B*), **58**, 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[2] Fan, J. and Peng, H. (2004) On Nonconcave Penalized Likelihood with Diverging Number of Parameters. *Annals of Statistics*, **32**, 928-961. https://doi.org/10.1214/009053604000000256

[3] Zhao, P. and Yu, B. (2006) On Model Selection Consistency of Lasso. *Journal of Machine Learning Research*, **7**, 2541-2563.

[4] Bunea, F., Tsybakov, A. and Wegkamp, M. (2007) Sparsity Oracle Inequalities for the Lasso. *Electronic Journal of Statistics*, **64**, 330-332. https://doi.org/10.1214/07-EJS008

[5] Zhang, C.H. and Huang, J. (2008) The Sparsity and Bias of the Lasso Selection in High-Dimensional Linear Regression. *Annals of Statistics*, **36**, 1567-1594. https://doi.org/10.1214/07-AOS520

[6] Lv, J. and Fan, Y. (2009) A Unified Approach to Model Selection and Sparse Recovery Using Regularized Least Squares. *Annals of Statistics*, **37**, 3498-3528. https://doi.org/10.1214/09-AOS683

[7] Fan, J. and Lv, J. (2011) Nonconcave Penalized Likelihood with NP-Dimensionality. *Journal IEEE Transactions on Information Theory*, **57**, 5467-5484. https://doi.org/10.1109/TIT.2011.2158486

[8] Kim, Y., Choi, H. and Oh, H.S. (2008) Smoothly Clipped Absolute Deviation on High Dimensions. *Journal of the American Statistical Association*, **103**, 1665-1673. https://doi.org/10.1198/016214508000001066

[9] Candes, E. and Tao, T. (2005) The Danzig Selector: Statistical Estimation When *p* Is Much Larger than *n*. *Annals of Statistics*, **35**, 2313-2351. https://doi.org/10.1214/009053606000001523

[10] Fan, J. and Lv, J. (2008) Sure Independence Screening for Ultrahigh Dimensional Feature Space. *Journal of the Royal Statistical Society*, **70**, 849-911. https://doi.org/10.1111/j.1467-9868.2008.00674.x

[11] Fan, J., Samworth, R. and Wu, Y. (2009) Ultrahigh Dimensional Feature Selection: Beyond the Linear Model. *Journal of Machine Learning Research*, **10**, 2013-2038.

[12] Zhao, S.D., Cai, T.T. and Li, H. (2014) Variance Estimation in High-Dimensional Linear Models. *Biometrika*, **2**, 269-284. https://doi.org/10.1093/biomet/ast065

[13] Reid, S., Tibshirani, R. and Friedman, J. (2016) A Study of Error Variance Estimation in Lasso Regression. *Statistica Sinica*, **26**, 35-67. https://doi.org/10.5705/ss.2014.042

[14] Hastie, T. and Tibshirani, R. (1993) Varying-Coefficient Models. *Journal of the Royal Statistical Society* (*Series B*), **55**, 757-796. https://doi.org/10.1111/j.2517-6161.1993.tb01939.x

[15] Wong, H., Ip, W.C. and Zhang, R. (2008) Varying-Coefficient Single-Index Model. *Statistics*, **52**, 1458-1476. https://doi.org/10.1016/j.csda.2007.04.008

[16] Ma, S. and Song, X.K. (2015) Varying Index Coefficient Models. *Journal of the American Statistical Association*, **110**, 341-356. https://doi.org/10.1080/01621459.2014.903185

[17] Xue, L. and Liang, H. (2008) Polynomial Spline Estimation for a Generalized Additive Coefficient Model. *Scandinavian Journal of Statistics*, **37**, 26-46. https://doi.org/10.1111/j.1467-9469.2009.00655.x

[18] Lv, J., Yang, H. and Guo, C. (2016) Robust Estimation for Varying Index Coefficient Models. *Computational Statistics*, **31**, 1-37. https://doi.org/10.1007/s00180-015-0595-5

[19] Fan, J., Guo, S. and Hao, N. (2012) Variance Estimation Using Refitted Cross-Validation in Ultrahigh Dimensional Regression. *Journal of the Royal Statistical Society* (*Series B*), **74**, 37-65. https://doi.org/10.1111/j.1467-9868.2011.01005.x

[20] Fan, J., Yang, F. and Wu, Y. (2010) High-Dimensional Variable Selection for Cox's Proportional Hazards Model. *Statistics*, **105**, 205-217. https://doi.org/10.1214/10-IMSCOLL606

[21] Huang, S.Y. (1985) Is There a Difference in the Severity of Primary Biliary Liver Hardening? *International Journal of Digestive Diseases*, No. 3, 186-187.