Journal of Econometrics 189 (2015) 272–284

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Functional index coefficient models with variable selection[☆]

Zongwu Cai[a,b], Ted Juhl[a], Bingduo Yang[c,*]

[a] Department of Economics, University of Kansas, Lawrence, KS 66045, USA
[b] Wang Yanan Institute for Studies in Economics and Fujian Key Lab of Statistical Sciences, Xiamen University, Xiamen, Fujian 361005, China
[c] School of Finance, Jiangxi University of Finance and Economics, Nanchang, Jiangxi 330013, China

ARTICLE INFO

Article history:
Available online 19 March 2015

JEL classification:
C140
C580
C520

Keywords:
Functional index coefficient autoregressive model
Model selection
Oracle property
Penalty function
Smoothly clipped absolute deviation

ABSTRACT

We consider model (variable) selection in a semi-parametric time series model with functional coefficients. Variable selection in the semi-parametric model must account for the fact that the parametric part of the model is estimated at a faster convergence rate than the nonparametric component. Our variable selection procedures employ a smoothly clipped absolute deviation penalty function and consist of two steps. The first is to select covariates with functional coefficients that enter in the semi-parametric model. Then, we perform variable selection for variables with parametric coefficients. The asymptotic properties, such as consistency, sparsity and the oracle property of these two-step estimators are established. A Monte Carlo simulation study is conducted to examine the finite sample performance of the proposed estimators and variable selection procedures. Finally, an empirical example exploring the predictability of asset returns demonstrates the practical application of the proposed functional index coefficient autoregressive models and variable selection procedures.

1. Introduction

Linear time series models such as linear autoregressive moving average models, hereafter ARMA models (Box and Jenkins, 1970), were well developed in last century. However, linear ARMA models may not capture some important and potentially nonlinear features of the data in economics and finance. Many nonlinear time series models have been proposed. The early work includes the bilinear models (Granger and Andersen, 1978), the threshold autoregressive (TAR) models (Tong, 1990), the smooth transition AR (STAR) models (Chan and Tong, 1986; Teräsvirta, 1994) and Markov switching models (Hamilton, 1989), among others. One of the popular semiparametric models is the functional coefficient autoregressive (FAR) model, which was proposed by Chen and Tsay (1993) and extended by Cai et al. (2000b). The coefficients in FAR model are in unknown vector functional form depending on lagged terms, which satisfy

$$r_t = \sum_{j=1}^{p} g_j(\mathbf{r}^*_{t-1}) r_{t-j} + \varepsilon_t ,$$

where $\mathbf{r}^*_{t-1} = (r_{t-i_1}, \ldots, r_{t-i_d})^T$ with $1 \leq i_1 < i_2 < \cdots < t_d$ and $g_j(\cdot)$ is an unknown function in $\mathbb{R}^d$ for $1 \leq j \leq p$. The above FAR model covers several traditional varying coefficient models as a special case, such as the threshold autoregressive models in Tong (1990) and the STAR models in Chan and Tong (1986) and Teräsvirta (1994).

Due to the curse of dimensionality, Chen and Tsay (1993) just considered one single threshold variable case $\mathbf{r}^*_{t-1} = r_{t-k}$ for some k, and they proposed an arranged local regression to estimate the functional coefficients $\{g_j(\cdot)\}$ with an iterative algorithm. In fact, their method is similar to the local constant semiparametric estimator as pointed out by Cai et al. (2000b). For efficient estimation of the FAR model, the reader is referred to the papers by Cai et al. (2000a) and Fan and Zhang (1999).

To overcome the curse of dimensionality and incorporate more variables in the functional coefficients $\{g_j(\cdot)\}$, we assume that $\mathbf{r}^*_{t-1}$ is a linear combination of r_{t-i_k}'s, e.g. $\mathbf{r}^*_{t-1} = \beta^T \mathbf{r}_{t-1}$, where $\mathbf{r}_{t-1} = (r_{t-1}, \ldots, r_{t-d})^T$. We denote this model as the functional index

[☆] We thank the Guest Editors, Professors Shiqing Ling, Michael McAleer and Howell Tong, and the anonymous referees for their insightful comments that greatly improved our paper. Cai's research was supported, in part, by the National Nature Science Foundation of China grants #71131008 (Key Project), #70871003 and #70971113. Yang's research was supported by the National Nature Science Foundation of China grant #71401066 and Specialized Research Fund for the Doctoral Program of Higher Education #20130161120023.

[*] Corresponding author.
E-mail addresses: caiz@ku.edu (Z. Cai), juhl@ku.edu (T. Juhl), bdyang2006@gmail.com (B. Yang).

http://dx.doi.org/10.1016/j.jeconom.2015.03.022
0304-4076/© 2015 Elsevier B.V. All rights reserved.

coefficient autoregressive (FIAR) model satisfying

$$r_t = \sum_{j=1}^{p} g_j(\beta^T \mathbf{r}_{t-1}) r_{t-j} + \varepsilon_t,$$

where $g_j(\cdot)$ is an unknown function in $\mathbb{R}$ for $1 \le j \le p$. In fact, the above FIAR model can be regraded as a case of functional index coefficient models of Fan et al. (2003) with

$$y_i = \sum_{j=1}^{p} g_j(\beta^T Z_i) X_{ji} + \varepsilon_i \equiv g(\beta^T Z_i)^T X_i + \varepsilon_i, \quad 1 \le i \le n, \tag{1}$$

where y_i is a dependent variable, $X_i = (X_{1i}, X_{2i}, \dots, X_{pi})^T$ is a $p \times 1$ vector of covariates, Z_i is a $d \times 1$ vector of local variables, ε_i are independently identically distributed (i.i.d.) with mean 0 and standard deviation σ, $\beta \in \mathbb{R}^d$ is a $d \times 1$ vector of unknown parameters and $g(\cdot) = (g_1(\cdot), \dots, g_p(\cdot))^T$ is a $p \times 1$ vector of unknown functional coefficients. We assume that $\|\beta\| = 1$ and the first element of β is positive for identification, where $\|\cdot\|$ is the Euclidean norm (L_2-norm). Note that both X_i and Z_i can include the lagged variables of y_i. In particular, if $X_{1i} \equiv 1$, then, model (1) contains an intercept function term.

Xia and Li (1999) studied the asymptotic properties of model (1) under mixing conditions when the index part of above model is not constrained to be a linear combination of Z_i. However, due to the efficiency of estimation and the accuracy of prediction, it is of importance to select variables in both Z_i and X_i, and to potentially exclude variables in Eq. (1). Fan et al. (2003) provided algorithms to estimate local parameters β and functional coefficients $g(\cdot)$. Meanwhile, they deleted the least significant variables in a given model according to t-value, and selected the best model in terms of the Akaike information criterion (AIC) of Akaike (1973) in multiple steps. However, as mentioned in Fan and Li (2001), this stepwise deletion procedure may suffer stochastic errors inherited in the multiple stages. Meanwhile, there is no theory on this variable selection procedure and the authors did not mention how to select the regressors X_i. These selection issues motivate us to consider variable selection on both local variables Z_i and covariates X_i in model (1).

The FIAR model reduces the curse of dimensionality since each of the nonparametric functions has only one argument. However, there still remain potential areas of dimension reduction. First, there are several nonparametric functions in the $p \times 1$ vector $g(\beta^\top Z)$. In addition, the vector Z is d-dimensional. Hence, by using model selection methods, there is potential to find a more parsimonious model that effectively captures the features of our data. Variable selection methods and their algorithms can be traced back to four decades ago. Pioneering contributions include the AIC and the Bayesian information criterion (BIC) of Schwarz (1978). Various shrinkage type methods have been developed recently, including but not limited to the nonnegative garrotte of Breiman (1995), bridge regression of Fu (1998), the least absolute shrinkage and selection operator (LASSO) of Tibshirani (1996), the smoothly clipped absolute deviation of Fan and Li (2001), the adaptive LASSO of Zou (2006), and so on. The reader is referred to the review paper by Fan and Lv (2010) for details. Here, we recommend the smoothly clipped absolute deviation (SCAD) penalty function of Fan and Li (2001) since it merits three properties of unbiasedness, sparsity and continuity. Furthermore, it has the oracle property; namely, the resulting procedures perform as well as those that correspond to the case when the true model is known in advance.

The shrinkage method has been successfully extended to semiparametric models; for example, variable selection in partially linear models in Liang and Li (2009), partially linear models in longitudinal data in Fan and Li (2004), single-index models in Kong and Xia (2007), semiparametric regression models in Brent et al. (2008) and Li and Liang (2008), varying coefficient partially linear models with errors-in-variables in Zhao and Xue (2010), and partially linear single-index models in Liang et al. (2010), and the references therein.

However, the aforementioned papers focused mainly on the variable selection with parametric coefficients. Also, the shrinkage method was extended to select significant variables with functional coefficients. Lin and Zhang (2006) proposed component selection and smoothing operator (COSSO) for model selection and model fitting in multivariate nonparametric regression models in the framework of smoothing spline analysis of variance. Meanwhile, they extended the COSSO to the exponential families (Zhang and Lin, 2006). Wang et al. (2008) proposed the variable selection procedures with basis function approximations and SCAD, which is similar to the COSSO, and they argued that their procedures can select significant variables with time-varying effect and estimate the nonzero smooth coefficient functions simultaneously. Huang et al. (2010) proposed to use the adaptive group LASSO for variable selection in nonparametric additive models based on a spline approximation, in which the number of variables and additive components may be larger than the sample size. By adopting the idea of the grouping method in Yuan and Lin (2006), Wang and Xia (2009) used kernel LASSO to apply shrinkage to functional coefficients in the varying coefficient models. Their pure nonparametric shrinkage procedure is different from approaches of using spline and basis functions (Lin and Zhang, 2006; Wang et al., 2008; Huang et al., 2010). For a comprehensive survey paper of variable selection in nonparametric and semiparametric regression models via shrinkage, the reader is referred to the paper by Su and Zhang (2013).

Almost all the variable selection procedures mentioned above are based on the assumption that the observations are independent and identically distributed (i.i.d.). To the best of our knowledge, there are few papers to consider variable selections under non i.i.d. settings. It might not be appropriate if it is applied to analyze financial and economic data directly, since most of the financial/economic data are weakly dependent. To address this issue, Wang et al. (2007) extended to the regression model with autoregressive errors via LASSO. In this paper, we consider variable selection in functional index coefficient models under very general dependence structure—the strong mixing context. Our variable selection procedures consist of two steps. The first is to select covariates with functional coefficients, and then we perform model selection for local variables with parametric coefficients.

The rest of this paper is organized as follows. In Section 2, we present the identification conditions for functional index coefficient models, our new two-step estimation procedures, and some properties of the SCAD penalty function and numerical implementations. In Section 3, we propose variable selection procedures for both covariates with functional coefficients and local variables with parametric coefficients. We then establish the consistency, the sparsity and the oracle property of all the proposed estimators. A simple bandwidth selection method is also discussed in the same section. Monte Carlo simulation results for the proposed two-step procedures are reported in Section 4. An empirical example of applying the functional index coefficient autoregressive model and its variable selection procedures is extensively studied in Section 5. Finally, the concluding remarks are given in Section 6 and all the regularity conditions and technical proofs are gathered in the Appendix.

2. Identification, estimation and penalty function

2.1. Identification

The identification problem in single index model was first investigated by Ichimura (1993), and extensively studied by

Li and Jeffrey (2007) and Horowitz (2009). Meanwhile, partial conditions for identification in functional index coefficient models were showed in Fan et al. (2003). Here we present the conditions for identification below.

Theorem 1 (*Identification in Functional Index Coefficient Models*). *Assume that dependent variable Y is generated by Eq. (1), X is a p-dimensional vector of covariates and Z is a d-dimensional vector of local variables. β is a d-dimensional vector of unknown parameters and $g(\cdot)$ is a p-dimensional vector of unknown functional coefficients. Then, β and $g(\cdot)$ are identified if the following conditions hold:*

Assumption I. I1. The vector functions $g(\cdot)$ are continuous and not constant everywhere.

I2. The components of Z are continuously distributed random variables.

I3. There exists no perfect multi-collinearity within each components of Z and none of the components of Z is constant.

I4. There exists no perfect multi-collinearity within each components of X.

I5. The first element of β is positive and $\|\beta\| = 1$, where $\|\cdot\|$ is the standard Euclidean norm.

I6. When $X = Z$, $E(Y|X,Z)$ becomes to $E(Y|X)$ and it cannot be expressed in the form as $E(Y|X) = \alpha^T X\beta^T X + \gamma^T X + c$, where $\alpha, \gamma \in \mathbb{R}^d$ and $c \in R$ are constant, and α and β are not parallel to each other.

Remark 1. Assumption I1 is a mild condition since continuous and bounded functions are commonly assumed in nonparametric estimation, and it is obvious that β cannot be identified if any element of $g(\cdot)$ is a constant. We can relax Assumption I2 with some components of Z being discrete random variables, however, two more conditions should be imposed, see Ichimura (1993) and Horowitz (2009) in detail. The perfect multi-collinearity problem in Assumptions I3 and I4 is similar to those in the classical linear models. In fact, it would be hard to get accurate estimates if high correlation of components exists in either Z or X. Meanwhile, it is not identified if any component of Z is constant. For example, if Z_1=1, $E(Y|X,Z) = g^T(\beta_1+\beta_2 Z_2+\cdots+\beta_d Z_d)X = f^T(\beta_2 Z_2+\cdots+\beta_d Z_d)X$. An alternative of Assumption I5 is to let the first coefficient be 1, i.e. $\beta_1 = 1$. However, it is infeasible to implement variable selection procedures, since we do not have any prior information that whether the coefficient β_1 of Z_1 is zero or not. Assumption I6 can be found in the paper by Fan et al. (2003).

2.2. Estimation procedures

Model (1) can be regarded as a semiparametric model. Therefore, to estimate both functions $g(\cdot)$ and parameters β, it is common to use a two-stage approach. To estimate $g(\cdot)$, one needs an initial estimator of $\hat\beta$ which might have little effect on the final estimation of $g(\cdot)$ if the sample size n is large enough, due to the fact that the convergence rate of the parametric estimator $\hat\beta$ is faster than the nonparametric estimator $\hat g(\cdot)$. Here, we propose variable selection and estimation in two steps:

Step One: Given an initial estimator $\hat\beta$ such that $\|\hat\beta - \beta\| = O_p(1/\sqrt{n})$, minimize the penalized local least squares $Q(\hat g, \hat\beta, h)$ to obtain $\hat g(\cdot)$, where

$$Q(\hat g, \hat\beta, h) = \sum_{j=1}^{n}\sum_{i=1}^{n}\left\{y_i - \hat g^T\left(\hat\beta^T Z_j\right)X_i\right\}^2 K_h\left(\hat\beta^T Z_i - \hat\beta^T Z_j\right) + n\sum_{k=1}^{p} P_{\lambda_n}\left(\|\hat g_{\cdot k}\|\right), \qquad (2)$$

with $K(\cdot)$ being a kernel function, $K_h(z) = K(z/h)/h$ and $P_{\lambda_n}(\cdot)$ being a penalty function. $\{\lambda_1,\ldots,\lambda_p\}$ are tuning parameters and $\hat g_{\cdot k} = [\hat g_k(\hat\beta^T Z_1),\ldots,\hat g_k(\hat\beta^T Z_n)]^T$ is the estimates of kth functional coefficient at corresponding sample points. As recommended, an initial estimator $\hat\beta$ can be obtained by various algorithms such as the method in Fan et al. (2003), or average derivative estimators such as Newey and Stoker (1993). As long as the initial estimator satisfies $\|\hat\beta - \beta\| = O_p(1/\sqrt{n})$, as expected, the parametric estimator $\hat\beta$ has little effect on the shrinkage estimation of functional coefficients $\hat g(\cdot)$ in the above equation if sample size n is large. We choose penalty term $P_{\lambda_n}(\cdot)$ as the SCAD function, which is described in Section 2.3, and the L_2 functional norm $\|\hat g_{\cdot k}\| = \sqrt{\hat g_k^2(\hat\beta^T Z_1)+\cdots+\hat g_k^2(\hat\beta^T Z_n)}$ has the same definition of standard Euclidean norm. The purpose of using the penalized locally weighted least squares is to select significant covariates X_i in model (1).

Note that when the penalty term $P_{\lambda_n}(z) = \lambda_n|z|$, the penalized local least squares becomes the Lasso type, so that the above object function in Eq. (2) is reduced to the case in the paper by Wang and Xia (2009).

Step Two: Given the estimator of function $\hat g(\cdot)$, minimize the penalized global least squares $Q(\beta, \hat g)$, where

$$Q(\beta, \hat g) = \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \hat g^T(\beta^T Z_i)X_i\right)^2 + n\sum_{k=1}^{d}\Psi_{\zeta_n}(|\beta_k|) \qquad (3)$$

with $\Psi(\cdot)$ being a penalty function. $\{\zeta_1,\ldots,\zeta_d\}$ are tuning parameters and $|\beta_k|$ takes the absolute value of β_k.

Clearly, the above general setting may cover several other existing variable selection procedures as a special case. For example, when $p = 1$ and the regressor $X = 1$, the above procedure becomes variable selection for the single-index model in Kong and Xia (2007), which provided an alternative variable selection method called separated cross validation. When $p = 2$ and the only regressor is market return, then the above model reduces to the case in the paper by Cai et al. (2014a) for an application in finance. In particular, they considered semiparametric estimates of time-varying betas and alpha in the conditional capital asset pricing model with variable selection. Furthermore, the model includes a special case of variable selection in partially linear single-index models as addressed in Liang et al. (2010), if only the first functional coefficient $g(\cdot)$ is nonlinear and all others are constant. Finally, it transforms to variable selection in semiparametric regression modeling by Li and Liang (2008), if the dimension of local variables $d = 1$ and some of the functional coefficients $g(\cdot)$ are constant and others are not.

2.3. Penalty function and implementation

As pointed out by Fan and Li (2001), a good penalty function should enjoy the following three desirable properties, e.g., unbiasedness for the large true unknown estimator, sparsity that can set small estimator to be zero automatically, and continuity of the resulting estimator to avoid instability in model prediction.

To achieve all the aforementioned three properties, Fan and Li (2001) proposed the following so called SCAD penalty function,

$$P_\lambda(|\beta|) = \begin{cases} \lambda|\beta|, & |\beta| \le \lambda, \\ -(|\beta|^2 - 2a\lambda|\beta| + \lambda^2)/[2(a-1)], & \lambda < |\beta| \le a\lambda, \\ (a+1)\lambda^2/2, & |\beta| > a\lambda. \end{cases} \qquad (4)$$

The important property for the SCAD penalty function is that it has the following first derivative,

$$P'_{\lambda}(|\beta|) = \begin{cases} \lambda, & |\beta| \leq \lambda, \\ (a\lambda - |\beta|)/(a-1), & \lambda < |\beta| \leq a\lambda, \\ 0, & |\beta| > a\lambda, \end{cases}$$

$$\text{for some } a > 2 \tag{5}$$

so that it makes the computational implementation easily. It can be clearly seen that $P_{\lambda}(|\beta|)$ is not differentiable at 0 with respect to β. Thus, it is not easy to minimize the penalized least squares due to its singularity. To make implementation easily, Fan and Li (2001) suggested to approximate the penalty function by a quadratic function as

$$P_{\lambda}(|\beta_j|) \approx P_{\lambda}(|\beta_j^{(0)}|) + \frac{1}{2}\{P'_{\lambda}(|\beta_j^{(0)}|)/|\beta_j^{(0)}|\}(\beta_j^2 - \beta_j^{(0)2})$$

$$\text{for } \beta_j \approx \beta_j^{(0)}. \tag{6}$$

Alternatively, Zou and Li (2008) proposed local linear approximation for non-concave penalty functions as

$$P_{\lambda}(|\beta_j|) \approx P_{\lambda}(|\beta_j^{(0)}|) + P'_{\lambda}(|\beta_j^{(0)}|)(|\beta_j| - |\beta_j^{(0)}|) \quad \text{for } \beta_j \approx \beta_j^{(0)} \tag{7}$$

which can reduce the computational cost without losing any statistical efficiency. Meanwhile, some other algorithms such as minorize-maximize algorithm (Hunter and Li, 2005) are also proposed.

In view of (6), given a good initial value $\beta^{(0)}$ we can find the one-step estimator as follow

$$\beta^{(1)} = \operatorname{argmin}\left[\frac{1}{2}(\beta - \beta^{(0)})^T[-\nabla^2\ell(\beta^{(0)})](\beta - \beta^{(0)}) + n\sum_{k=1}^{d}\frac{P'_{\lambda}(|\beta_k^{(0)}|)}{2|\beta_k^{(0)}|}\beta_k^2\right], \tag{8}$$

where $\ell(\cdot)$ is a loss function and $\nabla^2\ell(\beta^{(0)}) = \partial^2\ell(\beta^{(0)})/\partial\beta\partial\beta^T$. As argued in Fan and Li (2001), there is no need to iterate until it converges as long as the initial estimator is reasonable. Also, the MLE estimator from the full model without penalty term can be regarded as the reasonable initial estimator. For using the local linear approximation in Zou and Li (2008) and Eq. (7), the sparse one-step estimator given in Eq. (8) becomes

$$\beta^{(1)} = \operatorname{argmin}\left[\frac{1}{2}(\beta - \beta^{(0)})^T[-\nabla^2\ell(\beta^{(0)})](\beta - \beta^{(0)}) + n\sum_{k=1}^{d}P'_{\lambda}(|\beta_k^{(0)}|)|\beta_k|\right]. \tag{9}$$

As demonstrated in Zou and Li (2008), this one step estimator is as efficient as the fully iterative estimator, provided that the initial estimator is good enough. For example, we let $\beta^{(0)}$ be the maximal likelihood estimator without the penalty term.

3. Large sample theory

3.1. Penalized nonparametric estimator for functional coefficients

Let $\{(X_i, Z_i, y_i)\}$ be a strictly stationary and strong mixing sequence, and $f(\cdot, \beta)$ be the density function of $\beta^T Z$, where β is an interior point of the compact set $\mathbb{B}$. Assume δ is a small positive constant and define $\mathcal{A}_z = \{Z : f(\beta^T Z, \beta) \geq \delta, \beta \in \mathbb{B}$, there exist a and b such that $\beta^T Z \in [a, b]\}$ as the domain of Z. Then, $\beta^T Z$ is bounded and the density of $f(\cdot, \beta)$ is bounded away from 0. Also, define the domain of bandwidth h, $\mathcal{H}_n = \{h$: there exist C_1 and C_2 such as $C_1 n^{-1/5} < h < C_2 n^{-1/5}\}$. For $Z \in \mathcal{A}_z$, $\beta \in \mathbb{B}$, and $h \in \mathcal{H}_n$, define a $n \times p$ matrix penalized estimator as

$$\hat{G}\left(\hat{\beta}\right) = \left[\hat{g}\left(\hat{\beta}^T Z_1\right), \ldots, \hat{g}\left(\hat{\beta}^T Z_n\right)\right]^T = \left[\hat{g}_{\cdot 1}, \ldots, \hat{g}_{\cdot p}\right],$$

where

$$\hat{g}\left(\hat{\beta}^T Z\right) = \left[\hat{g}_1\left(\hat{\beta}^T Z\right), \ldots, \hat{g}_p\left(\hat{\beta}^T Z\right)\right]^T \in \mathbb{R}^p,$$

and

$$\hat{g}_{\cdot k} = \left[\hat{g}_k\left(\hat{\beta}^T Z_1\right), \ldots, \hat{g}_k\left(\hat{\beta}^T Z_n\right)\right]^T \in \mathbb{R}^n.$$

Similarly, we define the true value $G_0(\beta)$, $g_0\left(\beta^T Z\right)$ and $g_{0\cdot k}$, respectively. Without loss of generality, we assume that the first p_0 functional coefficients are non-zero, and other $p - p_0$ functional coefficients are zero, i.e. $\|g_{\cdot k}\| \neq 0$ and $g_{\cdot k}$ are not constant everywhere for $1 \leq k \leq p_0$, $\|g_{\cdot k}\| = 0$ for $p_0 < k \leq p$. Let $\alpha_n = \max\{P'_{\lambda}(\|g_{\cdot k}\|) : 1 \leq k \leq p_0\}$. Then, by minimizing the penalized local least squares $Q(\hat{g}, \hat{\beta}, h)$ in Eq. (2), one can obtain the penalized local least squares estimator $\hat{g}(\cdot)$ for $g(\cdot)$.

To study the asymptotic distribution of the penalized local least squares estimator, we impose some technical conditions as follows.

Assumption A. A1. The vector functions $g(\cdot)$ have continuous second order derivatives with respect to the support of $\mathcal{A}_z$.

A2. For any $\beta \in \mathbb{B}$ and $Z \in \mathcal{A}_z$, the density function $f(\cdot, \beta)$ is continuous and there exists a small positive δ such that $f(\cdot, \beta) > \delta$.

A3. The kernel function $K(\cdot)$ is a bounded density with a bounded support region. Let $\mu_2 = \int v^2 K(v)dv$ and $\nu_0 = \int K^2(v)dv$.

A4. $\lim_{n\to\infty}\inf_{\theta\to 0^+} P'_{\lambda_n}(\theta)/\lambda_n > 0$, $n^{-1/10}\lambda_n \to 0$, $h \propto n^{-1/5}$ and $\|\hat{\beta} - \beta_0\| = O_p(1/\sqrt{n})$.

A5. Define $\Omega(z, \beta) = E\left(X_i X_i^T | \beta^T Z_j = z\right)$. Assume that $\Omega(\cdot)$ is nonsingular and has bounded second order derivative on $\mathcal{A}_z$.

A6. $\{(X_i, Z_i, y_i)\}$ is a strictly stationary and strongly mixing sequence with mixing coefficient satisfying $\alpha(m) = O(\rho^m)$ for some $0 < \rho < 1$.

A7. Assume that the conditional density of $f\left(z_i, z_s | z_j\right)$ is continuous and has bounded second order derivative.

A8. Assume that $\Omega\left(z_i, z_s, z_j\right) = E\left(X_i X_i^T X_s X_s^T | \beta^T Z_i = z_i, \beta^T Z_s = z_s, \beta^T Z_j = z_j\right)$ is continuous and has bounded second order derivative. Define $\Omega_1\left(z_i, z_s, z_j\right) = \partial\Omega\left(z_i, z_s, z_j\right)/\partial z_i$ and $\Omega_2\left(z_i, z_s, z_j\right) = \partial\Omega\left(z_i, z_s, z_j\right)/\partial z_s$.

Remark 2. The conditions in A2 imply that the distances between two ranked values $\beta^T Z_{(i)}$ are at most order of $O_p(\log n/n)$ (Janson, 1987). For any value $Z \in \mathcal{A}_z$, we can find a closest value $\beta^T Z_j$ to $\Lambda = \beta^T Z$ such that $|\beta^T Z_j - \Lambda| = O_p(\log n/n)$. With the conditions in A1, $\|g(\beta^T Z_j) - g(\Lambda)\| = O_p(\log n/n)$, which is smaller order of nonparametric convergence rate $n^{-2/5}$. This implies that we only need to estimate $\hat{g}(\beta^T Z_i)$ for $i = 1, 2, \ldots, n$ rather than $\hat{g}(\Lambda)$ for all values in the domain $\mathcal{A}_z$. For the detailed arguments, we refer to the paper by Wang and Xia (2009). Assumption A3 is a common assumption in nonparametric estimation. The assumption $\|\hat{\beta} - \beta_0\| = O_p(1/\sqrt{n})$ in A4 implies that the estimators of $\hat{\beta}$ have little effect in the estimation of $\hat{g}(\cdot)$ if the sample size n is large, since the convergence rate of the parametric estimators $\hat{\beta}$ is faster than the nonparametric function estimators $\hat{g}(\cdot)$. The assumptions in A5–A8 are very standard and used for the proof under mixing conditions; see Cai et al. (2000b). In particular, Assumptions in A6 are the common conditions with weekly dependent data. Most financial models satisfy these conditions, such as ARMA, ARCH and GARCH models; see Cai (2002).

Define the nonparametric estimator $\hat{g}(z,\hat{\beta}) \equiv [\hat{g}_a(z,\hat{\beta}), \hat{g}_b(z,\hat{\beta})]^T$ where $z = \hat{\beta}^T Z$, $\hat{g}_a(z,\hat{\beta}) = [\hat{g}_1(z,\hat{\beta}), \ldots, \hat{g}_{p_0}(z,\hat{\beta})]^T \in \mathbb{R}^{p_0}$ and $\hat{g}_b(z,\hat{\beta}) = [\hat{g}_{p_0+1}(z,\hat{\beta}), \ldots, \hat{g}_p(z,\hat{\beta})]^T \in \mathbb{R}^{p-p_0}$. Analogously, we denote the true value $g_0(z,\beta_0) \equiv [g_{0a}(z,\beta_0), g_{0b}(z,\beta_0)]^T$. The following theorem presents the asymptotic properties for the penalized nonparametric estimator $\hat{g}(z,\hat{\beta})$, including the oracle property, sparsity and asymptotic normality of the estimator $\hat{g}(z,\hat{\beta})$.

Theorem 2 (*Oracle Property*)**.** *Let* (X_i, Z_i) *be a strong mixing and strictly stationary sequence. Under Assumptions* A1–A8, $\|\hat{\beta}-\beta_0\| = O_p(n^{-1/2})$, $\lim_{n\to\infty}\inf_{\theta\to 0^+} P'_{\lambda_n}(\theta)/\lambda_n > 0$, $h \propto n^{-1/5}$ *and* $n^{-1/10}\lambda_n \to 0$ *as* $n \to \infty$, *then*

(a) Sparsity: $\sup_{Z\in\mathcal{A}_z} \|\hat{g}_k(z,\hat{\beta})\| = 0$, for all $p_0 < k \le p$.
(b) Asymptotic normality:

$$\sqrt{nh}\left(\hat{g}_a(z,\hat{\beta}) - g_{0a}(z,\beta_0) - h^2 B(z,\beta_0) + o_p(h^2)\right) \sim N(0, V(z,\beta_0)),$$

where $V(z,\beta_0) = \nu_0 M^{-1}(z,\beta_0)\sigma^2$, and

$$B(z,\beta_0) = \mu_2 M^{-1}(z,\beta_0)\dot{M}(z,\beta_0)\dot{g}(z,\beta_0) + \frac{1}{2}\mu_2\ddot{g}(z,\beta_0)$$

with $M(z,\beta_0) = f(z,\beta_0)\Omega(z,\beta_0)$, $\dot{M}(z,\beta_0) = \partial M(z,\beta_0)/\partial z$, $\dot{g}(z,\beta_0) = \partial g(z,\beta_0)/\partial z$ and $\ddot{g}(z,\beta_0) = \partial\dot{g}(z,\beta_0)/\partial z$.

Remark 3. The unpenalized estimator can be written as

$$\hat{g}_u(z,\beta) = \left[\sum_{i=1}^{n} X_{ia}X_{ia}^T K_h(\beta^T Z_i - z)\right]^{-1}\left[\sum_{i=1}^{n} X_{ia}Y_i^T K_h(\beta^T Z_i - z)\right].$$

Similar to the argument in the paper by Wang and Xia (2009), under regular conditions, one can show that $\sup_{Z\in\mathcal{A}_z}\|\hat{g}_a(z,\hat{\beta}) - \hat{g}_u(z,\hat{\beta})\| = o_p(n^{-2/5})$. It suggests that the difference between penalized estimator $\hat{g}_a(z,\hat{\beta})$ and unpenalized estimator $\hat{g}_u(z,\hat{\beta})$ is smaller order of optimal nonparametric convergence rate of $n^{-2/5}$. Thus, the penalized estimator $\hat{g}_a(z,\hat{\beta})$ merits the same large sample properties as the unpenalized estimator $\hat{g}_u(z,\hat{\beta})$, as the sample size n goes to infinity.

Sparsity is an important statistical property in high-dimensional statistics. By assuming that only a small subset of the variables are important for dependent variable, it can reduce complexity so that it improves interpretability and predictability of the model. The sparsity property from Theorem 2 demonstrates that our penalized model can estimate zero components of the true parameter vector exactly as zero with probability one as sample size goes to infinity.

3.2. Penalized estimator for parametric coefficients

To perform variable selection for variables with parametric coefficients, we should minimize the penalized least squares listed in Eq. (3). We assume the first d_1 coefficients of β are nonzero and all rest of parameters are zero. That is, $\beta_0 = (\beta_{10}^T, \beta_{20}^T)^T$, where all elements of β_{10} with dimension d_1 are nonzero and $d - d_1$ dimensional coefficients $\beta_{20} = 0$. Finally, define $V_n = \sum_{i=1}^{n}(Z_i - E(Z_i|\beta_0^T Z_i))\dot{g}^T(\beta_0^T Z_i)X_i\varepsilon_i$, where vector $\dot{g}(\cdot)$ is the first derivative of function $g(\cdot)$ vector, and ε_i is independent and identically distributed (i.i.d.) with mean 0 and standard deviation σ. Let $\widetilde{V}_0 = \frac{1}{n}\mathrm{Var}(V_n)/\sigma^2$, and define $\mathbf{e}$ be an asymptotically standard normal random d-dimensional vector such that $V_n = n^{1/2}\sigma\widetilde{V}_0^{1/2}\mathbf{e}$. $V_{1n} = \sum_{i=1}^{n}(Z_{1i} - E(Z_{1i}|\beta_{10}^T Z_{1i}))\dot{g}^T(\beta_{10}^T Z_{1i})X_i\varepsilon_{1i}$, where ε_{1i} is the same as ε_i since $\beta_{20} = 0$. Similarly, we define $\widetilde{V}_{10} = \frac{1}{n}\mathrm{Var}(V_{1n})/\sigma^2$ and $\mathbf{e}_1$ be an asymptotically standard normal random d_1-dimensional vector such that $V_{1n} = n^{1/2}\sigma\widetilde{V}_{10}^{1/2}\mathbf{e}_1$.

To study the asymptotic distribution of the penalized least squares estimator $\hat{\beta}$, we impose some technique conditions as below.

Assumption B. B1. The vector functions $g(\cdot)$ have continuous second order derivatives with respect to the support of $\mathcal{A}_z$.
B2. For any $\beta \in \mathbb{B}$ and $Z \in \mathcal{A}_z$, the density function $f(\cdot,\beta)$ is continuous and there exists a small positive δ such that $f(\cdot,\beta) > \delta$.
B3. The kernel function $K(\cdot)$ is a bounded density with a bounded support region. Let $\mu_2 = \int v^2 K(v)dv$ and $\nu_0 = \int K^2(v)dv$.
B4. $\lim_{n\to\infty}\inf_{\theta\to 0^+} P'_{\zeta_n}(\theta)/\zeta_n > 0$, $\zeta_n \to 0$, $\sqrt{n}\zeta_n \to \infty$ and $h \propto n^{-1/5}$.
B5. Same as Assumption A6.
B6. $E(\varepsilon_i|X_i, Z_i) = 0$, $E(\varepsilon_i^2|X_i, Z_i) = \sigma^2$, $E|X_i|^m < \infty$ and $E|y_i|^m < \infty$ for all $m > 0$.

Remark 4. The assumptions in B4 indicate the oracle property in Theorem 4. An alternative condition for bandwidth in Ichimura (1993) is $nh^8 \to 0$. However, the condition $nh^8 \to 0$ is still satisfied with our condition $h \propto n^{-1/5}$ in B4. For Assumption B6, it is not hard to extend to the heteroscedasticity case, $E(\varepsilon_i^2|X_i, Z_i) = \sigma^2(X_i, Z_i)$, and it requires some higher moment conditions of X_i and y_i so that Chebyshev inequality can be applied.

Now, we have the asymptotic properties for the penalized least squares estimator $\hat{\beta}$.

Theorem 3. *Let* $\{(X_i, Z_i, y_i)\}$ *be a strictly stationary and strong mixing sequence,* $a_n = \max\{\Psi'_{\zeta_n}(\beta_k) : \beta_k \neq 0\}$, *and* $\hat{\beta} = \mathrm{argmin}_{\beta\in\mathbb{B}}\, Q(\beta,\hat{g})$. *Under Assumptions* B1–B6 *and if* $\max\{\Psi''_{\zeta_n}(\beta_k) : \beta_k \neq 0\} \to 0$, *then the order of* $\|\hat{\beta}-\beta_0\|$ *is* $O_p(n^{-1/2} + a_n)$. *If the penalty function is SCAD function,* $a_n = 0$ *as sample size* $n \to \infty$, *and* $\|\hat{\beta}-\beta_0\| = O_p(n^{-1/2})$.

Theorem 4 (*Oracle Property*)**.** *Let* $\{(X_i, Z_i, y_i)\}$ *be a strictly stationary and strong mixing sequence. Under Assumptions B1–B6, by assuming* $\zeta_n \to 0$ *and* $\sqrt{n}\zeta_n \to \infty$ *as* $n \to \infty$, *then,*
(a) ***Sparsity:***

$$\hat{\beta}_2 = 0.$$

(b) ***Asymptotic normality:***

$$\sqrt{n}(\hat{\beta}_1 - \beta_{10}) \to N\left(0, \widetilde{V}_{10}^{-1}V_{10}\widetilde{V}_{10}^{-1}\right),$$

where $\widetilde{V}_{10}$ *is defined earlier and* $V_{10} = \Gamma(0) + 2\sum_{\ell=1}^{\infty}\Gamma(\ell)$ *with* $\Gamma(\ell) = \mathrm{Cov}(\Gamma_i, \Gamma_{i-\ell})$ *and* $\Gamma_i = \left(Z_{1i} - E(Z_{1i}|\beta_{10}^T Z_{1i})\right)\dot{g}^T(\beta_{10}^T Z_{1i}) X_i\varepsilon_{1i}$.

When the random variables $\{\Gamma_i\}_{i=1}^{\infty}$ are either i.i.d. or martingale difference sequence, V_{10} becomes $V_{10} = \Gamma(0) = \mathrm{Var}(\Gamma_i)$. Otherwise, the autocovariance function $\Gamma(\ell)$ may not be zero at least for some lag orders $\ell > 0$ due to the serial correlation. Theorem 4 shows that our variable selection procedures of minimizing penalized least squares enjoy the oracle property.

3.3. Choosing bandwidth and tuning parameters

To do the nonparametric estimation and variable selection simultaneously, we should choose suitable regularization parameters, bandwidth h for nonparametric estimator and λ's for penalty terms. For simplicity, we just consider global bandwidth selection rather than pointwise selection. Recent literature reveals that the

BIC-type selector identifies the true model consistently and the resulting estimator possesses the oracle property. In contrast, the AIC-type selector tends to be less efficient and over fitting in the final model; see the papers by Wang et al. (2007) and Zhang et al. (2010). This motivates us to select the bandwidth h and tuning parameters λ's simultaneously with BIC-type criterion. We define our BIC criterion as

$$\mathrm{BIC}(h,\lambda) = \log(\mathrm{SSE}(h,\lambda)) + \mathrm{df}(h,\lambda)\log(n)/n,$$

where $\mathrm{SSE}(h,\lambda)$ is the sum of squared errors obtained from the penalized least squares with parameters (h,λ), and $\mathrm{df}(h,\lambda)$ is the number of nonzero coefficients of $\hat{\beta}$ conditional on parameters h and λ. This BIC criterion is reasonable since it can balance the trade-off between the variance and the number of non-zero coefficients in terms of the bandwidth h and tuning parameters λ's. Further, it enjoys the property of consistency, which indicates that it can select the correct model with the probability one as the sample size goes to infinity (Zhang et al., 2010). However, it is still computationally expensive to choose d-dimensional tuning parameters $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_d)^T$. By adopting the idea of Fan and Li (2004), to reduce the dimension of λ, we let $\lambda_n = \lambda_0\hat{\sigma}(\hat{\beta}_k^{(0)})$, where $\hat{\sigma}(\hat{\beta}_k^{(0)})$ is the standard deviation of unpenalized estimator $\hat{\beta}_k^{(0)}$. The theoretical properties of $\mathrm{BIC}(h,\lambda)$ and dimension reduction technique with $\lambda_k = \lambda_0\hat{\sigma}(\hat{\beta}_k^{(0)})$ need further research and they can be regarded as the future research topics. The reader is referred to the papers by Cai et al. (2000a) and Fan and Li (2001) more on choosing the bandwidth in nonparametric estimation and the tuning parameters in variable selection.

4. Monte Carlo simulations

Example 1. In this example, we study the finite sample performance of the variable selection for covariates with functional coefficients. In our simulations, the optimal bandwidth and the tuning parameter λ_n are chosen by BIC criterion in Section 3.3. The Epanechnikov kernel $K(x) = 0.75(1-x^2)$ if $|x| \le 1$ is used. We choose the value of a in SCAD to be 3.7 as suggested in Fan and Li (2001).

In this example, we assume that the data are generated by

$$y_i = (Z_{1i}+Z_{2i}) + (Z_{1i}+Z_{2i})^2 X_{1i} + \sigma\varepsilon_i, \quad 1 \le i \le n,$$

and the working model is

$$y_i = g_0(\beta^T Z_i) + \sum_{k=1}^{6} g_k(\beta^T Z_i) X_{ki} + e_i,$$

where ε_i is generated from standard normal distribution and $Z = (Z_1, Z_2)^T$ with $Z_1 = \Phi(Z_1^*)$, $Z_2 = \Phi(Z_2^*)$ and $\Phi(\cdot)$ being the cumulative standard normal distribution function. The eight dimensional vector $(Z_1^*, Z_2^*, X_1, \ldots, X_6)^T$ follows the following vector autoregressive process

$$\begin{pmatrix} Z_i^* \\ X_i \end{pmatrix} = \mathbb{A}\begin{pmatrix} Z_{i-1}^* \\ X_{i-1} \end{pmatrix} + \xi_i,$$

where $Z^* = (Z_1^*, Z_2^*)^T$, $X = (X_1, X_2, \ldots, X_6)^T$ and $\mathbb{A}$ is an 8×8 matrix with the diagonal elements being 0.15 and all others being 0.05. The initial value of $(Z_1^*, X_1)^T$ and each component of the random vector term ξ_i are generated from i.i.d. standard normal distribution. Note that for this set up, the data generated by the above autoregressive process are weekly dependent. We consider three sample sizes as $n = 200$, $n = 400$ and $n = 1000$ and two standard deviations as $\sigma = 2$ and $\sigma = 4$. Sample sizes $n = 200, 400, 1000$ are corresponding to about one year, two years and four years trading days, respectively. For each setting, we replicate 1000 times. The "Shrinkage Rate" and "Keeping rate" are reported in Table 1, in which "Shrinkage rate" represents the percentage that five zero functional coefficients correctly shrink to 0 and "Keeping rate" stands for the percentage that two nonzero functional coefficients do not set to 0 correctly. Clearly, one can see from Table 1 that "Shrinkage rate" and "Keeping rate" produce better results with larger sample size and smaller noise. Meanwhile, it shows that the proposed estimator performs as good as the oracle estimator if the sample size $n = 1000$ as well as the case of $n = 400$ and $\sigma = 2$. This simulation shows that the proposed variable selection procedures perform fairly well for a finite sample.

Table 1
Simulation results for the covariates with functional coefficients.

	$\sigma = 4$	$\sigma = 2$
$n = 200$		
Shrinkage rate	79.4%	93.4%
Keeping rate	92.0%	99.8%
$n = 400$		
Shrinkage rate	94.5%	100%
Keeping rate	98.6%	100%
$n = 1000$		
Shrinkage rate	100%	100%
Keeping rate	100%	100%

Table 2
Simulation results for the local variable with parametric coefficients.

	$\sigma = 15$	$\sigma = 7.5$
$n = 200$		
Shrinkage rate	83.2%	91.1%
Keeping rate	93.4%	96.9%
$n = 400$		
Shrinkage rate	92.3%	100%
Keeping rate	97.5%	100%
$n = 1000$		
Shrinkage rate	100%	100%
Keeping rate	100%	100%

Example 2. To examine the performance of the variable selection for local variables with parametric coefficients, similar to Tibshirani (1996) and Fan and Li (2001), our data generating process is given below

$$y_i = u_i + u_i^2 X_i + \sigma\varepsilon_i,$$

where $u_i = Z_i^T\beta$, $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$ and ε_i is generated from standard normal distribution. Furthermore, the nine dimensional vector $(Z_i^T, X_i)^T$ is generated from the following vector autoregressive process

$$\begin{pmatrix} Z_i \\ X_i \end{pmatrix} = \mathbb{A}^*\begin{pmatrix} Z_{i-1} \\ X_{i-1} \end{pmatrix} + e_i,$$

where $\mathbb{A}^*$ is a 9×9 matrix with the diagonal elements being 0.15 and all others being 0.05. The initial value of $(Z_1^T, X_1)^T$ and each element of the random vector e_i are generated from i.i.d. standard normal. Similar to the previous example, we consider three sample sizes as $n = 200$, 400 and 1000 and for each simulation, we replicate 1000 times. We also consider two values for σ as $\sigma = 7.5$ and $\sigma = 15$. Table 2 displays the simulation results of SCAD variable selection for the local variables with parametric coefficients. Similar to the conclusions from Table 1, it can be seen from Table 2 that the "Shrinkage rate" for irrelevant local variables and "Keeping rate" for relevant local variables perform better with larger sample

Table 3
Simulation results for the two-step selection procedures.

	$\sigma = 2$	$\sigma = 1$
$n = 200$		
Shrinkage rate for nonsignificant covariates	79.1%	89.7%
Keeping rate for significant covariates	92.0%	96.3%
Shrinkage rate for local nonsignificant covariates	77.0%	82.0%
$n = 400$		
Shrinkage rate for nonsignificant covariates	81.8%	91.2%
Keeping rate for significant covariates	95.7%	99.3%
Shrinkage rate for local nonsignificant covariates	82.5%	93.8%
$n = 1000$		
Shrinkage rate for nonsignificant covariates	88.9%	95.7%
Keeping rate for significant covariates	100%	100%
Shrinkage rate for local nonsignificant covariates	93.6%	100%

size and smaller noise. Specifically, it performs as good as the oracle estimator for the cases where $n = 400$ and $\sigma = 7.5$ as well as sample size $n = 1000$. The Monte Carlo simulation results indicate that our variable selection for local variables merits good finite sample properties.

Example 3. To investigate the performance of variable selection for covariates and local variables simultaneously, we do one more step with variable selection for local variables in Example 1. All the settings are the same as in Example 1, except the true model is defined as

$$y_i = g_0(Z_{1i}) + g_1(Z_{1i})X_{1i} + \sigma\,\varepsilon_i, \quad 1 \le i \le n,$$

where $g_0(u) = u$ and $g_1(u) = u^2$. We assume that the index coefficient depends only on local variable Z_1 in this true model. Local variable Z_2 and five covariates $(X_2, \ldots, X_6)$ are not included in the model but estimated in the working model; see Example 1 for details. Two-step selection procedures are employed in this simulation. The first is to select six covariates $(X_1, \ldots, X_6)$ with functional coefficients as well as constant term. Then, we perform variable selection for local variables Z_1 and Z_2 with parametric coefficients. The simulation results for these two-step selection procedures are tabulated in Table 3. Table 3 shows that with larger sample size and smaller noise, shrinkage rates for both nonsignificant covariates and local nonsignificant covariates become larger. These indicate that our two-step procedures perform quite well so that the proposed methods are efficient.

5. Empirical example

In the previous section, we conduct Monte Carlo simulation studies to illustrate the effectiveness of the proposed estimation methods. In this section, to demonstrate the practical usefulness of the proposed model and its estimation methods, we apply these methodologies to consider the predictability of the asset return.

Our data consist of daily, weekly and monthly returns on the three indexes of Dow Jones Industrial Average, NASDAQ Composite and S&P 500 Index. The sample of these three indexes comprise over 30 years between May 1, 1994 and April 30, 2014. They end in 30 April due to the fact that most listing corporations post their annual reports at the end of April. The sample size up to 30 years are considered so that there are enough data for nonparametric estimation in the model. All the data are downloaded from the Wind Information database.[1] Table 4 shows the summary statistics of returns for one day horizon, one week horizon and one month horizon. All horizons show the negative skewness, which indicates that a relatively long lower tail exists. For one day and one week horizons, as expected, they appear to have high sample kurtosis, and it demonstrates that more sample points are further away from the sample mean and their tails are heavier. The Box–Pierce tests show that the autocorrelations of monthly return of three indexes are not significantly different from zero. However, others among daily and weekly horizons are significantly different from zero. This phenomena suggests that most financial variables are not i.i.d. To be precise, they are weakly dependent.

To explore the performance of functional index coefficient autoregressive models, we assume that our working model is established below

$$r_t = g_0(\mathbf{z}_t) + \sum_{j=1}^{p} g_j(\mathbf{z}_t) r_{t-j} + \varepsilon_t$$

$$\varepsilon_t = \sigma_t e_t \quad e_t \sim \text{skewed-}t(\lambda, \nu)$$

$$\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \gamma\varepsilon_{t-1}^2 I_{t-1} + \rho\sigma_{t-1}^2$$

where $\mathbf{z}_t = \beta_1 r_{t-1} + \beta_2 r_{t-2} + \beta_3 r_{t-3}$ and we assume $\beta_1^2 + \beta_2^2 + \beta_3^2 = 1$ in order to satisfy the identification condition. The standardized residuals e_t is skewed-t distributed with skewness parameter λ and degree of freedom ν, γ captures the leverage effect. The indicator function I_{t-1} takes value of 1 for $\varepsilon_t \le 0$ and 0 otherwise. This model can be viewed as an extension of the model by Chen and Tsay (1993). We use the two-step variable selection procedures to select variables and to estimate unknown coefficient functions simultaneously. Firstly, we select covariates based on penalized local least squares and then do variable selection for local important variables based on penalized global least squares. After the two-step variable selection procedures are employed in above model, the estimated coefficients of local variables and the norms of covariates are reported in Table 5. Note that $\mathbf{z}_t$ may include other financial or state economy variables as in Cai et al. (2014a).

In the columns of local variables, both one day lagged return and two days lagged return have effect on the daily return of these three indexes. Three days lagged return does not have any effect on the daily return of both NASDAQ and S&P 500. Only one week lagged return contributes to the weekly return of both NASDAQ and S&P 500. However, two weeks lagged return does not have any contribution. Specifically, one week lagged return and three weeks lagged return perform similar for the weekly return of DOW. For the monthly horizon, one month lagged return has a significant effect on these three indexes, two months lagged return for NASDAQ and three months lagged return for both DOW and S&P 500.

In the columns of covariates, only one day lagged covariate and two days lagged covariate contribute to the daily return of three indexes. Meanwhile, for the weekly horizon, only one week lagged covariate and three weeks lagged covariate are important factors for the weekly return of three indexes. Further, for the monthly horizon, one month lagged covariate has contribution for the monthly return of three indexes, two months lagged return for NASDAQ, and three months lagged return for DOW and S&P 500.

The coefficients of GJR-GARCH model for error terms are tabulated in Table 6. The significance for the skewness λ and the degree of freedom μ for all horizons leads to the non-normality of standardized residuals. It is interesting that leverage effects exist in both one day horizon and one week horizon. However, it cannot be observed in one month horizon. Meanwhile, we cannot find any heteroscedasticity in terms of GJR-GARCH model from one month horizon.

6. Conclusion

Variable selection technology and its algorithms are well developed for models with i.i.d. data and for many fully parametric models. Specifically, variable selection in both semi-parametric

[1] The Web Site for Wind Information is http://www.wind.com.cn/En/Default.aspx.

Table 4
Summary statistics of returns for different horizons.

	Sample size	Mean	Median	StdDev	Skewness	Kurtosis	Min	Max	ρ_1	Box–Pierce test
	One day horizon									
DOW	7572	0.0414	0.0520	1.1229	−1.0822	29.6289	−22.6100	11.0800	−0.0349	0.0000
NASDAQ	7572	0.0471	0.1064	1.4084	−0.0250	8.2972	−11.3500	14.1700	0.0150	0.0000
S&P 500	7572	0.0392	0.0588	1.1522	−0.8293	21.4268	−20.4700	11.5800	−0.0408	0.0000
	One week horizon									
DOW	1566	0.1958	0.3499	2.2885	−0.6595	5.3837	−18.1500	11.2900	−0.0682	0.0001
NASDAQ	1566	0.2249	0.3348	2.9844	−0.7072	7.3782	−25.3000	18.9800	0.0209	0.0304
S&P 500	1566	0.1843	0.3133	2.3041	−0.5709	5.2818	−18.2000	12.0300	−0.0703	0.0004
	One month horizon									
DOW	360	0.8371	1.1460	4.3900	−0.8078	2.8407	−23.2200	13.8200	0.0181	0.9318
NASDAQ	360	0.9957	1.7370	6.4387	−0.5703	1.9663	−27.2300	21.9800	0.1001	0.5826
S&P 500	360	0.7866	1.1430	4.4202	−0.7755	2.3396	−21.7600	13.1800	0.0517	0.9723

Table 5
Coefficients for local variables and covariates.

	Local variables			Covariates[a]						
	r_{t-1}	r_{t-2}	r_{t-3}	1	r_{t-1}	r_{t-2}	r_{t-3}	r_{t-4}	r_{t-5}	r_{t-6}
	One day horizon									
DOW	0.7087	0.6385	−0.3000	3.1532	2.8250	2.6166	0	0	0	0
NASDAQ	0.6609	−0.7505	0	5.0515	3.2321	3.7374	0	0	0	0
S&P 500	0.4183	−0.9083	0	4.4862	1.8567	3.9876	0	0	0	0
	One week horizon									
DOW	0.7716	0	0.6360	16.4212	8.0355	0	6.6187	0	0	0
NASDAQ	1	0	0	30.9023	14.9169	0	4.3741	0	0	0
S&P 500	1	0	0	15.5780	9.5644	0	2.1025	0	0	0
	One month horizon									
DOW	0.8638	0	−0.5034	70.1698	59.3783	0	34.5686	0	0	0
NASDAQ	0.8488	0.5286	0	189.735	68.8333	42.7579	0	0	0	0
S&P 500	0.8697	0	0.4934	98.7581	48.1937	0	27.6172	0	0	0

[a] We calculate the norm of the functional coefficients for covariates.

Table 6
Estimation results of GJR-GARCH model for error terms. The skewness λ and the degree of freedom μ are parameters of skewed-t distribution of standardized residuals e_t. Four parameters ω, α, γ and ρ are from GJR-GARCH model. The corresponding t-ratios based on robust standard errors are reported in parentheses.

	λ	μ	ω	α	γ	ρ
	One day horizon					
DOW	0.8594*	8.4413*	0.0134*	0.0000	0.1786*	0.9009*
	(32.02)	(5.02)	(3.20)	(0.00)	(5.53)	(49.04)
NASDAQ	0.8592*	12.8651*	0.0263*	0.0000	0.1524*	0.9068*
	(31.02)	(3.30)	(2.44)	(0.02)	(3.83)	(38.27)
S&P 500	0.8534*	8.2578*	0.0201*	0.0000	0.1882*	0.8935*
	(28.55)	(5.45)	(3.14)	(0.48)	(5.02)	(44.84)
	One week horizon					
DOW	0.8670*	9.1538*	0.2994	0.0131	0.1913*	0.8166*
	(29.45)	(4.48)	(1.79)	(0.59)	(2.96)	(10.40)
NASDAQ	0.8886*	7.6217*	0.2626*	0.0657*	0.1351*	0.8262*
	(24.22)	(5.47)	(1.97)	(0.03)	(0.06)	(0.06)
S&P 500	0.8709*	10.4121*	0.2680	0.0045	0.2322*	0.8119*
	(30.73)	(3.98)	(1.75)	(0.24)	(3.19)	(10.55)
	One month horizon					
DOW	1.0221*	2.0107*	0.0000	0.4868	1.0000	0.0013
	(84.93)	(177.50)	(0.08)	(0.52)	(0.59)	(0.20)
NASDAQ	1.0304*	2.0100*	0.0036	0.0140	1.0000	0.2310
	(69.11)	(234.73)	(0.70)	(0.09)	(0.51)	(0.34)
S&P 500	1.0056*	2.0344*	0.0918	0.1551	0.4581	0.6138
	(53.03)	(37.72)	(0.61)	(0.16)	(0.32)	(0.72)

* Denotes significance at confidence level 5%.

and nonparametric models has become popular in recent years. In contrast to the i.i.d. setting in those papers with variable selection, we considered variable selection in functional index coefficient models under strong mixing context. Most weakly dependent financial time series can be analyzed in our procedures under the general conditions considered in this paper. Our variable selection procedures select both covariates with functional coefficients and local variables with parametric coefficients in two steps. Theoretical properties such as consistency, sparsity, and the oracle property of these two-step estimators are derived. Monte Carlo simulations show that our two-step procedures perform fairly well. To address the issue of stock return predictability, an example of functional index coefficient autoregressive models is extensively studied and it can be viewed as an extension of the model in Chen and Tsay (1993).

In financial economics, many of the regressions may suffer spurious regression due to the presence of highly persistent regressors. Persistence can be found in many financial variables, such as book-to-market ratios, dividend–price ratio, earning–price ratio, short-term Treasury bill rate and yield spread (Campbell and Yogo, 2006; Phillips and Lee, 2013). The theories for regression model with persistent variables are very different from the model with stationary variables; see, for example, Cai and Wang (2014) and Cai et al. (forthcoming). And there is little literature regarding variable selection in the model with persistent regressors. For future research, it would be interesting to consider variable selection for the linear and nonlinear time series prediction models with persistent and/or nonstationary variables.

Appendix. Mathematical proofs

In this Appendix, we present briefly the derivations of the main results given in previous sections. Before embracing on the proofs, we define some notations and list some lemmas that will be used throughout this appendix. First, let C be a finite positive constant and R_m denotes ignorable small order term. Both of them might be different in different appearances. Now, we present Lemmas 1 and 2.

Lemma 1. *Let* $\{X_i, Z_i, y_i\}$ *be a strong mixing and strictly stationary sequence. Under Assumptions* A1–A8. *Assume that* $h \propto n^{-1/5}$,

$n^{-1/10}\alpha_n \to 0$ and $\|\hat\beta - \beta_0\| = O_p(1/\sqrt{n})$, we have

$$n^{-1}\sum_{i=1}^{n}\|\hat g(\hat\beta^T Z_i) - g_0(\beta_0^T Z_i)\|^2 = O_p(n^{-4/5}).$$

Proof. By the triangle inequality $n^{-1}\sum_{i=1}^{n}\|\hat g(\hat\beta^T Z_i) - g_0(\beta_0^T Z_i)\|^2 \le n^{-1}\sum_{i=1}^{n}\|\hat g(\hat\beta^T Z_i) - g_0(\hat\beta^T Z_i)\|^2 + n^{-1}\sum_{i=1}^{n}\|g_0(\hat\beta^T Z_i) - g_0(\beta_0^T Z_i)\|^2$.

The second term on the right hand side

$$\begin{aligned}
n^{-1}&\sum_{i=1}^{n}\|g_0(\hat\beta^T Z_i) - g_0(\beta_0^T Z_i)\|^2\\
&= n^{-1}\sum_{i=1}^{n}\|\dot g_0(\beta_0^T Z_i)(\hat\beta - \beta_0)^T Z_i + o_p(n^{-1/2})\|^2\\
&\text{(by Taylor expansion)}\\
&\le n^{-1}C\sum_{i=1}^{n}(\hat\beta - \beta_0)^T Z_i Z_i^T(\hat\beta - \beta_0) + o_p(n^{-1})\\
&\text{(by Assumption A1)}\\
&= C(\hat\beta - \beta_0)^T E(Z_i Z_i^T)(\hat\beta - \beta_0) + o_p(n^{-1})\\
&= O_p(n^{-1}).
\end{aligned}$$

In the above equation, C is the maximum value of $\|\dot g_0(\beta_0^T Z_i)\|^2$. We can conclude that the order of the second term is of order as $O_p(n^{-1})$. Now, it suffices to show that $n^{-1}\sum_{i=1}^{n}\|\hat g\left(\hat\beta^T Z_i\right) - g_0\left(\hat\beta^T Z_i\right)\|^2 = O_p(n^{-4/5})$. Following the proof in Wang and Xia (2009), we let $u = (u_{ik}) \in \mathbb{R}^{n\times p}$ be an arbitrary $n \times p$ matrix with rows $u_{i\cdot}$ and columns $u_{\cdot k}$ and $u = (u_{1\cdot}, u_{2\cdot}, \dots, u_{n\cdot})^T = (u_{\cdot 1}, u_{\cdot 2}, \dots, u_{\cdot p})$. Set $\|u\| = \sqrt{\sum_{i,k}u_{i,k}^2}$ to be the L_2-norm for an arbitrary matrix $u = (u_{ik})$. For any small $\varepsilon > 0$, if we can show that there is a large constant C such that $P\{\inf_{n^{-1}\|u\|^2=C} Q(G_0 + (nh)^{-1/2}u, \hat\beta) > Q(G_0, \hat\beta)\} > 1 - \varepsilon$, then the proof is finished. To this end, define

$$\begin{aligned}
D &\equiv n^{-1}h\{Q(G_0 + (nh)^{-1/2}u, \hat\beta) - Q(G_0, \hat\beta)\}\\
&= n^{-1}h\left\{\sum_{j=1}^{n}\sum_{i=1}^{n}\left[y_i - g_0^T(\hat\beta^T Z_j)X_i - (nh)^{-1/2}u_{j\cdot}^T X_i\right]^2\right.\\
&\quad \times K_h\left(\hat\beta^T Z_i - \hat\beta^T Z_j\right) - \sum_{j=1}^{n}\sum_{i=1}^{n}\left[y_i - g_0^T(\hat\beta^T Z_j)X_i\right]^2\\
&\quad \left.\times\ K_h\left(\hat\beta^T Z_i - \hat\beta^T Z_j\right)\right\}\\
&\quad + h\sum_{k=1}^{p}\left[P_{\lambda_n}\left(\|g_{0\cdot k} + (nh)^{-1/2}u_{\cdot k}\|\right) - P_{\lambda_n}\left(\|g_{0\cdot k}\|\right)\right]\\
&\ge n^{-1}\sum_{j=1}^{n}\left[u_{j\cdot}^T\hat\Sigma(\hat\beta^T Z_j)u_{j\cdot} - 2u_{j\cdot}^T\hat e_j\right]\\
&\quad + h\sum_{k=1}^{p_0}\left[P_{\lambda_n}\left(\|g_{0\cdot k} + (nh)^{-1/2}u_{\cdot k}\|\right) - P_{\lambda_n}\left(\|g_{0\cdot k}\|\right)\right],
\end{aligned}$$

where $\hat\Sigma(\hat\beta^T Z_j) = n^{-1}\sum_{i=1}^{n}X_i X_i^T K_h(\hat\beta^T Z_i - \hat\beta^T Z_j)$ and $\hat e_j = n^{-1/2}h^{1/2}\sum_{i=1}^{n}[X_i X_i^T(g_0(\beta_0^T Z_i) - g_0(\hat\beta^T Z_i)) + X_i X_i^T(g_0(\hat\beta^T Z_i) - g_0(\hat\beta^T Z_j)) + X_i\varepsilon_i]K_h(\hat\beta^T Z_i - \hat\beta^T Z_j)$. Let $\hat\lambda_j^{\min}$ be the smallest eigenvalue of $\hat\Sigma(\hat\beta^T Z_j)$, $\hat\lambda_{\min} = \min\{\hat\lambda_j^{\min}, j = 1, \dots, n\}$ and $\hat e = (\hat e_1, \dots, \hat e_n)^T = \mathbb{R}^{n\times p}$. Then, $D \ge n^{-1}\sum_{j=1}^{n}(\|u_{j\cdot}\|^2\hat\lambda_j^{\min} - 2\|u_{j\cdot}\|\|\hat e_j\|) - n^{-1/2}h^{1/2}\sum_{k=1}^{p_0}P'_{\lambda_n}(\|g_{0\cdot k}\|)\|u_{\cdot k}\|$, where the first term on the right hand side is followed by Cauchy–Schwarz inequality and the second term is followed by Taylor expansion and triangle inequality. Therefore,

$$\begin{aligned}
D &\ge \hat\lambda_{\min}n^{-1}\sum_{j=1}^{n}\|u_{j\cdot}\|^2 - 2(n^{-1}\|u\|^2)^{1/2}(n^{-1}\|\hat e\|^2)^{1/2}\\
&\quad - n^{-1/2}h^{1/2}\alpha_n\sum_{k=1}^{p_0}\|u_{\cdot k}\|\\
&\ge \hat\lambda_{\min}n^{-1}\|u\|^2 - 2(n^{-1}\|u\|^2)^{1/2}(n^{-1}\|\hat e\|^2)^{1/2}\\
&\quad - h^{1/2}\alpha\sqrt{p_0}\left(n^{-1}\sum_{k=1}^{p_0}\|u_{\cdot k}\|^2\right)^{1/2}\\
&= \hat\lambda_{\min}C - 2\sqrt{C}(n^{-1}\|\hat e\|^2)^{1/2} - h^{1/2}\alpha_n\sqrt{p_0}\sqrt{C}.
\end{aligned}$$

As we will show later that

$$n^{-1}\|\hat e\|^2 = O_p(1) \quad\text{and}\quad \hat\lambda_{\min} \to^P \lambda_0^{\min} \quad \text{as } n \to \infty,$$

where $\lambda_0^{\min} = \inf_{z\in[0,1]}\lambda_{\min}(f(\hat\beta Z)\Omega(\hat\beta Z))$, $\lambda_{\min}(\cdot)$ denotes the minimal eigenvalue of an arbitrary positive definite matrix. By Assumptions A2 and A4, as $\lambda_0^{\min} > 0$ and $h^{1/2}\alpha_n \to 0$, we can show that $D > 0$ for a sufficient large C. Then, this proof is complete.

To show $n^{-1}\|\hat e\|^2 = O(1)$, it is easy to see that

$$n^{-1}\|\hat e\|^2 \to^P E\|\hat e_j\|^2$$

and

$$\begin{aligned}
E\|\hat e_j\|^2 &\le n^{-1}hE\left\|\sum_{i=1}^{n}[X_i X_i^T(g_0(\hat\beta^T Z_i) - g_0(\hat\beta^T Z_j))]\right.\\
&\quad \left.\times\ K_h(\hat\beta^T Z_i - \hat\beta^T Z_j)\right\|^2\\
&\quad + n^{-1}hE\left\|\sum_{i=1}^{n}[X_i\varepsilon_i K_h(\hat\beta^T Z_i - \hat\beta^T Z_j)]\right\|^2\\
&\quad + n^{-1}hE\left\|\sum_{i=1}^{n}[X_i X_i^T(g_0(\beta_0^T Z_i) - g_0(\hat\beta^T Z_i))]\right.\\
&\quad \left.\times\ K_h(\hat\beta^T Z_i - \hat\beta^T Z_j)\right\|^2\\
&\equiv A + B + \tilde D
\end{aligned}$$

where A denotes the first term, B is for the second term and $\tilde D$ stands for the last term. We introduce notations as $z_i = \beta^T Z_i$, $z_s = \beta^T Z_s$ and $z_j = \beta^T Z_j$

$$\begin{aligned}
A &= n^{-1}hE\left\{\sum_{i\ne s\ne j}[(g_0(\hat\beta^T Z_i) - g_0(\hat\beta^T Z_j))^T X_i X_i^T X_s X_s^T(g_0(\hat\beta^T Z_s)\right.\\
&\quad - g_0(\hat\beta^T Z_j))\\
&\quad \left.\times\ K_h(\hat\beta^T Z_i - \hat\beta^T Z_j)K_h(\hat\beta^T Z_s - \hat\beta^T Z_j)]\right\} + n^{-1}hE\sum_{(i=s)\ne j}\{\cdots\}\\
&\equiv n^{-1}hE\left\{\sum_{i\ne s\ne j}[(g_0(z_i) - g_0(z_j))^T X_i X_i^T X_s X_s^T(g_0(z_s) - g_0(z_j))\right.\\
&\quad \left.\times\ K_h(z_i - z_j)K_h(z_s - z_j)]\right\} + n^{-1}hE\sum_{(i=s)\ne j}\{\cdots\}\\
&\equiv A_1 + A_2,
\end{aligned}$$

where A_1 denotes the first term and A_2 is for the second term. It is easy to show that

$$\begin{aligned} A_1 &\equiv nhE\{(g_0(z_i)-g_0(z_j))^T X_iX_i^T X_sX_s^T(g_0(z_s) \\ &\quad -g_0(z_j))K_h(z_i-z_j)K_h(z_s-z_j)\}+R_m \\ &= nhE\{(g_0(z_i)-g_0(z_j))^T\Omega(z_i,z_s,z_j)(g_0(z_s)\\ &\quad -g_0(z_j))K_h(z_i-z_j)K_h(z_s-z_j)\}+R_m\\ &= nh\int E\{(g_0(z_i)-g_0(z_j))^T\Omega(z_i,z_s,z_j)\\ &\quad\times(g_0(z_s)-g_0(z_j))K_h(z_i-z_j)K_h(z_s-z_j)|z_j\}f(z_j)dz_j+R_m\\ &\equiv A_{11}+R_m,\end{aligned}$$

where the definition of A_{11} is apparent and R_m is an ignorable small order term, which might be different in different appearances. Let $z_i=z_j+wh$ and $z_s=z_j+vh$. Then,

$$\begin{aligned} A_{11} &= nh\int\left\{\int\int\left(\dot{g}_0(z_j)wh+\frac{1}{2}C_1w^2h^2\right)^T\right.\\ &\quad\times\Omega(z_j+wh,z_j+vh,z_j)\left(\dot{g}_0(z_j)vh+\frac{1}{2}C_2v^2h^2\right)\\ &\quad\left.\times\ k(w)k(v)f((z_j+wh,z_j+vh)|z_j)dwdv\right\}f(z_j)dz_j\\ &\equiv nh\int A_{12}(z_j)f(z_j)dz_j,\end{aligned}$$

where

$$\begin{aligned} A_{12}(z_j) &= \int\int\left(\dot{g}_0(z_j)wh+\frac{1}{2}C_1w^2h^2\right)^T[\Omega(z_j,z_j,z_j)\\ &\quad+\Omega_1(z_j,z_j,z_j)wh+\Omega_2(z_j,z_j,z_j)vh\\ &\quad+o_p(w^2h^2)+o_p(v^2h^2)]\\ &\quad\times\left(\dot{g}_0(z_j)vh+\frac{1}{2}C_2v^2h^2\right)[f((z_j,z_j)|z_j)\\ &\quad+f_1((z_j,z_j)|z_j)wh+f_2((z_j,z_j)|z_j)vh\\ &\quad+o_p(w^2h^2)+o_p(v^2h^2)]k(w)k(v)dwdv\\ &= I_{12}(Z_j)h^4\int w^2v^2k(w)k(v)dwdv+o_p(h^4)\end{aligned}$$

where $I_{12}(Z_j)$ is an integrable function. Then, $A_1=O_p(nh^5)=O_p(1)$. Also, we can show that

$$\begin{aligned} A_2 &= n^{-1}hE\left\{\sum_{i\neq j}[(g_0(z_i)-g_0(z_j))^T X_iX_i^T X_iX_i^T\right.\\ &\quad\left.\times\ (g_0(z_i)-g_0(z_j))K_h^2(z_i-z_j)]\right\}\\ &= hE\{(g_0(z_i)-g_0(z_j))^T X_iX_i^T X_iX_i^T(g_0(z_i)-g_0(z_j))\\ &\quad\times K_h^2(z_i-z_j)\}+R_m\\ &= hE\{(g_0(z_i)-g_0(z_j))^T\Omega(z_i,z_j)(g_0(z_i)-g_0(z_j))\\ &\quad\times K_h^2(z_i-z_j)\}+R_m\\ &= h\int E\{(g_0(z_i)-g_0(z_j))^T\Omega(z_i,z_j)(g_0(z_i)\\ &\quad-g_0(z_j))K_h^2(z_i-z_j)|z_j\}f(z_j)dz_j+R_m\\ &\equiv h\int A_{21}f(z_j)dz_j+R_m,\end{aligned}$$

where

$$\begin{aligned} A_{21}(z_j) &\equiv \int(g_0(z_i)-g_0(z_j))^T\Omega(z_i,z_j)(g_0(z_i)-g_0(z_j))\\ &\quad\times K_h^2(z_i-z_j)f(z_i|z_j)dz_i.\end{aligned}$$

Let $z_i=z_j+wh$. Then,

$$\begin{aligned} A_{21}(z_j) &= \frac{1}{h}\int(\dot{g}_0(z_j)wh+Cw^2h^2)^T\Omega(z_j+wh,z_j)\\ &\quad\times(\dot{g}_0(z_j)wh+Cw^2h^2)k^2(w)f(z_j+wh|z_j)dw\\ &= I_{21}(Z_j)h\int w^2k^2(w)dw+R_m\end{aligned}$$

where $I_{21}(Z_j)$ is an integrable function, then $A_2=O_p(h^2)=o_p(1)$. Hence, $A=O_p(1)$, Now, we consider the term B as follows.

$$\begin{aligned} B &= n^{-1}hE\left\{\left(\sum_{i=1}^n X_i\varepsilon_iK_h(z_i-z_j)\right)^T\left(\sum_{s=1}^n X_s\varepsilon_sK_h(z_s-z_j)\right)\right\}\\ &= n^{-1}hE\left\{\sum_{(i=s)\neq j}X_iX_s^T\varepsilon_i\varepsilon_sK_h(z_i-z_j)K_h(z_s-z_j)\right\}\\ &\quad+2n^{-1}hE\left\{\sum_{(i=j)\neq s}X_iX_s^T\varepsilon_i\varepsilon_sK_h(z_i-z_j)K_h(z_s-z_j)\right\}\\ &\quad+n^{-1}hE\left\{\sum_{(i\neq s)\neq j}X_iX_s^T\varepsilon_i\varepsilon_sK_h(z_i-z_j)K_h(z_s-z_j)\right\}\\ &\quad+n^{-1}hE\left\{\sum_{i=s=j}X_iX_s^T\varepsilon_i\varepsilon_sK_h(z_i-z_j)K_h(z_s-z_j)\right\}\\ &\equiv B_1+B_2+B_3+B_4,\end{aligned}$$

where the definitions of B_j's are apparent. Now,

$$\begin{aligned} B_1 &= hE[X_iX_i^T\varepsilon_i^2K_h^2(z_i-z_j)]+R_m\\ &= hE[X_iX_i^TK_h^2(z_i-z_j)E(\varepsilon_i^2|X_i,z_i,z_j)]+R_m\\ &= h\sigma^2E[X_iX_i^TK_h^2(z_i-z_j)]+R_m\\ &= h\sigma^2E[\Omega(z_i,z_j)K_h^2(z_i-z_j)]+R_m\\ &= h\sigma^2E\{E[\Omega(z_i,z_j)K_h^2(z_i-z_j)|z_j]\}+R_m.\end{aligned}$$

Let $z_i=z_j+wh$. Then, we have

$$\begin{aligned} &E[\Omega(z_i,z_j)K_h^2(z_i-z_j)|z_j]\\ &\quad=\frac{1}{h^2}\int\Omega(z_i,z_j)k^2\left(\frac{z_i-z_j}{h}\right)f_{z_i|z_j}(z_i|z_j)dz_i\\ &\quad=\frac{1}{h}\int\Omega(z_j+wh,z_j)k^2(w)f_{z_i|z_j}(z_j+wh|z_j)dw\\ &\quad=I_{B2}(z_j)O_p(1/h)\int k^2(w)dw,\end{aligned}$$

where $I_{B1}(z_j)$ is an integral function of z_j, so that $B_1=O_p(1)$,

$$\begin{aligned} B_2 &= 2n^{-1}hE\left\{\sum_{s\neq j}X_jX_s^T\varepsilon_j\varepsilon_sK_h(0)K_h(z_s-z_j)\right\}\\ &= 2n^{-1}h\sum_{\ell=-\infty}^{\infty}E[X_{s+\ell}X_s^T\varepsilon_{\ell+s}\varepsilon_sK_h(0)K_h(z_{s+\ell}-z_s)]+R_m\\ &= 2n^{-1}h\sum_{\ell=-\infty}^{\infty}E[E(X_{s+\ell}X_s^T\varepsilon_{\ell+s}\varepsilon_s|z_{\ell+s},z_s)\\ &\quad\times K_h(0)K_h(z_{s+\ell}-z_s)]+R_m\\ &= O_p(1),\end{aligned}$$

$$\begin{aligned} B_3 &= n^{-1}hE\left\{\sum_{(i\neq s)\neq j} X_iX_s^T\varepsilon_i\varepsilon_sK_h(z_i-z_j)K_h(z_s-z_j)\right\} \\ &= n^{-1}h\sum_{\ell=-\infty}^{\infty} E[X_iX_{i-\ell}^T\varepsilon_i\varepsilon_{i-\ell}K_h(z_i-z_j)K_h(z_{i-\ell}-z_j)] \\ &= n^{-1}h\sum_{\ell=-\infty}^{\infty} E[E(X_iX_{i-\ell}^T\varepsilon_i\varepsilon_{i-\ell}|z_i,z_{i-\ell},z_j) \\ &\quad\times K_h(z_i-z_j)K_h(z_{i-\ell}-z_j)] \\ &= O_p(h), \end{aligned}$$

and

$$\begin{aligned} B_4 &= n^{-1}hE[X_j^TX_j\varepsilon_j^2K_h^2(0)] = n^{-1}hE[X_j^TX_jE(\varepsilon_j^2|X_j)K_h^2(0)] \\ &= n^{-1}h\sigma^2K_h^2(0)E[X_j^TX_j] = O_p(n^{-4/5}). \end{aligned}$$

Thus, $B = O_p(1)$. Now,

$$\begin{aligned} \tilde{D} &= n^{-1}hE\left\|\sum_{i=1}^{n}[X_iX_i^T(g_0(\beta_0^TZ_i)-g_0(\hat{\beta}^TZ_i))]K_h(\hat{\beta}^TZ_i-\hat{\beta}^TZ_j)\right\|^2 \\ &\le hE\|[X_iX_i^T(g_0(\beta_0^TZ_i)-g_0(\hat{\beta}^TZ_i))]K_h(\hat{\beta}^TZ_i-\hat{\beta}^TZ_j)\|^2 \\ &= hE\|[X_iX_i^T(\dot{g}_0(\beta_0^TZ_i)(\hat{\beta}-\beta_0)^TZ_i+o_p(n^{-1/2}))] \\ &\quad\times K_h(\hat{\beta}^TZ_i-\hat{\beta}^TZ_j)\|^2 \\ &\le C(h/n)E\|X_iX_i^TK_h(\hat{\beta}^TZ_i-\hat{\beta}^TZ_j)\|^2 \\ &= O_p(1/n). \end{aligned}$$

This proves the lemma. □

Lemma 2. *Let $\{X_i, Z_i, y_i\}$ be a strong mixing and strictly stationary sequence, $h \propto n^{-1/5}$, $\lim_{n\to\infty}\inf_{\theta\to 0^+}P'_{\lambda_n}(\theta)/\lambda_n > 0$, and $n^{-1/10}\lambda_n \to 0$. Then, $\|\hat{g}_{.k}\| = 0$ as $n \to \infty$ for $k > d_0$.*

Proof. Assume $\|\hat{g}_{.k}\| \neq 0$, then,

$$\frac{\partial Q(G,\hat{\beta},h)}{\partial g_{.k}} = J_1 + J_2 = 0,$$

where $J_1 = (J_{11}, J_{12}, \ldots, J_{1n})^T$, $J_{1j} = -2\sum_{i=1}^{n}X_{ik}\left(y_i - \hat{g}^T(\hat{\beta}^TZ_j)X_i\right)$ $K_h(\hat{\beta}^TZ_i-\hat{\beta}^TZ_j)$, and $J_2 = nP'_{\lambda_n}(\|g_{.k}\|)\frac{g_{.k}}{\|g_{.k}\|}$. Similar to the proof of (A.7) in Wang and Xia (2009), by Lemma 1, we can derive that $\|J_1\| = O_p(nh^{-1/2})$ and we know $\|J_2\| = nP'_{\lambda_n}(\|g_{.k}\|) = \frac{P'_{\lambda_n}(\|g_{.k}\|)}{\lambda_n}\cdot\sqrt{h}\lambda_n\cdot nh^{-1/2}$. Since $\frac{P'_{\lambda_n}(\|g_{.k}\|)}{\lambda_n} > 0$ and $\sqrt{h}\lambda_n \to 0$, then $P(\|J_2\| < \|J_1\|) \to 1$ as $n \to \infty$. It contradicts with the assumption. Hence, $\|\hat{g}_{.k}\| = 0$ as $n \to \infty$. □

Proof of Theorem 2. (a) Following the similar steps in the Proof of Theorem 1 by Wang and Xia (2009), with Lemma 2 and Hunter and Li (2005), we can conclude $\sup_{Z\in\mathcal{A}_z}\|\hat{g}_k(z,\hat{\beta})\| = 0$, for all $d_1 < k \le d$.

(b) We want to show that there exists a $\hat{G}_a$ such that it is the minimizer of $Q((G_a, 0), \hat{\beta}, h)$. Taking the first derivative of $Q((G_a, 0), \hat{\beta}, h)$ with respective to $\hat{g}_a(\hat{\beta}^TZ_j)$, we can get the normal equation as

$$\sum_{i=1}^{n}X_{ia}\left(y_i - \hat{g}_a^T(\hat{\beta}^TZ_j)X_{ia}\right)K_h(\hat{\beta}^TZ_i-\hat{\beta}^TZ_j) + n\Pi_j = 0,$$

where Π_j is a a-dimensional vector with its kth component given by

$$P'_{\lambda_n}(\|\hat{g}_{.k}\|)\frac{\hat{g}_k(\hat{\beta}^TZ_j)}{\|\hat{g}_{.k}\|}.$$

Since $P'_{\lambda_n}(\|\hat{g}_{.k}\|) = 0$ when $\|\hat{g}_{.k}\| \neq 0$ and n is large, then, $\Pi = 0$ follows when n is large. Note that

$$\sum_{i=1}^{n}X_{ia}\left(y_i - \hat{g}_a^T(\hat{\beta}^TZ_j)X_{ia}\right)K_h(\hat{\beta}^TZ_i-\hat{\beta}^TZ_j) = 0.$$

In fact, the above normal equation also holds for all $z = \hat{\beta}^TZ$, $Z \in \mathcal{A}_z$, $\hat{\beta} \in \mathbb{B}$. It turns out

$$\sum_{i=1}^{n}X_{ia}\left(y_i - \hat{g}_a^T(z,\hat{\beta})X_{ia}\right)K_h(\hat{\beta}^TZ_i - z) = 0$$

and

$$\hat{g}_a(z,\hat{\beta}) = \left[\sum_{i=1}^{n}X_{ia}X_{ia}^TK_h(\hat{\beta}^TZ_i-z)\right]^{-1}\sum_{i=1}^{n}X_{ia}y_iK_h(\hat{\beta}^TZ_i-z).$$

Then,

$$\begin{aligned} \hat{g}_a(z,\hat{\beta}) - g_{0a}(z,\beta_0) &= \{\hat{g}_a(z,\hat{\beta}) - \hat{g}_a(z,\beta_0)\} \\ &\quad+ \{\hat{g}_a(z,\beta_0) - g_{0a}(z,\beta_0)\}. \end{aligned}$$

By Taylor expansion, the first term in the right hand side of the above equation is the order of $O_p(n^{-1/2})$ and the second term in the right hand side is the order of $O_p(n^{-2/5})$. Thus the asymptotic property of the $\hat{g}_a(z,\hat{\beta}) - g_{0a}(z,\beta_0)$ is the same as the second term. And the asymptotic property of the second term $\hat{g}_a(z,\beta_0) - g_{0a}(z,\beta_0)$ can be found in the proof of Theorem 3 by Xia and Li (1999). □

Proof of Theorem 3. It follows from Theorem 1 in Xia and Li (1999) that

$$\hat{Q}_1(\beta,h) = \tilde{S}(\beta) + T(h) + R_1(\beta,h) + R_2(h),$$

where $\hat{Q}_1(\beta,h) = \sum_{i=1}^{n}(y_i - \hat{g}^T(\beta^TZ_i)X_i)^2$, $T(h)$ and $R_2(h)$ do not depend on β, and $R_1(\beta,h)$ is an ignorable term. Furthermore,

$$\begin{aligned} \tilde{S}(\beta) &= n[\tilde{V}_0^{1/2}(\beta-\beta_0) - n^{-1/2}\sigma\varepsilon]^T[\tilde{V}_0^{1/2}(\beta-\beta_0) - n^{-1/2}\sigma\varepsilon] \\ &\quad+ R_3 + R_4(\beta), \end{aligned}$$

where R_3 does not depend on β and h, and $R_4(\beta)$ is an ignorable term.

Let $\delta_n = n^{-1/2} + a_n$, $t = (t_1, \ldots, t_d)^T$. For any small $\varepsilon > 0$, if we can show there exists a large constant C, such that

$$P\{\inf_{\|t\|=C}Q(\beta_0+\delta_nt,\hat{g}) > Q(\beta_0,\hat{g})\} > 1-\varepsilon,$$

then

$$\|\hat{\beta}-\beta_0\| = O_p(\delta_n).$$

Define $D_n = Q(\beta_0+\delta_nt,\hat{g}) - Q(\beta_0,\hat{g})$. Then,

$$\begin{aligned} D_n &\ge \frac{1}{2}\sum_{i=1}^{n}(y_i - \hat{g}^T(\beta_0^TZ_i+\delta_nt^TZ_i)X_i)^2 \\ &\quad- \frac{1}{2}\sum_{i=1}^{n}(y_i - \hat{g}^T(\beta_0^TZ_i)X_i)^2 \\ &\quad+ n\sum_{k=1}^{d_1}\Psi_{\zeta_n}(|\beta_{10k}+\delta_nt_k|) - n\sum_{k=1}^{d_1}\Psi_{\zeta_n}(|\beta_{10k}|) \\ &\quad(\text{by } \beta_{20} = 0) \end{aligned}$$

and

$$
\begin{aligned}
& n\sum_{k=1}^{d_1}\Psi_{\zeta_n}(|\beta_{10k}+\delta_n t_k|) - n\sum_{k=1}^{d_1}\Psi_{\zeta_n}(|\beta_{10k}|) \\
&\quad = n\sum_{k=1}^{d_0}\left[\delta_n\Psi'_{\zeta_n}(|\beta_{10k}|)sgn(\beta_{10k})t_k + \frac{1}{2}\delta_n^2\Psi''_{\zeta_n}(|\beta_{10k}|)t_k^2\right] \\
&\qquad + o_p(n\delta_n^2) \\
&\quad \le \sqrt{d_1}n\delta_n a_n\|t\| + \frac{1}{2}n\delta_n^2 max_{1\le k\le d_0}\{\Psi''_{\zeta_n}(|\beta_{10k}|)\}\|t\|^2 + o_p(n\delta_n^2) \\
&\qquad \text{(by Cauchy–Schwarz inequality)} \\
&\quad \le n\delta_n^2\sqrt{d_0}C + O_p(n\delta_n^2) \\
&\qquad \text{as } n\to\infty \text{ and } max_{1\le k\le d_0}\{\Psi''_{\zeta_n}(|\beta_{10k}|)\}\to 0
\end{aligned}
$$

and

$$
\begin{aligned}
& \frac{1}{2}\sum_{i=1}^{n}(y_i-\hat{g}^T(\beta_0^T Z_i+\delta_n t^T Z_i)X_i)^2 - \frac{1}{2}\sum_{i=1}^{n}(y_i-\hat{g}^T(\beta_0^T Z_i)X_i)^2 \\
&\quad = \frac{1}{2}n[\tilde{V}_0^{1/2}\delta_n t - n^{-1/2}\sigma\varepsilon]^T[\tilde{V}_0^{1/2}\delta_n t - n^{-1/2}\sigma\varepsilon] \\
&\qquad - \frac{1}{2}n[n^{-1/2}\sigma\varepsilon]^T[n^{-1/2}\sigma\varepsilon] \\
&\qquad + R_1(\beta_0+\delta_n t, h) - R_1(\beta_0, h) + o_p(1) \\
&\qquad \text{(by the theorem in Xia and Li, 1999)} \\
&\quad = \frac{1}{2}n\delta_n^2 t^T\tilde{V}_0 t - n^{1/2}\delta_n t^T\tilde{V}_0^{1/2}\sigma\varepsilon + R_1(\beta_0+\delta_n t, h) \\
&\qquad - R_1(\beta_0, h) + o_p(1) \\
&\quad = \frac{1}{2}n\delta_n^2 t^T\tilde{V}_0 t - \delta_n t^T V_n + R_1(\beta_0+\delta_n t, h) - R_1(\beta_0, h) + o_p(1).
\end{aligned}
$$

Since R_1 are negligible terms as $n\to\infty$ and $\frac{1}{\sqrt{n}}V_n = O_p(1)$. then $-\delta_n t^T V_n = C\cdot O_p(\delta_n\sqrt{n}) = C\cdot O_p(\delta_n^2 n)$. By choosing a sufficient large C, the term $\frac{1}{2}n\delta_n^2 t^T\tilde{V}_0 t$ will dominate others. Hence, $D_n\ge 0$ holds. □

Proof of Theorem 4. Let $\hat{\beta}_1-\beta_{10} = O_p\left(n^{-1/2}\right)$. We want to show that $\left(\hat{\beta}_1, 0\right)^T = \text{argmin}_{\left(\beta_1^T,\beta_2^T\right)^T\in\mathbb{B}} Q\left(\left(\beta_1^T,\beta_2^T\right)^T, \hat{g}\right)$. It suffices to show that for some constant C and $k = q_0+1,\ldots,q$,

$$
\begin{aligned}
\frac{\partial Q\left(\left(\beta_1^T,\beta_2^T\right)^T,\hat{g}\right)}{\partial\beta_k} &> 0 \quad \text{for } 0<\beta_k<Cn^{-1/2} \\
&< 0 \quad \text{for } -Cn^{-1/2}<\beta_k<0.
\end{aligned}
$$

Note that

$$
\begin{aligned}
\frac{\partial\hat{Q}_1(\beta,h)}{\partial\beta_k} &= \frac{\partial\tilde{S}(\beta)}{\partial\beta_k} + R_m \\
&= e_k^T\frac{\partial\tilde{S}(\beta)}{\partial\beta} + R_m \\
&= 2ne_k^T\tilde{V}_0(\beta-\beta_0) - 2n^{1/2}\sigma e_k^T\tilde{V}_0^{1/2}\varepsilon + R_m \\
&= 2ne_k^T\tilde{V}_0(\beta-\beta_0) - 2e_k^T V_n + R_m
\end{aligned}
$$

where R_m represents small order term and e_k is a d-dimensional vector with kth element being one and all others being zero. Since $\beta-\beta_0 = O_p\left(1/\sqrt{n}\right)$ and $V_n = O_p\left(\sqrt{n}\right)$, then,

$$
\frac{\partial\hat{S}(\beta,h)}{\partial\beta_k} = O_p\left(\sqrt{n}\right)
$$

and

$$
\begin{aligned}
\frac{\partial Q\left(\left(\beta_1^T,\beta_2^T\right)^T,\hat{g}\right)}{\partial\beta_k} &= \frac{1}{2}\frac{\partial\hat{Q}_1(\beta,h)}{\partial\beta_k} + n\Psi'_{\zeta_n}(|\beta_k|)\,sgn(\beta_k) \\
&= n\zeta_n\left[O_p\left(\frac{1}{\sqrt{n}\zeta_n}\right) + \frac{\Psi'_{\zeta_n}(|\beta_k|)}{\zeta_n}sgn(\beta_k)\right].
\end{aligned}
$$

Since $\sqrt{n}\zeta_n\to\infty$ and $\lim\inf_{n\to\infty,\beta_k\to 0^+}\frac{\Psi'_{\zeta_n}(|\beta_k|)}{\zeta_n} > 0$, the sign of $\frac{\partial Q}{\partial\beta_k}$ is determined by the sign of β_k. It follows from Part (a) that

$$
\left.\frac{\partial Q\left(\left(\beta_1^T,\beta_2^T\right)^T,\hat{g}\right)}{\partial\beta}\right|_{\beta=\binom{\hat{\beta}_1}{0}} = 0
$$

and

$$
\frac{1}{2}\frac{\partial\hat{S}\left(\left(\hat{\beta}_1,0\right),h\right)}{\partial\beta_1} + n\Delta\Psi_{\zeta_n}^{d_1} = 0
$$

where $\Delta\Psi_{\zeta_n}^{d_1} = \{\Psi'_{\zeta_n}(|\beta_1|)\,sgn(\beta_1),\ldots,\Psi'_{\zeta_n}\left(|\beta_{d_1}|\right)sgn(\beta_{d_1})\}^T$. Note that as $n\to\infty$ and $\zeta_n\to 0$, $\Psi'_{\zeta_n}(|\beta_k|) = 0$ for $k = 1,\ldots,d_1$ and

$$
\frac{1}{2}\frac{\partial\hat{S}\left(\left(\hat{\beta}_1,0\right),h\right)}{\partial\beta_1} = 0,
$$

which implies that

$$
n\widetilde{V}_{10}\left(\hat{\beta}_1-\beta_{10}\right) - n^{1/2}\sigma\widetilde{V}_{10}^{1/2}\mathbf{e}_1 + o_p(1) = 0
$$

$$
\begin{aligned}
\sqrt{n}\left(\hat{\beta}_1-\beta_{10}\right) &= \widetilde{V}_{10}^{-1}(1/\sqrt{n})V_{1n} \\
&= \widetilde{V}_{10}^{-1}(1/\sqrt{n})\sum_{i=1}^{n}(Z_{1i}-E(Z_{1i}|\beta_{10}^T Z_{1i}))\dot{g}^T(\beta_{10}^T Z_{1i})X_i\varepsilon_{1i}
\end{aligned}
$$

so that

$$
\sqrt{n}\left(\hat{\beta}_1-\beta_{10}\right)\rightarrow^D N\left(0,\widetilde{V}_{10}^{-1}\sum_{\ell=-\infty}^{\infty}\Gamma(\ell)\widetilde{V}_{10}^{-1}\right)
$$

where $\Gamma(\ell) = E\Gamma_i\Gamma_{i-\ell}^T$ with $\Gamma_i = \left(Z_{1i}-E(Z_{1i}|\beta_{10}^T Z_{1i})\right)\dot{g}^T(\beta_{10}^T Z_{1i})X_i\varepsilon_{1i}$. □

References

Akaike, H., 1973. Maximum likelihood identification of gaussian autoregressive moving average models. Biometrika 60, 255–265.

Box, G.E.P., Jenkins, G.M., 1970. Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco.

Breiman, L., 1995. Better subset regression using the nonnegative garrote. Technometrics 37, 373–384.

Brent, A.J., Lin, D.Y., Zeng, D., 2008. Penalized estimating functions and variable selection in semiparametric regression models. J. Amer. Statist. Assoc. 103, 672–680.

Cai, Z., 2002. Regression quantiles for time series. Econometric Theory 18, 169–192.

Cai, Z., Fan, J., Li, R., 2000a. Efficient estimation and inferences for varying coefficient models. J. Amer. Statist. Assoc. 95, 888–902.

Cai, Z., Fan, J., Yao, Q., 2000b. Functional-coefficient regression models for nonlinear time series. J. Amer. Statist. Assoc. 95, 941–956.

Cai, Z., Ren, Y., Yang, B., 2014a. A semiparametric conditional capital asset pricing model. Working paper, The Wang Yanan Institute for Studies in Economics, Xiamen University.

Cai, Z., Wang, Y., 2014. Testing predictive regression models with nonstationary regressors. J. Econometrics 178, 4–14.

Cai, Z., Wang, Y., Wang, Y., 2014b. Testing instability in predictive regression model with nonstationary regressors. Econometric Theory, (forthcoming). http://dx.doi.org/10.1017/S0266466614000590.

Campbell, J.Y., Yogo, M., 2006. Efficient tests of stock return predictability. J. Financ. Econ. 81, 27–60.

Chan, K.S., Tong, H., 1986. On estimating thresholds in autoregressive models. J. Time Ser. Anal. 7, 179–190.

Chen, R., Tsay, R.S., 1993. Functional coefficient autoregressive model. J. Amer. Statist. Assoc. 88, 298–308.

Fan, J., Li, R., 2001. Variable selection via non-concave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 96, 1348–1360.

Fan, J., Li, R., 2004. New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. J. Amer. Statist. Assoc. 99, 710–723.

Fan, J., Lv, J., 2010. A selective overview of variable selection in high dimensional feature space. Statist. Sinica 20, 101–148.

Fan, J., Yao, Q., Cai, Z., 2003. Adaptive varying-coefficient linear models. J. R. Stat. Soc. Ser. B 65, 57–80.

Fan, J., Zhang, W.Y., 1999. Statistical estimation in varying coefficient models. Ann. Statist. 27, 1491–1518.

Fu, W.J., 1998. Penalized regressions: the bridge versus the LASSO. J. Comput. Graph. Statist. 7, 397–416.

Granger, C.W.J., Andersen, A.P., 1978. An Introduction to Bilinear Time Series Models. Vanderhoek and Ruprecht, Gottingen.

Hamilton, J.D., 1989. A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica 57, 357–384.

Horowitz, J.L., 2009. Semiparametric and Nonparametric Methods in Econometrics. Springer-Verlag, New York.

Huang, J., Joel, L.H., Wei, F.R., 2010. Variable selection in nonparametric additive models. Ann. Statist. 38, 2282–2313.

Hunter, D.R., Li, R., 2005. Variable selection using MM algorithms. Ann. Statist. 33, 1617–1642.

Ichimura, H., 1993. Semiparametric least squares (SLS) and weighted SLS estimation of single index model. J. Econometrics 58, 71–120.

Janson, S., 1987. Maximal spacing in several dimensions. Ann. Probab. 15, 274–280.

Kong, E., Xia, Y., 2007. Variable selection for the single index model. Biometrika 94, 217–229.

Li, Q., Jeffrey, S.R., 2007. Nonparametric Econometrics: Theory and Practice. Princeton University Press, Princeton.

Li, R., Liang, H., 2008. Variable selection in semiparametric regression modeling. Ann. Statist. 36, 261–286.

Liang, H., Li, R., 2009. Variable selection for partially linear models with measurement errors. J. Amer. Statist. Assoc. 104, 234–248.

Liang, H., Liu, X., Li, R., Tsai, C.L., 2010. Estimation and testing for partially linear single-index models. Ann. Statist. 38, 3811–3836.

Lin, Y., Zhang, H., 2006. Component selection and smoothing in multivariate nonparametric regression. Ann. Statist. 34, 2272–2297.

Newey, W.K., Stoker, T.M., 1993. Efficiency of weighted average derivative estimators and index models. Econometrica 61, 1199–1223.

Phillips, P.C.B., Lee, J.H., 2013. Predictive regression under various degrees of persistence and robust long-horizon regression. J. Econometrics 177, 250–264.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Statist. 6, 461–464.

Su, L., Zhang, Y., 2013. Variable selection in nonparametric and semiparametric regression models. In: Handbook in Applied Nonparametric and Semi-Nonparametric Econometrics and Statistics. Research Collection School of Economics.

Teräsvirta, T., 1994. Specification, estimation, and evaluation of smooth transition autoregressive models. J. Amer. Statist. Assoc. 89, 208–218.

Tibshirani, R., 1996. Regression shrinkage and selection via the LASSO. J. R. Stat. Soc. Ser. B 58, 267–288.

Tong, H., 1990. Non-linear Time Series: A Dynamical System Approach. Oxford University Press, Oxford, UK.

Wang, H., Li, G., Tsai, C.L., 2007. Regression coefficient and autoregressive order shrinkage and selection via LASSO. J. R. Stat. Soc. Ser. B 69, 63–68.

Wang, L.F., Li, H.Z., Huang, J.H., 2008. Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. J. Amer. Statist. Assoc. 103, 1556–1569.

Wang, H., Xia, Y., 2009. Shrinkage estimation of the varying coefficient model. J. Amer. Statist. Assoc. 104, 747–757.

Xia, Y., Li, W.K., 1999. On single-index coefficient regression models. J. Amer. Statist. Assoc. 94, 1275–1285.

Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. J. R. Stat. Soc. Ser. B 68, 49–57.

Zhang, Y., Li, R., Tsai, C.L., 2010. Regularization parameter selections via generalized information criterion. J. Amer. Statist. Assoc. 105, 312–323.

Zhang, H.H., Lin, Y., 2006. Component selection and smoothing for nonparametric regression in exponential families. Statist. Sinica 16, 1021–1041.

Zhao, P.X., Xue, L., 2010. Variable selection for semi-parametric varying coefficient partially linear errors-in-variables models. J. Multivariate Anal. 101, 1872–1883.

Zou, H., 2006. The adaptive LASSO and its oracle properties. J. Amer. Statist. Assoc. 101, 1418–1429.

Zou, H., Li, R., 2008. One-step sparse estimates in non-concave penalized likelihood models. Ann. Statist. 36, 1509–1533.