INTERVAL ESTIMATION FOR SEMIPARAMETRIC PREDICTIVE REGRESSION

by

Shaoxin Hong

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Applied Mathematics

Charlotte

2018

Approved by:

Dr. Jiancheng Jiang

Dr. Zhiyi Zhang

Dr. Shaoyu Li

Dr. Weidong Tian

ABSTRACT

SHAOXIN HONG. Interval Estimation for Semiparametric Predictive Regression.
(Under the direction of DR. JIANCHENG JIANG)

Predictive regression is an important research topic in financial econometrics. Various estimation methods have been proposed for it, but they suffer from complicated asymptotic limits which depend on whether or not the predicting variable is stationary. This makes inference for the predictability difficult. In this paper we employ a nonlinear projection to deal with endogeneity of the state variable which results in a new semiparametric predictive regression model for describing the relationship between the state variables and the asset return. We propose a weighted profile estimation equation method to estimate the parameters and an empirical likelihood ratio test to examine the predictability of state variables. We establish the asymptotic normality of the proposed estimator and show the Wilks theorem holds for the test statistic regardless of predicting variables being stationary or not. This provides a unifying method for constructing confidence regions of the coefficients of state variables. Simulations demonstrate favorable finite sample performance of the proposed method over some existing approaches. Real examples illustrate the value of our methodology.

ACKNOWLEDGMENTS

Foremost, I would like to express my sincere gratitude to my advisor Dr. Jiancheng Jiang for the continuous support of my Ph.D study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D study.

I would also like to thank my committee members, Dr. Zhiyi Zhang, Dr. Shaoyu Li and Dr. Weidong Tian for their encouragement and insightful comments.

I gratefully acknowledge the financial support received towards my Ph.D. from the Graduate School and Mathematics department. Thanks to Professor Joel Avrin and Shaozhong Deng for their encouragement and supervisory role and valuable input.

Last but not the least, I would like to thank my family: my parents Diaodong Hong and Xiuyi Dai, for giving birth to me at the first place and supporting me spiritually throughout my life.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

As an important research topic in economics and finance, the predictability of stock returns has been studied in decades. In many financial applications, for example, the conditional capital asset pricing, the mutual fund performance, and the optimal asset allocations, the predictability problem is routinely examined. An enormous amount of empirical research effort demonstrates the predictability of stock returns using various lagged financial variables, such as the book-to-market ratio, the dividend yield, the dividend-price (D-P) ratio, the earning-price (E-P) ratio, the interest rates, and the term spread and default premia, among others. An essential question is often asked about whether the returns are predictable in a specific financial application. Because many of the predictive financial variables are highly persistent and even nonstationary, it is challenging to answer this question.

In the literature, many works have been devoted to addressing the above question through predictive regression. Let y_t be the predictable variable, say excess stock returns, in period t, and let x_{t-1} be the state variable, say log D-P ratio, in period $t-1$. A nice framework of the predictive regression is

$$y_t = \beta_0 + \beta_1 x_{t-1} + \varepsilon_t, \qquad x_t = \rho x_{t-1} + u_t, \qquad 1 \le t \le n, \tag{1.1}$$

where $E(u_t|x_{t-1}) = 0$, but $E(\varepsilon_t|x_{t-1})$ may be nonzero. In many applications, the correlation between innovations ε_t and u_t is nonzero (Table 1 in Torous et al., 2004;

Table 4 in Campbell and Yogo, 2006), which brings the nonzero correlation between x_{t-1} and ε_t and creates the so-called "endogeneity". Hence, directly regress y_t on x_{t-1} may yield a biased ordinary least squares (OLS) estimator of β. The parameter ρ is the unknown degree of persistence of variable x_t. When $|\rho| < 1$, x_t is stationary (Viceira, 1997; Amihud and Hurvich, 2004; Amihud et al., 2009); when $\rho = 1$, it is first order integrated (I(1) for short); when $\rho = 1 + c/n$ with $c < 0$, it is local-to unity or nearly first order integrated (NI(1) for short); when $\rho = 1 + c/n$ with $c > 0$, it is mildly explosive. See Elliott and Stock (1994), Cavanagh, Elliott, and Stock (1995), Phillips (1988), Torous, Valkanov, and Yan (2004), Campbell and Yogo (2006), Polk, Thompson, and Vuolteenho (2006), Rossi (2007), and Cai and Wang (2014), etc. These references validate that predictor x_t can be stationary or nonstationary and highly persistent. This brings much difficulty in modeling. In the following let us first review three common estimation strategies for model (1.1) and then introduce our new methodology for alleviating the modeling difficulty.

The first one is the bias correction approach using information conveyed by the AR(1) process of x_t. For example, Kothari and Shanken (1997) and Stambaugh (1999) suggested the first order bias-corrected OLS estimation, Amihud and Hurvich (2004) proposed the second order bias-correction method, and Lewellen (2004) studied the conservative bias-correction vehicle which assumes the true autoregressive coefficient of AR(1) to be close to one. The second approach is the maximum likelihood estimation in Campbell and Yogo (2006), which assumes that innovations $\{(\varepsilon_t, u_t)\}$ are independently distributed bivariate normal $N(0, \Sigma)$. The third one is based on linear projection and least squares in Amihud and Hurvich (2004), Amihud et al.

(2009), and Cai and Wang (2014). In such a way, the endogeneity may be removed from the model by the projection of ε_t onto u_t. For $|\rho| < 1$, using the linear projection of ε_t onto u_t, $\varepsilon_t = \gamma\, u_t + v_t$, Amihud and Hurvich (2004) reexpressed model (1.1) as

$$y_t = \beta_0 + \beta_1 x_{t-1} + \gamma u_t + v_t, \tag{1.2}$$

where v_t is white noise independent of x_t and u_t at all leads and lags. If u_t were known, the error in model (1.2) would satisfy the classical assumption of OLS without endogeneity. They applied the two-stage least squares regression (TSR) method as follows: first obtain the OLS estimator $\hat{\rho}$ of ρ, then calculate the fitted residuals $\hat{u}_t$, and finally regress y_t on x_{t-1} and $\hat{u}_t$ to obtain an estimate of β. For $\rho = 1 + c/n$ with $c \leq 0$, Cai and Wang (2014) investigated the TSR method and established its limiting distribution. These works show that the limiting distributions of the estimator of β are different for I(0), I(1) and NI(1) cases, which makes inference for β difficult, since one has to decide which limiting distribution is used for inference. In particular, when ρ is close to one, the variance estimate of the limiting distribution behaves erratically (Chan, Li and Peng, 2012; Zhu, Cai and Peng, 2014). Hence, it is desired to establish a unifying inference tool for β. In addition, the above works cannot assess joint predictability of multiple state variables and thus may suffer from severe under-fitting problems.

To surmount the above difficulty and to cope with multiple predictive variables simultaneously, in this paper we employ a nonlinear projection of ε_t onto u_t, which results in a new semiparametric predictive regression model. This model contains the above models and incorporates multiple state variables. As indicated above, it is de-

sired to develop a unifying inference tool which does not need to choose which limiting distribution to use. To this end, we propose a weighted estimation equation method for the proposed model. This approach gives us a unifying limiting distribution of the resulting estimators no matter if the state variables are stationary, I(0), I(1), NI(1), or slightly explosive. Theoretically, we will establish the asymptotic normality distribution of the proposed estimator.

Similar weighted estimation was previously studied in the literature, including the Cauchy estimator in So and Shin (1999), the weighted absolute deviation estimation in Ling (2005) and Chan and Peng (2005), the weighted least squares estimation in Chan, Li and Peng (2012) and Zhu, Cai and Peng (2014). However, these works can deal with linear predictive regression with only one predictor. Further, the resulting estimator is not efficient for I(0) or stationary processes because of the effect of weights. Our weighted method borrows the strength of the previous techniques which provide a unifying limiting distribution, but is different from them. On one hand, our method can deal with multiple state variables (see Theorems 1-2 and Example 5.3). On the other hand, it leads to the most efficient estimator with probability going to one when the state variable is I(0) or stationary under the mixing condition.

Since our aim is to check predictability of state variables, we have to construct confidence region of β. Our asymptotically normal distribution furnishes us a Wald type of statistic for the confidence region, but it requires estimating the asymptotic variance matrix of our estimator. In our experience and also noted in the literature (Chan, Li, Peng, 2012; Fu, Cai and Peng, 2014), this variance estimator behaves irregularly for nonstationary cases. This motivates us to propose an empirical likelihood

method for constructing the confidence region, based on our weighted estimation equations. The empirical likelihood was initially introduced by Owen (1988). It has been demonstrated as a powerful nonparametric tool for interval estimates. See Owen (2001) for an overview. This method has many advantages over the normal approximation-based method and the bootstrap method for constructing confidence intervals, such as the transformation respecting, the range of parameter respecting, and no predetermined shape requirement (Hall and La Scala, 1990; Hall, 1992).

The remainder of this article is organized as follows. In Section 2, we introduce our semiparametric predictive regression model and the corresponding estimation procedures. In Section 3, we establish our unified asymptotic distribution of the proposed estimator. In Section 4, we present the empirical likelihood ratio for constructing confidence region and derive the Wilks theorem. In Section 5, we conduct extensive simulations to compare finite sample performance of our method with others. In Section 6, real examples are used to illustrate the value of the proposed methodology. Proofs of our theorems are relegated to the Appendix.

CHAPTER 2: SEMIPARAMETRIC PREDICTIVE MODELS

The linear projection method for motivating model (1.2) assumes that $E[\varepsilon_t \,|\, u_t] = \gamma u_t$. This may not hold in general. Using nonlinear projection $\varepsilon_t = \phi(u_t) + v_t$, we extend model (1.2) to

$$y_t = \beta x_{t-1} + \phi(u_t) + v_t, \tag{2.1}$$

where $\phi(\cdot)$ is an unknown function. However, u_t is not observed. To estimate the parameters, one needs to use observed residual $\hat{u}_t = x_t - \hat{\rho} x_{t-1}$ to replace u_t, and then estimate the parameters in the model. However, the resulting estimators of parameters are difficult to study, because their distributions depend on that of $\hat{u}_t$. Instead of model (2.1) one may consider the model

$$y_t = \beta x_{t-1} + \phi(z_t) + v_t$$

with $\{z_t\}$ being an observable time series. In this model, the predicted variable x_t is one-dimensional, so it cannot be used to study joint effects of multiple predictors. In theory and practice, it may be more sensible to incorporate multiple predictors for correcting model misspecification and for the fact: the predicted variables are usually correlated and as in multiple linear models a variable may not have predictability after other useful correlated variables are taken into account. Therefore, in this paper we

study the following semiparametric model:

$$y_t = x_{t-1}^\top \beta + \theta(z_t) + v_t, \quad x_t = (x_{t,1}, \ldots, x_{t,k})^\top, \tag{2.2}$$

where $\theta(\cdot)$ is a unknown function, z_t is a stationary time series, v_t is white noise uncorrelated with x_{t-1} and z_t, and $x_{t,i}$ is a stationary process or an AR(1) process satisfying that $x_{t,i} = \rho_i\, x_{t-1,i} + u_{t,i}$, with $u_{t,i}$ being a martingale difference, and $|\rho_i| < 1 \quad \text{or} \quad \rho_i = 1 + \gamma_i/n, \quad \gamma_i \in \mathcal{R}$, for $i = 1, \ldots, k$. When $|\rho_i| < 1$, $x_{t,i}$ is called an I(0) process; when $\rho_i = 1$, it is an I(1) process in the sense that its first order differences are I(0); when $\rho_i = 1 + \gamma_i/n$ and $\gamma_i < 0$, it is NI(1); when $\rho_i = 1 + \gamma_i/n$ and $\gamma_i > 0$, it is slightly explosive, which is also called the NI(1) case to avoid abuse of notation since ρ_i is nearly one. In this setting, x_t has finite variance or infinite variance. The interesting parameters are vector β.

Model (2.2) contains many existing models studied in the literature. When $\{x_t, y_t, z_t\}$ are independent and identically distributed, it reduces to the semiparametric model widely studied in statistics (Speckman, 1988; Carroll et al., 1997). When $\theta(\cdot) = 0$ and $x_{t-1} = y_{t-1}$, it is the AR(1) model studied in Chan and Wei (1987), Phillips and Han (2008), Chan and Zhang (2009), and Chan, Li and Peng (2012). For univariate x_t, when $\theta(\cdot)$ is linear and $z_t = \hat{u}_t$, it becomes the model studied in Cai and Wang (2014). Since $u_t = x_t - \rho x_{t-1}$, model (1.2) is equivalent to

$$y_t = \gamma_0 + \gamma_1 x_{t-1} + \gamma_2 x_t + v_t. \tag{2.3}$$

Hence, model (2.3), the equivalent of model (1.2) studied in Amihud and Hurvich (2004), is also a special case of model (2.2) with $\theta(\cdot) = 0$. These indicate that our

proposed model is an important family of predictive regression models.

2.1 Profile Least Squares Estimation

There are different approaches to estimating the unknown parameters in partly linear regression model (2.2) in i.i.d. cases. The profile least squares method is a useful one and is semiparametrically efficient (Carroll et al., 1997). We now employ it in the current setting. Specifically, for any given β, model (2.2) becomes

$$y_t^* = \theta(z_t) + v_t, \tag{2.4}$$

where $y_t^* = y_t - x_{t-1}^\top \beta$. The local linear regression technique is then applied to estimate the function $\theta(\cdot)$ by minimizing

$$\sum_{t=1}^{n} [y_t^* - a - b(z_t - z)]^2 K_h(z_t - z)$$

over (a, b), where $K_h(\cdot) = h^{-1}K(\cdot/h)$ with $K(\cdot)$ being a kernel function and h being a bandwidth used to control the amount of data in smoothing. The resulting estimate of $\theta(z)$ admits the close form (Fan and Gijbels, 1996)

$$\hat{\theta}(z;\beta) = \hat{a} = \sum_{t=1}^{n} w_t(z) y_t^* / \sum_{t=1}^{n} w_t(z) \equiv \sum_{t=1}^{n} \xi_t(z) y_t^*,$$

where $w_t(z) = K_h(z_t - z)\{S_{n,2}(z) - (z_t - z)S_{n,1}(z)\}$ with $S_{n,j}(z) = n^{-1}\sum_{t=1}^{n} K_h(z_t - z)(z_t - z)^j$, and $\xi_t(z) = w_t(z)/\sum_{t=1}^{n} w_t(z)$. Note that $\sum_{t=1}^{n} \xi_t(z) = 1$. Substituting $\hat{\theta}(z_t;\beta)$ into (2.2), we obtain

$$\sum_{s=1}^{n} \xi_s(z_t)(y_t - y_s) \approx \beta^\top \sum_{s=1}^{n} \xi_s(z_t)(x_{t-1} - x_{s-1}) + v_t, \;\; t = 1, \ldots, n. \tag{2.5}$$

Let

$$\tilde{x}_{t-1} = x_{t-1} - \sum_{s=1}^{n} \xi_s(z_t) x_{s-1} \tag{2.6}$$

and $\tilde{y}_t = y_t - \sum_{s=1}^{n} \xi_s(z_t) y_s$. Then the final estimate of β is then the least-squares estimate

$$\hat{\beta} = (\sum_{t=1}^{n} \tilde{x}_{t-1}^{\otimes 2})^{-1} \sum_{t=1}^{n} \tilde{x}_{t-1} \tilde{y}_t, \tag{2.7}$$

where $a^{\otimes 2} = aa^{\top}$ for any matrix a. Function $\theta(z)$ can simply be estimated by $\hat{\theta}(z; \hat{\beta})$.

It is interesting to investigate asymptotic properties of the above estimators, However, when $\{x_t\}$ is $NI(1)$, $I(1)$ or mildly explosive, the asymptotic behaviors should be different from the traditional ones (see Cai, Li and Park, 2007). As expected, $\hat{\beta}_i$ should be n-consistent if $x_{t,i}$ is I(1) or NI(1) and $\sqrt{n}$-consistent for I(0) or stationary cases (Cai and Wang, 2014), and $\hat{\theta}(\cdot)$ should be $\sqrt{nh}$-consistent. In classical settings where $\{x_t\}$ is stationary, under the mixing condition, $\hat{\beta}$ is also $\sqrt{n}$-consistent. This again creates difficulty in statistical inference for β, since the above complicated limiting properties of $\hat{\beta}$ depend on whether the predicted variable is stationary or not. In other words, it leads to different reference distributions for hypothesis testing problems about β. Hence, one needs to judge stationarity of each component of x_t before conducting hypothesis testing. Alternatively, one can use a bootstrap procedure to obtain critical values, but the full sample bootstrap method is inconsistent in NI(1) or infinite variance settings (Datta, 1996; Hall and Jing, 1998). This motivates us to suggest a weighted estimation equation procedure for the proposed model.

2.2 Weighted Profile Score Equation Estimation

The main idea of weighted estimation is to weigh down the contribution of those observations according to the value of the corresponding predictor so as to eliminate the effect of infinite variance and non-stationarity on the limiting distribution. To illustrate our weighting scheme, we begin with the profile least squares score equations for β.

Note that (2.5) can be rewritten as

$$v_t \approx y_t - x_{t-1}^\top \beta - \hat{\theta}(z_t;\beta) = \tilde{y}_t - \tilde{x}_{t-1}^\top \beta. \tag{2.8}$$

Then the score equations for the profile least squares estimation for β are

$$\sum_{t=1}^{n} \tilde{x}_{t-1}(\tilde{y}_t - \tilde{x}_{t-1}^\top \beta) = 0. \tag{2.9}$$

Solving the above equations results in the least squares estimator $\hat{\beta}$ in (2.7). However, as noted before, the limiting distributions of $\hat{\beta}$ are different in stationary and nonstationary cases.

For simple predictive regression with $k = 1$, it was shown that the weighted least squares estimation could remove the effect of infinite variance of a predictor (Ling, 2005) and the effect of nonstationarity of a predictor (Chan, Li and Peng, 2012; Zhu, Cai and Peng, 2014), but for model (2.2) with multiple predictors, the weighted least squares estimation does not work. This motivates us to propose the following weighted profile estimation equations:

$$\sum_{t=1}^{n} \Omega_t \tilde{x}_{t-1}(\tilde{y}_t - \tilde{x}_{t-1}^\top \beta) = 0, \tag{2.10}$$

where $\Omega_t = \text{diag}(\omega_{t,1}, \ldots, \omega_{t,k})$ is a sequence of non-negative definite diagonal matrices to be chosen and is used to weight down the contributions of data points to the score equations in (2.9). Solving the above equations leads to our weighted profile score equation estimator of β:

$$\hat{\beta}_\omega = (\sum_{t=1}^{n} \Omega_t \tilde{x}_{t-1}^{\otimes 2})^{-1} \sum_{t=1}^{n} \Omega_t \tilde{x}_{t-1} \tilde{y}_t. \tag{2.11}$$

With the estimator of β at hand, we estimate $\theta(z)$ by $\hat{\theta}_\omega(z) = \hat{\theta}(z; \hat{\beta}_\omega)$.

It is worthy to point out that $\hat{\beta}_\omega$ is not any kind of weighted least squares estimator of model (2.8) even though it is motivated from the score equations in (2.9). In constructing $\hat{\beta}_\omega$, we used the local linear estimator $\hat{\theta}(z; \beta)$, which is not a weighted one. This is very critical.

The above method requires specifying $\omega_{t,i}$ for $i = 1, \ldots, k$. When $k = 1$, one viable choice for ω_t is

$$\omega_t^* = (\delta + x_{t-1}^2)^{-1/2}, \tag{2.12}$$

where δ is a nonnegative constant, which was previously suggested by several authors. So and Shin (1999) used ω_t^* with $\delta = 0$ and obtained the Cauchy estimators of AR(1) models. Chan, Li and Peng (2012) employed the weight for united interval estimation of AR(1) models and found that $\delta = 1$ works well in general. Zhu, Cai and Peng (2014) applied the weight with $\delta = 1$ for united interval estimation of simple linear predictive regression. When the underlying process x_t is stationary and satisfies the mixing condition, the weight sequence in (2.12) cannot lead to efficient estimation, since the unweighted least squares principle is optimal. That is, for the stationary

cases one should not employ the weighting scheme. This motivates us to propound the following weights for $i = 1, \ldots, k$:

$$\omega_{t,i} = \begin{cases} 1 & \text{if} \quad i \in \mathcal{I}, \\ (1 + \|\tilde{x}_{t-1,\mathcal{I}^c}\|^2)^{-1/2} & \text{otherwise}, \end{cases}$$

where $\mathcal{I} = \{i : \max_{1 \le t \le n} n^{-\frac{1}{2}} \log(n) |\tilde{x}_{t-1,i}| < c^*\}$, $\tilde{x}_{t-1,\mathcal{I}^c}$ is the subvector of $\tilde{x}_{t-1}$ with indexes not in $\mathcal{I}$, and c^* is a positive constant chosen to maximize the efficiency of $\hat{\beta}_\omega$. See Section 3. When $n^{-\frac{1}{2}} \log(n) \max_t |\tilde{x}_{t-1,i}| \ge c^*$, we set $\omega_{t,i}$ like the one with $\delta = 1$ in (2.12), which is used to control the contribution of the ith score equation in (2.10). For stationary cases with more than 2nd moments, given a positive constant c^*, we generally have

$$P(\max_{1 \le t \le n} n^{-\frac{1}{2}} \log(n) |\tilde{x}_{t-1,i}| < c^*) \to 1, \text{ as } n \to \infty.$$

Hence, we set $\omega_{t,i}$ be one if $n^{-\frac{1}{2}} \log(n) \max_t |\tilde{x}_{t-1,i}| < c^*$, so that it leads to the unweighted least squares score equation with probability going to one. If all components of x_t are stationary and mixing and have more than second moments, then with probability going to one all score equations in (2.10) become the profile score equations in (2.9) and the resulting estimator should be semiparametrically efficient.

Our weight sequence involves constant c^*. Different values of c^* lead to different estimators. For a given sample, as c^* gets small enough, it leads to the weight in Ling (2005), Chan, Li and Peng (2012), and Zhu, Cai and Peng (2014) when $k = 1$; as c^* gets sufficiently large, all weights are equal to one, which results in unweighted least squares estimator $\hat{\beta}$. Therefore, our weighted score estimation bridges the gap

between the previous weighting estimation and the ordinary least squares method. In practice, one can choose c^* to maximize efficiency of $\hat{\beta}_\omega$. We will explore this choice later. For the purpose of minimizing prediction errors, one can regard c^* as tuning parameter and choose it by cross validation.

CHAPTER 3: ASYMPTOTICS OF THE WEIGHTED ESTIMATION

Throughout the paper, "$\Rightarrow$"denotes weak convergence and "$\xrightarrow{D}$" represents convergence in distribution. Let $U_{n,i}(r) = n^{-1/2}x_{[nr],i}$, where $r = t/n$ and $[x]$ denotes the integer part of x. Then, under general conditions on $u_{t,i}$, for example, the regularity ones in Phillips (1988), it holds that

$$U_{n,i}(r) \Rightarrow U_{\gamma_i}(r) \tag{3.1}$$

as $n \to \infty$, where $U_{\gamma_i}(r) = \int_0^r \exp\{(r-s)\gamma_i\}\, dW_u^{(i)}(s)$ is a diffusion process and $W_u^{(i)}(s)$ is a one-dimensional Brownian motion with variance $\sigma_{u,i}^2 = \mathrm{Var}(u_{1,i}) + 2\sum_{s=2}^{\infty} E(u_{1,i}u_{s,i})$. As demonstrated in Lemma A.1, the weak convergence in (3.1) can be strengthened to a strong one, which is key to the derivation of our theoretic results.

The following notation and regularity conditions are needed for our asymptotic results.

(A_0) For $i = 1, \ldots, k$, if $x_{t,i}$ is not an AR(1) process, then it is ρ-mixing with mixing coefficients $\rho_i^*(s)$ satisfying $\sum_\ell \rho_i^*(\ell) < \infty$ or is α-mixing with mixing coefficients $\alpha_i^*(s)$ satisfying $\sum_\ell \ell^a \{\alpha_i^*(\ell)\}^b < \infty$, for some $0 < b < 1$ and $a > b$.

(A_1) For $i = 1, \ldots, k$, if $x_{t,i}$ is an AR(1) process, then $E(u_{0,i}) = 0$, $E|u_{t,i}|^{k_1+k_2} < \infty$ for some $k_1 > 2$ and $k_2 > 0$, and $\{u_{t,i}\}_{t=0}^{\infty}$ is α-mixing with mixing coefficients

$\alpha_i(s)$ satisfying $\sum_{s=1}^{\infty}\{\alpha_i(s)\}^{1-2/k_1} < \infty$.

(A_2) Bandwidth h satisfies that $h \to 0$ and $nh \to \infty$, as n goes to ∞.

(A_3) Kernel function $K(\cdot)$ is a continuous density with bounded support $[0, 1]$. Let $\mu_i = \int_{-1}^{1} u^i K(u)\, du$ and $\nu_i = \int_{-1}^{1} u^i K^2(u)\, du$.

(A_4) Time series $\{z_t\}$ is stationary and has continuous stationarity density $f(z)$ with bounded support $supp(f)$.

(A_5) Function $\theta(z)$ has a continuous second order derivative $\theta''(z)$ for $z \in supp(f)$.

(A_6) The errors $\{v_t\}$ are white noises with variance $\sigma_v^2 = \mathrm{Var}(v_t)$ and satisfy that $E|v_t|^{2+\delta} < \infty$ for some $\delta > 0$.

(A_7) If $x_{t,i}$ is stationary or an AR(1) process with$|\rho_i| < 1$, then the condition expectation $\alpha_i(z) = E(x_{t-1,i}|z_t = z)$ has continuous second order derivative and the conditional variance $\sigma_i(z) = \mathrm{Var}(x_{t-1,i}|z_t = z)$ is continuous on $z \in supp(f)$. Furthermore, $E|x_{t,i}|^4 < \infty$.

Condition (A_0) is very general for a stationary process and assumed in Jiang and Mack (2001), condition (A_1) was used in Phillips (1988), conditions (A_2)-(A_5) are standard in local smoothing, and conditions (A_6)-(A_7) are mild. The finite fourth moment in (A_7) can be relaxed to a $2 + \delta$ moment for achieving robustness but with some efficiency loss as in traditional robustness estimation.

The following theorem sets up a unifying limit of the proposed estimator.

Theorem 3.1. Suppose model (2.2) holds with $x_{t,i}$ being stationary or an AR(1) process with $|\rho_i| < 1$ or $\rho_i = 1 + \gamma_i/n$ for some $\gamma_i \in \mathcal{R}$. Under the conditions of (A_0)-(A_7), if $nh^4 \to 0$, then

$$\{\sum_{t=1}^{n}(\Omega_t\tilde{x}_{t-1})^{\otimes 2}\}^{-1/2}(\sum_{t=1}^{n}\Omega_t\tilde{x}_{t-1}^{\otimes 2})(\hat{\beta}_\omega - \beta_0) \xrightarrow{D} N(0, \sigma_v^2 I_k),$$

where β_0 is the true value of β and I_k is a $k \times k$ identity matrix.

Corollary 1. For $i = 1, \ldots, k$, if $x_{t,i}$ is a stationary process satisfying condition (A_0) and $E(x_{t,i}^{2+\delta}) < \infty$, then $n^{-\frac{1}{2}}\log(n)\max_{1\le t\le n}|\tilde{x}_{t-1,i}| = o_p(1)$ and $P\{\omega_{t,i} = 1 \text{ for all } t\} \to 1$, as $n \to \infty$. This indicates that, when all $x_{t,i}$ satisfy condition (A_0) and have more than second moments, with probability going to one, $\hat{\beta}_\omega$ becomes the unweighted profile least squares estimator β, which is semiparametrically efficient (Carroll et al., 1997).

Remark By Theorem 3.1, $\hat{\beta}_\omega$ is asymptotically unbiased. Its variance can be approximated by

$$[\{\sum_{t=1}^{n}(\Omega_t\tilde{x}_{t-1})^{\otimes 2}\}^{-1/2}(\sum_{t=1}^{n}\Omega_t\tilde{x}_{t-1}^{\otimes 2})^2\{\sum_{t=1}^{n}(\Omega_t\tilde{x}_{t-1})^{\otimes 2}\}^{-1/2}]^{-1}\sigma_v^2.$$

Since σ_v^2 does not depend on c^*, we choose c^* to maximize the efficiency or equivalently the generalized variance of $\hat{\beta}_\omega$. Let $\hat{c}^* = \arg\max_{c^*} D_n(c^*)$, where

$$D_n(c^*) = \|\{\sum_{t=1}^{n}(\Omega_t\tilde{x}_{t-1})^{\otimes 2}\}^{-1/2}\big(\sum_{t=1}^{n}\Omega_t\tilde{x}_{t-1}^{\otimes 2}\big)\|_F,$$

where $\|\cdot\|_F$ is the Hilbert norm of a matrix.

Using the limiting distribution in Theorem 3.1, one can construct a Wald type of confidence region of β. This requires estimating σ_v^2 by the the standard error of the

model

$$\hat{\sigma}_v^2 = n^{-1} \sum_{t=1}^{n} \{y_t - x_{t-1}^\top \hat{\beta}_\omega - \hat{\theta}_\omega(z_t)\}^2.$$

However, in our experience and also as noted in Chan, Li and Peng (2012), such an interval estimate of β has unacceptable coverage probabilities for NI(1), I(1) and explosive cases, since $\hat{\sigma}_v^2$ behaves erratically. As pointed out in Zhu, Cai and Peng (2014), a bootstrap method may be used for the interval estimate, but the full sample bootstrap method is inconsistent for an NI(1) and infinite variance AR process. The problem may be solved by employing the subsample bootstrap method, but the subsample size is difficult to choose (Hall and Jing, 1998; Datta, 1996). For this approach, theoretical justification seems difficult, and associated computational burden is also heavy. This motivates us to consider the empirical likelihood confidence interval (Owen, 2001) in the next section.

CHAPTER 4: TESTING PREDICTABILITY

As we discussed before, an important application of model (2.2) is testing predictability of x_{t-1}. In the following we suggest an empirical likelihood method to construct a confidence region for β or test $H_0 : \beta = \beta_0$. This approach avoids estimating the asymptotic variance and works for stationary, non-stationary and infinite variance cases.

For $t = 1, \ldots, n$, let $Z_t(\beta) = \Omega_t \tilde{x}_{t-1}(\tilde{y}_t - \beta^\top \tilde{x}_{t-1})$. Then our weighted score equation estimator $\hat{\beta}_\omega$ solves the weighted score equation $\sum_{t=1}^n Z_t(\beta) = 0$. Following Owen (1990) and Chan, Li and Peng (2012), we define the empirical likelihood ratio as follows:

$$L(\beta) = \sup\Big\{\prod_{t=1}^n (np_t) : p_1 \geq 0, \ldots, p_n \geq 0, \sum_{t=1}^n p_t = 1, \sum_{t=1}^n p_t Z_t(\beta) = 0.\Big\}$$

Using the Lagrange multiplier technique, we obtain that $p_t = n^{-1}\{1 + \lambda^\top Z_t(\beta)\}^{-1}$. Then the log empirical likelihood ratio is

$$\ell(\beta) = -2 \log L(\beta) = 2 \sum_{t=1}^n \log\{1 + \lambda^\top Z_t(\beta)\},$$

where $\lambda = \lambda(\beta)$ satisfies

$$\sum_{t=1}^n Z_t(\beta)/\{1 + \lambda^\top Z_t(\beta)\} = 0. \tag{4.1}$$

Note that the objective function $\ell(\beta)$ is concave in λ, the computational cost to

evaluate the log empirical likelihood ratio is not expensive. Our next theorem demonstrates that the Wilks result holds for the above empirical likelihood ratio. This extends Owen's (1988, 1990) empirical likelihood ratio to our semiparametric predictive model.

Theorem 4.1. Suppose conditions in Theorem 3.1 hold and $E|v_t|^3 < \infty$. Then $\ell(\beta_0)$ converges in distribution to $\chi^2(k)$, a chi squared distribution with degrees of freedom k, as n goes to ∞.

Using Theorem 4.1, we construct a $100(1-\alpha)\%$ confidence region for β_0 as

$$I_\alpha = \{\beta : \ell(\beta) \leq \chi^2_{k,\alpha}\},$$

where $\chi^2_{k,\alpha}$ is the α-quantile of $\chi^2(k)$.

For linear predictive regression, a similar unifying interval estimate was previously proposed by Zhu, Cai and Peng (2014), based on the empirical likelihood. However, their method works only for simple predictive regression with a single predictor. Our procedure allows for multiple predictors and nonlinearity, which provides us a simultaneous inference tool for the coefficients of predictors. In addition, our empirical likelihood is built upon more efficient score equation estimation than theirs (see Section 2.2).

CHAPTER 5: SIMULATIONS

To investigate the finite sample performance of the proposed weighted estimation and empirical likelihood (WEEL) method and to compare it with other procedures while applicable, we run $1,000$ simulations for the model

$$y_t = x_{t-1}^T \beta + \theta(z_t) + v_t, \tag{5.1}$$

$$x_{t,i} = \rho_i \, x_{t-1,i} + u_{t,i}, \quad u_{t,i} \sim N(0,1),$$

through the following examples, where β is a k-dimensional vector. We set $\theta(z) = \sin(\pi z)$ and consider three distributions of v_t: $v_t^{(1)} = N(0,1)$, $v_t^{(2)} = t(3)$, and $v_t^{(3)} = 0.95N(0,1) + 0.05N(0,3^2)$. Our estimation involves the choice of bandwidth h and kernel function $K(\cdot)$. We use the Gaussian kernel and select the rule-of-thumb bandwidth $h = 1.06Sn^{-1/3}$ for estimation of β, where S is the sample standard deviation of $\{z_t\}$. This is not optimal, but it works since Theorem 3.1 holds under a large range of bandwidth. Note that such an h satisfies $nh^4 \to 0$ and $nh \to \infty$. With the optimal bandwidth the WEEL should perform better.

Example 5.1. Generate data from model (5.1) with $k = 1$, $\beta = 0.5$, $\rho = 1$, and z_t equal to $z_t^{(1)} = x_t - x_{t-1}$ or $z_t^{(2)} = 0.5z_{t-1}^{(2)} + \epsilon_t$, where $\epsilon_t \sim N(0,1)$. This is I(1) case.

Example 5.2. The setting is the same as in Example 5.1 but with $\rho = 0.99$, an NI(1) or I(0) case.

Example 5.3. Generate data from model (5.1) with $k = 3$, $\beta = (0.5, 1, -0.7)^T$, and z_t equal to $z_t^{(1)} = x_{t,1} - x_{t-1,1}$ or $z_t^{(2)} = 0.5z_{t-1}^{(2)} + \epsilon_t$, where $\epsilon_t \sim N(0, 1)$. Both $x_{t,1}$ and $x_{t,2}$ are AR(1) processes with $\rho_1 = 0.95$ and $\rho_2 = 1$, respectively, and $x_{t,3} = x_{t-1,3} + u_{t,3} + 0.5u_{t-1,3}$ is a nonstationary ARIMA(0,1,1) process. This is a mixing case of one I(0) variable and two I(1) variables.

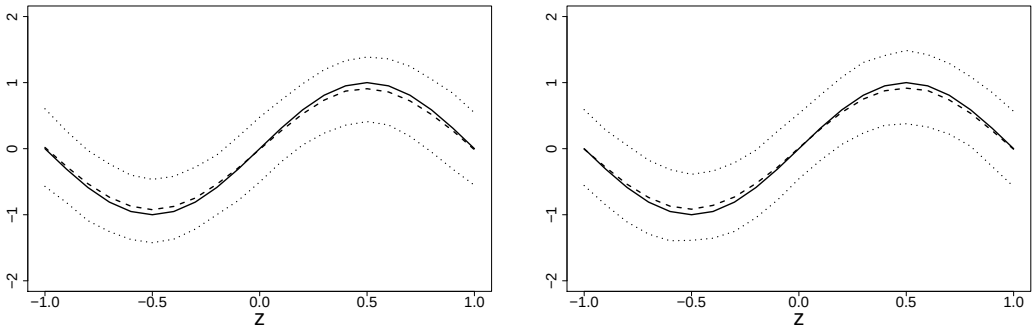
In each simulation we draw a sample of size $n = 200$ for Examples 5.1 and 5.2 and size $n = 400$ for Example 5.3. Tables 1-2 present summarized results from simulations for our method, including the average (ave) and standard deviation (std) of $\hat{\beta}_\omega$ and the coverage probability (cp) of the 95% confidence interval among $1,000$ simulations. It is evident that our WEEL procedure has very good performance in these examples in terms of bias, standard deviation and coverage probability.

For the case of I(1) in Example 5.1 and for the case of I(0) or NI(1) in Example 5.2, we also run simulations with unweighted profile least squares estimation for comparison, and the results were almost same and thus omitted for saving space. This exemplifies semiparametric efficiency of our estimation in the I(0) case, which is consistent with the theoretical result in Corollary 1. For the I(1) case, the results from the two estimation methods are also quite similar. This is also expected from the fact that c^* is chosen to maximize the efficiency of our estimator. Figure 1 displays $\hat{\theta}_\omega(\cdot)$ for Example 5.3 with $v_t \sim t(3)$. It is seen that the estimator captures the true curve very well. For other scenarios, the estimated curves are quite similar and not reported for saving space.

Example 5.4. To investigate if our WEEL method works for linear predictive models

Table 1: Simulation results for estimators of β

		Example 1		Example 2	
		$z_t = z_t^{(1)}$	$z_t = z_t^{(2)}$	$z_t = z_t^{(1)}$	$z_t = z_t^{(2)}$
$v_t^{(1)}$	ave	0.5000	.05001	0.4997	0.4993
	std	0.0173	0.0165	0.0199	0.0198
	cp	94.9%	95.5%	94.8%	95.5%
$v_t^{(2)}$	ave	0.5002	0.5003	0.4982	0.5002
	std	0.0288	0.0298	0.0356	0.0337
	cp	93.6%	94.0%	93.9%	94.5%
$v_t^{(3)}$	ave	0.5003	0.4998	0.4997	0.4996
	std	0.0196	0.0198	0.0226	0.0234
	cp	94.4%	94.7%	94.7%	94.2%



(a) (b)

Figure 1: Example 3 with $v_t \sim t(3)$. (a) - $z_t = z_t^{(1)}$, (b) - $z_t = z_t^{(2)}$. Solid - true, dashed - median curve, dotted - 2.5% and 97.5% pectentiles.

Table 2: Simulation results for estimators of β in Example 3

		$z_t = z_t^{(1)}$			$z_t = z_t^{(2)}$		
$v_t^{(1)}$	ave	(0.5007,	0.9997,	-0.7002)	(0.5009,	1.0000,	-0.7002)
	std	(0.0190,	0.0100,	0.0072)	(0.0189,	0.0099,	0.0067)
	cp		94.1%			95.1%	
$v_t^{(2)}$	ave	(0.5001,	0.9996,	-0.7004)	(0.5008,	1.0004,	-0.7004)
	std	(0.0321,	0.0174,	0.0113)	(0.0336,	0.0177,	0.0122)
	cp		93.5%			93.9%	
$v_t^{(3)}$	ave	(0.5002,	0.9999,	-0.7002)	(0.5006,	0.9997,	-0.6998)
	std	(0.0233,	0.0116,	0.0077)	(0.0230,	0.0120,	0.0082)
	cp		94.1%			94.0%	

and to compare with the two-stage regression (TSR) method in Cai and Wang (2014), we generate data from model (5.1) with $k = 1$, $\beta = 0.5$, $\rho = 1$, $z_t = x_t - x_{t-1}$, and $\theta(z) = z$, where $\epsilon_t \sim N(0, 1)$. This is the I(1) case in Example 5.1 but with $\theta(\cdot)$ being linear so that the TSR can be used. Summarized results are reported in Table 3. We also compare typical performance of the WEEL and the TSR, based on two typical samples. Typical sample I leads to the estimate of β equal to the median estimates in simulations by using the WEEL, and Typical sample II is the one with median performance by using the TSR. Table 4 reports the typical estimates and the 95% confidence intervals. Clearly both estimators have ignorable biases and close variances. This again shows that our WEEL does not lose much efficiency in the I(1) case, even though it is a linear predictive model that TSR works. However, the

Table 3: Simulation results for estimators of β in Example 4

	$v_t^{(1)}$			$v_t^{(2)}$			$v_t^{(3)}$		
	ave	std	cp	ave	std	cp	ave	std	cp
WEEL	0.5001	0.0165	95.2%	0.5002	0.0282	94.1%	0.5003	0.0194	94.5%
TSR	0.4999	0.0161	99.6%	0.5002	0.0280	99.5%	0.5004	0.0189	99.1%

Table 4: Estimates of β with two typical samples in Example 4

		WEEL		TSR	
	Typical samples	$\hat{\beta}_\omega$	95% CI	$\hat{\beta}_{cw}$	95% CI
$v_t^{(1)}$	I	0.5001	[0.4825, 0.5179]	0.4972	[0.4347, 0.5598]
	II	0.4991	[0.4816, 0.5172]	0.5001	[0.4433, 0.5568]
$v_t^{(2)}$	I	0.5007	[0.4332, 0.5633]	0.4752	[0.3547, 0.5957]
	II	0.4969	[0.4689, 0.5263]	0.5008	[0.3867, 0.6149]
$v_t^{(3)}$	I	0.4993	[0.4809, 0.5177]	0.4986	[0.4463, 0.5509]
	II	0.4986	[0.4702, 0.5232]	0.5001	[0.4245, 0.5757]

advantage of our interval estimate over the TSR is substantial in terms of the length and the coverage probability at 95% significance level. Since the standard error of the estimator in Cai and Wang (2014) was hard to estimate, they suggested to calculate the critical value for their confidence interval by simulation. This generated erratic estimates of the standard error in simulations and led to unacceptable coverage probability.

CHAPTER 6: REAL EXAMPLES

In these examples, we apply our methodology to examine the predictability of equity returns. We revisit the data in Campbell and Yogo (2006). We consider 4 log returns and 8 predictors. The 4 log returns are those for the annual S&P 500 index data (1880-2002) and the annual, quarterly and monthly NYSE/AMEX value-weighted index data (1926-2002) from the Center for Research in Security Prices (CRSP). The first two predictors are the S&P log D-P and log E-P ratios, and the remaining six predictors are annual, quarterly, monthly CRSP log D-P and log E-P ratios. For each series of the four log returns, there is a log D-P ratio series and a log E-P ratio series associated with it.We calculate the excess return for each log return according to Campbell and Yogo (2006). Through the following examples we analyze the data using our methodology and compare it with those in Campbell and Yogo (2006) and Cai and Wang (2014).

Example 6.1. Consider the predictability of the log E-P ratio for the corresponding excess return. Campbell and Yogo (2006, Table 4) found that the 8 predictors are highly persistent and some of them, such as the monthly log E-P ratio and the annual, quarterly and monthly log D-P ratios, are I(1) processes. Campbell and Yogo (2006) modeled the data using the predictive regression model

$$r_t = \alpha + \beta x_{t-1} + \epsilon_t, \tag{6.1}$$

where r_t is one of the 4 excess returns and x_{t-1} represents the associated log E-P ratio. Denoted by $\hat{u}_t$ the OLS residuals from the AR(1) model, $x_t = \gamma + \rho x_{t-1} + u_t$, and $\hat{\epsilon}_t$ the OLS residuals from model (6.1). According to Table 4 of Campbell and Yogo (2006), the sample correlation coefficients between the two innovations $\hat{\epsilon}_t$ and $\hat{u}_t$ are all in $(-0.987, -0.957)$ for the four predictive models. These non-zero correlations imply the existence of endogeneity, which may lead to biased estimates. In order to deal with the endogeneity, we consider our semiparametric predictive regression model

$$r_t = \beta x_{t-1} + \theta(\hat{u}_t) + v_t, \tag{6.2}$$

which reduces to Cai and Wang's (2014) model if $\theta(\cdot)$ is linear. Table 5 reports the point estimator of β, the 90% confidence intervals of β, and the corresponding standard error (se) of the model for each of the three estimation methods: our weighted estimator $\hat{\beta}_\omega$ for model (6.2), Campbell and Yogo's (2006) estimator $\hat{\beta}_{cy}$ for model (6.1), and Cai and Wang's (2014) estimator $\hat{\beta}_{cw}$ for model (6.2) with $\theta(\hat{u}_t) = \hat{u}_t$. Except for the S&P 500 excess return with the TSR method, all of the confidence intervals lie above zero, indicating that the log E-P ratio has predictability for its return. Our WEEL gives the shortest confidence intervals.

Example 6.2. In this example, we check the predictability of the log D-P ratio using models (6.1) and (6.2). The endogeneity of all 4 log D-P ratio series was confirmed by Table 4 in Campbell and Yogo (2006). It is expected that model (6.2) will provide a better fit than model (6.1). The estimated β coefficients, 90% confidence intervals, and the standard errors of the models are listed in Table 6. Unlike the other two procedures, the WEEL identifies predictability of all predictors. Again, the lengths

Table 5: Tests of Predictability with Log E-P Ratio

		SP500	Annual	Quarterly	Monthly
WEEL	$\hat{\beta}_\omega$	0.127	0.164	0.049	0.014
	90% CI	[0.112, 0.141]	[0.129, 0.195]	[0.042, 0.055]	[0.012, 0.015]
	se	0.048	0.053	0.017	0.009
Campbell	$\hat{\beta}$	0.131	0.169	0.049	0.014
and	90% CI	[0.042, 0.224]	[0.042, 0.277]	[0.010, 0.066]	[0.002, 0.018]
Yogo	se	0.176	0.187	0.105	0.054
TSR	$\hat{\beta}_{cw}$	0.127	0.162	0.047	0.013
	90% CI	[-0.002, 0.256]	[0.041, 0.284]	[0.025, 0.070]	[0.008, 0.018]
	se	0.048	0.054	0.018	0.009

of confidence intervals by WEEL are the smallest. To save space, we only report the estimates of the excess stock return for quarterly CRSP series in Figure 3. The estimated excess return $\hat{r}_t$ by WEEL for model (6.2) is essentially better than that by Campbell and Yogo (2006). This reflects existence of endogeneity and a poor fit of model (6.1). The WEEL and TSR methods have similar estimates of the β parameter and similar standard errors of the model, but the former is better than the latter because of shorter confidence intervals.

Example 6.3. To illustrate the use of the proposed semiparametric predictive regression, we inspect joint predictability of multiple predictors. The predictability of the long-short yield spread has been widely discussed in empirical studies. Campbell and Yogo (2006) showed that it has predictability to stock return in the sub time period, 1952-2002. Following Campbell and Yogo (2006), we use Moody's seasoned Aaa corporate bond yield as the long yield and the one-month T-bill rate as the short

Table 6: Tests of Predictability with Log D-P Ratio

		SP500	Annual	Quarterly	Monthly
WEEL	$\hat{\beta}_\omega$	0.086	0.171	0.039	0.008
	90% CI	[0.030, 0.128]	[0.125, 0.218]	[0.028, 0.048]	[0.006, 0.011]
	se	0.092	0.129	0.034	0.016
Campbell	$\hat{\beta}$	0.093	0.125	0.034	0.009
and	90% CI	[-0.033, 0.114]	[0.014, 0.188]	[-0.009, 0.044]	[-0.005, 0.010]
Yogo	se	0.179	0.189	0.106	0.054
TSR	$\hat{\beta}_{cw}$	0.083	0.162	0.034	0.008
	90% CI	[-0.134, 0.300]	[-0.472, 0.787]	[0.002, 0.066]	[0.001, 0.016]
	se	0.093	0.131	0.036	0.016

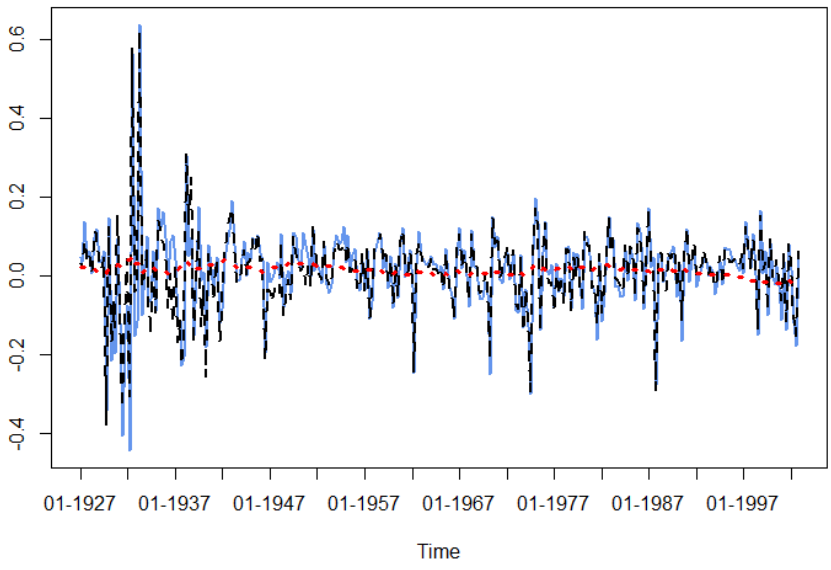


Figure 2: Estimated excess return for quarterly CRSP series, log D-P ratio

Table 7: Predictability of the log E-P ratio and yield spread

$\hat{\beta}_1$	$\hat{\beta}_2$	Ljung-Box Test	ADF Test
0.846	-3.537	0.181	<0.01

yield. Using the ADF test (p-value= 0.1149), we found that long-short yield spread s_t in the full-sample time period 1926-2002 is an I(1) process. Both CRSP log E-P ratio x_t and spread s_t are I(1) variables, but there is a cointegrating relationship between them, i.e. there are constants a and b such that $e_t = x_t - (a + bs_t)$ is stationary. Using the OLS, we obtain estimators $(\hat{a}, \hat{b})$ of (a, b) and the residuals $\hat{e}_t$. Motivated by the error-correction model in Engle and Granger (1987), we consider the predictive model

$$r_t = \beta_1 x_{t-1} + \beta_2 s_{t-1} + \theta(\hat{e}_t) + v_t, \tag{6.3}$$

where r_t is the monthly CRSP excess return and $\theta(\hat{e}_t)$ is a nonlinear error-correction term. The estimated coefficients and residual diagnostics are recorded in Table 7. From the Ljung-Box test and the ADF test, we see that the residuals from model (6.3) are stationary and there is no significant autocorrelation in the residuals. This suggests that model (6.3) is appropriate. The 95% confidence region of β in Figure 3 does not include the origin, so we conclude that the return is predictable jointly with the log E-P ratio and the yield spread.

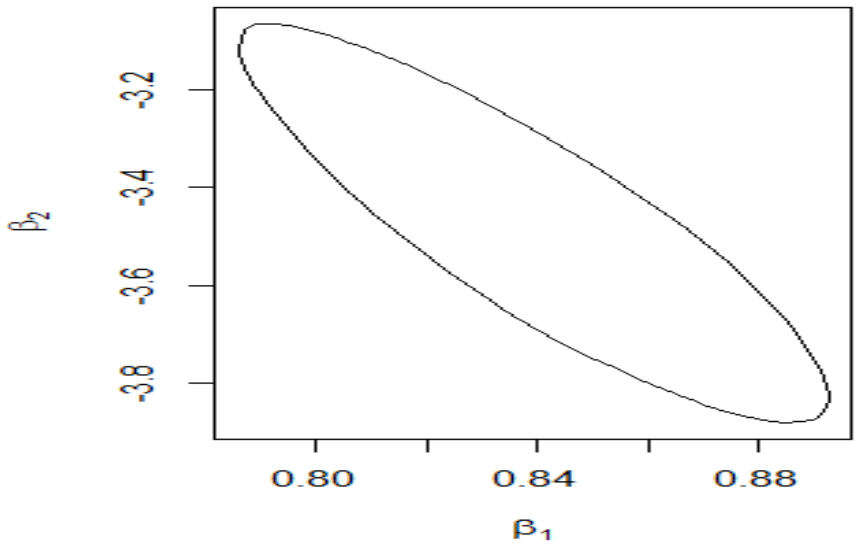


Figure 3: 95% confidence region of β in model (6.3)

REFERENCES

[1] Amilud, Y. and Hurvich, C. M. Predictive Regressions: A Reduced-Bias Estimation Method. *The Journal of Financial and Quantitative Analysis*, 39:813–841, 2004.

[2] Amilud, Y., Hurvich, C. M. and Wang, Y. Multiple-Predictor Regressions: Hypothesis Testing. *The Review of Financial Studies*, 22:413–434, 2009.

[3] Berkes, I. and Horváth, L. Convergence of integral functionals of stochastic processes. *Econometric Theory*, 22:304–322, 2006.

[4] Bickel, P. J. One-step Huber estimates in linear models. *Journal of the American Statistics Association*, 70:428–433, 1975.

[5] Campbell, J. Y. and Yogo, M. Efficient tests of stock return predictability. *Journal of Financial Economics*, 81:27–60, 2006.

[6] Cai, Z. and Li, Q. and Park, Joon Y. Functional-coefficient models for nonstationary time series data. *Journal of Econometrics*, 148:101–113, 2009.

[7] Cai, Z. and Wang, Y. Testing predictive regression models with nonstationary regressors. *Econometric Theory*, 178:4–14, 2014.

[8] Cai, Z. and Wang, Y. and Wang, Y. Testing instability in a predictive regression model with nonstationary regressions. *Econometric Theory*, 31:953-980, 2015.

[9] Carroll, R. J., Fan, J., Gijbels, I. and Wand, M. P. Generalized partially single-index models. *Journal of the American Statistical Association*, 92:477–489, 1997.

[10] Cavanagh, C. L. and Elliott, G. and Stock, J. H. Inference in Models with Nearly Integrated Regressors. *Econometric Theory*, 11:1131–1147, 1995.

[11] Chan, N. H. and Li, D. and Peng, L. Toward a unified interval estimation of autoregressions. *Econometric Theory*, 28:705–717, 2012.

[12] Chan, N. H. and Wei, C. Z. Asymptotic inference for nearly nonstationary AR(1) process. *The Annals of Statistics*, 15:1050–1063, 1987.

[13] Chan, N. H. and Zhang, R. M. Inference for nearly nonstationary processes under strong dependence and infinite variance. *Statistica Sinica*, 19:925–947, 2009.

[14] Datta, S. On asymptotic properties of bootstrap for AR(1) processes. *Journal of Statistical Planning and Inference*, 53:361–374, 1996.

[15] Elliott, G. and Stock, J. H. Inference in time series regression when the order of integration of a regressor is unknown. *Econometric theory*, 10:672–700, 1994.

[16] Engle, R. F. and Granger, C.W. J. Co-integration and error-correction: Representation, estimation and testing. *Econometrica*, 55:251–276, 1987.

[17] Fan, J. and Gijbels, I. Local Polynomial Modelling and Its Applications. *CRC Press*, 1996.

[18] Fan, J. and Jiang, J. Nonparametric inference for additive models. *Journal of the American Statistics Association*, 100:890–907, 2005.

[19] Fan, J. and Yao, Q. Nonlinear Time Series: Nonparametric and Parametric Methods. *Springer-Verlag*, 2003.

[20] Hall, P. The Bootstrap and Edgeworth Expansion. *New York: Springer*, 1992.

[21] Hall, P. and Heyde, C. C. Martingale Limit Theory and Its Application. *Academic Press*, 1980.

[22] Hall, P. and Jing, B. Y. Comparison of bootstrap and asymptotic approximations to the distribution of a heavy-tailed mean. *Statistica Sinica*, 8:887–906, 1998.

[23] Hall, P. and La Scala, B. Methodology and algorithms of empirical likelihood. *International Statistical Review*, 58:109–127, 1990.

[24] Jiang, J. and Mack, Y. P. Robust local polynomial regression for dependent data. *Statistica Sinica*, 11:705–722, 2001.

[25] Kothari, S. P. and Shanken, J. Book-to-market, dividend yield, and expected market returns: A time-series analysis. *Journal of Financial Economics*, 44:169–203, 1997.

[26] Ling, S. Self-weighted least absolute deviation estimation for infinite variance autoregressive models. *Journal of the Royal Statistical Society*, 67:381–393, 2005.

[27] Owen, A. B. Empirical Likelihood Ratio Confidence Intervals for a Single Functional. *Biometrika*, 75:237–249, 1988.

[28] Owen, A. B. Empirical Likelihood Ratio Confidence Regions. *The Annals of Statistics*, 18:90–120, 1990.

[29] Owen, A. B. Empirical Likelihood. *CRC Press*, 2001.

[30] Paye, B. S. and Timmermann, A. Instability of return prediction models. *Journal of Empirical Finance*, 13:274–315, 2006.

[31] Phillips, P. C. B. Regression Theory for Near-Integrated Time Series. *Econometrica*, 56:1021–1043, 1988.

[32] Phillips, P. C. B. and Han, C. Gaussian inference in AR(1) times series with or without a unit root. *Econometric Theory*, 63:1023–1078, 2008.

[33] Polk, C. and Thompson, S. and Vuolteenaho, T. Cross-sectional forecasts of the equity premium. *Journal of Financial Economics*, 81:101–141, 2006.

[34] Rossi, B. Expectations hypotheses tests at Long Horizons. *Econometrics Journal*, 10:1–26, 2007.

[35] So, B. S. and Shin, D. W. Cauchy estimators for autoregressive processes. *Econometric Theory*, 15:165–176, 1999.

[36] Stambaugh, R. F. Predictive regressions. *Journal of Financial Economics*, 54:375–421, 1999.

[37] Speckman, P. Kernel smoothing in partial linear models. *The Journal of the Royal Statistical Society*, 50:413–436, 1988.

[38] Torous, W. and Valkanov, R. and Yan, S. On Predicting Stock Returns with Nearly Integrated Explanatory Variables. *The Journal of Business*, 77:937–966, 2004.

[39] Viceira, L. M. Testing for structural change in the predictability of asset returns. *Working Paper, Harvard University*, 1997.

[40] Zhu, F. and Cai, Z. and Peng, L. Predictive regressions for macroeconomic data. *The Annals of Applied Statistics*, 8:577–594, 2014.

APPENDIX A: SKETCH OF PROOFS

Lemma

To facilitate our arguments for proofs, we first introduce three lemmas.

Lemma A.1. *(Lemma A.1 of Cai, Wang and Wang, 2015) Assume $u_{t,i}$ is a stationary α-mixing process with mixing coefficient $\alpha_i(n)$ satisfying that $E(|u_{t,i}|^r) < \infty$ and $\sum_{n=1}^{\infty} \alpha_i^s(n) < \infty$, where $s = 1/(2+\delta_*) - 1/r$, $r > 2+\delta_*$, and $0 < \delta_* \leq 2$. Let $\theta_* = 1/2 - 1/(2+\delta_*)$ and $\lambda_* > 0$ is a function of δ_*. Then the NI(1) or I(1) process $U_{n,i}(r) = n^{-1/2} x_{[nr],i}$ for $0 \leq r \leq 1$ admits the following strong approximation*

$$\sup_{0 \leq r \leq 1} |U_{n,i}(r) - U_{\gamma_i}(r)| = O[n^{-\theta_*} \{\log(n)\}^{\lambda_*}]$$

holds almost surely, where $U_{\gamma_i}(\cdot)$ is the diffusion process in (3.1).

Lemma A.2. *Let $\tilde{\theta}(z) = \theta(z) - \sum_{s=1}^{n} \xi_s(z)\theta(z_s)$. Suppose Conditions ($A_2$)-($A_5$) hold. Then $\sup_{z \in supp(f)} |\tilde{\theta}(z)| = O_p(h^2)$.*

Proof. By using the nonparametric regression technique for mixing processes (Proposition 6.2 in Fan and Yao, 2003; Lemma 7.2 in Jiang and Mack, 2001), it is easy to obtain that

$$S_{n,j}(z) = \mu_j f(z) + O_p(h) \tag{A.1}$$

and

$$n^{-1} \sum_{t=1}^{n} w_t(z) = S_{n,0}(z) S_{n,2}(z) - S_{n,1}^2(z) = (\mu_0 \mu_2 - \mu_1^2) f^2(z) + o_p(1), \tag{A.2}$$

uniformly for $z \in \text{supp}(f)$. Note that the weights w_t satisfy that

$$n^{-1}\sum_{t=1}^{n} w_t(z)(z_t - z) = S_{n,1}(z)S_{n,2}(z) - S_{n,2}(z)S_{n,1}(z) = 0.$$

It follows that

$$\begin{aligned}\sum_{s=1}^{n}\xi_s(z)\theta(z_s) - \theta(z) &= \sum_{s=1}^{n}\xi_s(z)\{\theta(z_s) - \theta(z)\} \\ &= \sum_{s=1}^{n}\xi_s(z)\{\theta(z_s) - \theta(z) - \theta'(z)(z_s - z)\}.\end{aligned}$$

By Taylor's expansion, we have

$$\sum_{s=1}^{n}\xi_s(z)\theta(z_s) - \theta(z) = 0.5h^2\sum_{s=1}^{n}\xi_s(z)\theta''(\eta_s)h^{-2}(z_s - z)^2, \tag{A.3}$$

where η_s is between z and z_s. Since K has bounded support $[0, 1]$, the weight $w_s(z)$ and $\xi_s(z)$ do not vanish only when $|z_s - z| \le h$. Hence, for all non-vanishing terms on the right hand side of (A.3), we have $\max_s |\eta_s - z| \le h \to 0$. Let

$$L_{n,j}(z) = n^{-1}\sum_{s=1}^{n} K_h^{(j)}(z_s - z)\theta''(\eta_s),$$

where $K_h^{(j)}(z_s - z) = h^{-j}(z_s - z)K_h(z_s - z)$. Then (A.3) becomes

$$\sum_{s=1}^{n}\xi_s(z)\theta(z_s) - \theta(z) = 0.5h^2\{n^{-1}\sum_{t=1}^{n} w_t(z)\}^{-1}\{S_{n,2}(z)L_{n,2}(z) - S_{n,1}(z)L_{n,3}(z)\}. \tag{A.4}$$

Similar to (A.1),

$$L_{n,j}(z) = f(z)\mu_j\theta''(z) + O(h), \tag{A.5}$$

uniformly for $z \in supp(f)$. Hence, by the definition of $\tilde{\theta}(z)$, (A.1-(A.2) and (A.4)-

(A.5),

$$\sup_{z \in supp(f)} |\tilde{\theta}(z)| = O_p(h^2). \tag{A.6}$$

□

Lemma A.3. *Without loss of generality, assume that* $x_t = (X_{t,1}^\top, X_{t,2}^\top)^\top$, *where* $X_{t,1}$ *is a* $d \times 1$ *vector of NI(1) or I(1) variables and* $X_{t,2}$ *is a* $(k-d) \times 1$ *vector of stationary or I(0) variables (*$0 \le d \le k$*). Let* $\alpha(z) = E(X_{t-1,2}|z_t = z)$, $U_\gamma(r) = \{U_{\gamma_1}(r), \ldots, U_{\gamma_d}(r)\}^\top$, $B(r) = \int_0^1 \{U_\gamma(r) - U_\gamma(s)\}\, ds$, *and* $B^*(r) \equiv B(r)/\|B(r)\|$. *Assume conditions* (A_0)*-*(A_7) *and* $E|v_t|^3 < \infty$. *Then, with probability going to one, we have*

(i) $V_n \equiv n^{-1}\sum_{t=1}^n Z_t^{\otimes 2}(\beta_0) \xrightarrow{p} V$, *where* $V = \mathrm{diag}(V_1, V_2)$, *where* $V_1 = \sigma_v^2 \int_0^1 B^*(r)^{\otimes 2}\, dr$ *and* $V_2 = \sigma_v^2 E\{X_{t-1,2} - \alpha(z_t)\}^{\otimes 2}$.

(ii) $Z_n^* = \max_{1 \le t \le n} \|Z_t(\beta_0)\| = o_p(n^{1/2})$.

(iii) If $nh^4 = O(1)$, *then* $\|\bar{Z}\| = O_p(n^{-\frac{1}{2}})$, *where* $\bar{Z} = n^{-1}\sum_{t=1}^n Z_t(\beta_0)$.

(iv) $n^{-1}\sum_{t=1}^n \|Z_t(\beta_0)\|^3 = o_p(n^{1/2})$.

Proof. **of Lemma A.3.** For $i = 1, \ldots, d$, $x_{t,i}$ is NI(1) or I(1). For $(t-1)/n \le r \le t/n$, define $U_{n,i}(r) = U_{ni,t} = n^{-1/2} x_{t-1,i}$. Then

$$\begin{aligned}
\sum_{s=1}^n \xi_s(z) x_{s-1,i} &= n^{1/2}\{\sum_{t=1}^n w_t(z)\}^{-1} \sum_{s=1}^n K_h(z_s - z)\{S_{n,2}(z) - h^{-1}(z_s - z)S_{n,1}(z)\}U_{ni,s} \\
&= n^{1/2}\{n^{-1}\sum_{t=1}^n w_t(z)\}^{-1}\{S_{n,2}(u)F_{n,0}(z) - S_{n,1}(z)F_{n,1}(z)\},
\end{aligned}$$

where for $j = 0, 1$, $F_{n,j}(z) = n^{-1}\sum_{s=1}^n K_h^{(j)}(z_s - z)U_{ni,s}$, with $K_h^{(j)}(z_s - z) = h^{-j}(z_s - z)K_h(z_s - z)$. Using the same argument as that for $F_{n,j,1}$ in equation (A.11) of Cai,

Li and Park (2009), we obtain that $F_{n,j}(z) = f(z)\mu_j W_u^{(i)} + o_p(1)$, where $W_u^{(i)} = \int_0^1 U_{\gamma_i}(r)\,dr$. By Bickel's (1975) chaining argument, this result holds uniformly for $z \in \text{supp}(f)$. Therefore,

$$\sum_{s=1}^{n} \xi_s(z) x_{s-1,i} = n^{1/2} W_u^{(i)} + o_p(n^{1/2}),$$

uniformly for $z \in \text{supp}(f)$. Then

$$\tilde{x}_{t-1,i} = n^{1/2}[U_{ni,t} - W_u^{(i)}] + o_p(n^{1/2}) \tag{A.7}$$

uniformly for $t = 1, \ldots, n$. By the strong approximation in Lemma A.1, we have $U_{n,i}(r) \xrightarrow{a.s} U_{\gamma_i}(r)$. Therefore,

$$n^{-1/2}\tilde{x}_{[nr],i} = U_{\gamma_i}(r) - W_u^{(i)} + o_p(1)$$

or equivalently

$$n^{-1/2}\tilde{X}_{[nr],1} = U_\gamma(r) - W_u + o_p(1) \equiv B(r) + o_p(1), \tag{A.8}$$

uniformly for $0 \le r \le 1$, where $W_u = \{W_u^{(1)}, \ldots, W_u^{(d)}\}^\top = \int_0^1 U_\gamma(r)\,dr$. Thus, $B(r) = \int_0^1 \{U_\gamma(r) - U_\gamma(s)\}\,ds$. Then, for $i = 1, \ldots, d$,

$$\begin{aligned} P\{\omega_{t,i} = (1 + \tilde{X}_{t-1,1})^{-1/2},\, \forall t = 1, \ldots, n\} &= P\{\max_{1\le t\le n} n^{-1/2}\log(n)\|\tilde{X}_{t-1,1}\| \ge c^*\} \\ &\to P\{\sup_{r\in[0,1]} \|B(r)\| \ge 0\} = 1. \end{aligned} \tag{A.9}$$

Let $\alpha(z) = E(X_{t-1,2}|z_t = z)$ and $\sigma^2(z) = \text{Var}(X_{t-1,2}|z_t = z)$. Then

$$X_{t-1,2} = \alpha(z_t) + \sigma(z_t)\epsilon_t,$$

where ϵ_t satisfies that $E(\epsilon_t|z_t) = 0$ and $\mathrm{Var}(\epsilon_t|z_t) = I_{k-d}$. The local linear estimator of $\alpha(z)$ is

$$\hat{\alpha}(z) = \sum_{s=1}^{n} \xi_s(z) X_{s-1,2}.$$

Since $\alpha(\cdot)$ has a continuous 2nd order derivative and $\sigma(z)$ is continuous, by Theorem 6.5 in Fan and Yao (2003), we have

$$\sup_{z \in \mathrm{supp}(f)} \|\hat{\alpha}(z) - \alpha(z)\| = O_p[h^2 + \{\log(1/h)/(nh)\}^{1/2}].$$

It follows that

$$\tilde{X}_{t-1,2} = X_{t-1,2} - \alpha(z_t) + O_p[h^2 + \{\log(1/h)/(nh)\}^{1/2}], \tag{A.10}$$

uniform for $t = 1, \ldots, n$. Furthermore, $n^{-1/2}\log(n)\max_{1\le t\le n}\|X_{t,2}\| = o_p(1)$, if $E\|X_{t,2}\|^{2+\delta} < \infty$. Hence, for $i = d+1, \ldots, k$,

$$\begin{aligned} P(\omega_{t,i} = 1,\, \forall t = 1, \ldots, n) &= 1 - P(\max_{1\le t\le n} n^{-1/2}\log(n)|\tilde{x}_{t-1,i}| \ge c^*) \\ &\to 1. \end{aligned} \tag{A.11}$$

Let $\Omega_{t,1} = (1 + \|\tilde{X}_{t-1,1}\|^2)^{-1/2} I_d$ and $\Omega_{t,2} = I_{k-d}$. It follows from (A.9) and (A.11) that with probability going to one

$$\Omega_t = \mathrm{diag}(\Omega_{t,1}, \Omega_{t,2}). \tag{A.12}$$

We adopt Phillips' (1988) notation and represent $\rho_i = \exp(\gamma_i/n)$ for $\gamma_i \in R$ and for $i = 1, \ldots, d$. Without loss of generality, it is assumed that $E(x_{0,i}) = 0$. By the

definition of $Z_t(\beta)$ and (2.8), we have

$$\begin{aligned} Z_t(\beta_0) &= \Omega_t \tilde{x}_{t-1}(\tilde{y}_t - \beta_0^\top \tilde{x}_{t-1}) \\ &= \Omega_t \tilde{x}_{t-1}\{y_t - x_{t-1}^\top \beta_0 - \hat{\theta}(z_t;\beta_0)\} \\ &= \Omega_t \tilde{x}_{t-1} v_t - \delta_t, \end{aligned} \tag{A.13}$$

where $\delta_t = \Omega_t \tilde{x}_{t-1}\{\hat{\theta}(z_t;\beta_0) - \theta(z_t;\beta_0)\}$. For the model $y_t - \beta_0^\top x_{t-1} = \theta(z_t) + v_t$, the local linear estimator $\hat{\theta}(z;\beta_0)$ satisfies that [see Theorem 6.5 in Fan and Yao (2003)]

$$\sup_{z \in supp(f)} |\hat{\theta}(z;\beta_0) - \theta(z;\beta_0)| = O_p[h^2 + \{nh/\log(1/h)\}^{-1/2}]. \tag{A.14}$$

By (A.13), we have

$$\{Z_t(\beta_0)\}^{\otimes 2} = (\Omega_t \tilde{x}_{t-1})^{\otimes 2} v_t^2 - (\Omega_t \tilde{x}_{t-1}\delta_t^\top + \delta^\top \Omega_t \tilde{x}_{t-1}) v_t + \delta_t^{\otimes 2}.$$

Then

$$\begin{aligned} V_n &= n^{-1}\sum_{t=1}^n (\Omega_t \tilde{x}_{t-1})^{\otimes 2} v_t^2 - n^{-1}\sum_{t=1}^n (\Omega_t \tilde{x}_{t-1}\delta_t^\top + \delta^\top \Omega_t \tilde{x}_{t-1}) v_t + n^{-1}\sum_{t=1}^n \delta_t^{\otimes 2} \\ &\equiv W_{n1} - W_{n2} + W_{n3}. \end{aligned}$$

It is easy to see from (A.14) that $W_{n3} = o_p(W_{n1})$. Then, by the Hölder inequality, W_{n2} is also dominated by W_{n1}. Hence, $V_n = W_{n1} + o_p(1)$. Let $W_{n1}^* = n^{-1}\sum_{t=1}^n (\Omega_t \tilde{x}_{t-1})^{\otimes 2}\sigma_v^2$. Then $E(W_{n1} - W_{n1}^*) = 0$. By (A.10) and Condition (A_7), we have $\mathrm{Var}(W_{n1} - W_{n1}^*) = o(1)$. Therefore, $W_{n1} = W_{n1}^* + o_p(1)$ and

$$V_n = \sigma_v^2 n^{-1}\sum_{t=1}^n (\Omega_t \tilde{x}_{t-1})^{\otimes 2} + o_p(1) \equiv \sigma_v^2 V_n^* + o_p(1). \tag{A.15}$$

(i) The result can be proven through the following steps:

Step 1. Let $\xi_{n1} = n^{-1}\sum_{t=1}^{n}(\Omega_{t,1}\tilde{X}_{t-1,1})^{\otimes 2}$ and $\xi_{n2} = n^{-1}\sum_{t=1}^{n}(\Omega_{t,2}\tilde{X}_{t-1,2})^{\otimes 2}$. Then

$$\begin{aligned}\xi_{n1} &= n^{-1}\sum_{t=1}^{n}(1/n + \|n^{-1/2}\tilde{X}_{t-1,1}\|^2)^{-1}(n^{-1/2}\tilde{X}_{t-1,1})^{\otimes 2} \\ &= n^{-1}\sum_{t=1}^{n}(\|n^{-1/2}\tilde{X}_{t-1,1}\|^2)^{-1}(n^{-1/2}\tilde{X}_{t-1,1})^{\otimes 2} + o_p(1).\end{aligned}$$

Let $B^*(r) = B(r)/\|B(r)\|$. Using (A.8) and Theorem 1.2 in Berkes and Horváth (2006), we establish that

$$\xi_{n1} = \int_0^1 \|B(r)\|^{-2}B(r)^{\otimes 2}\,dr + o_p(1) = \int_0^1 B^*(r)^{\otimes 2}\,dr + o_p(1).$$

By (A.10) and (A.12), with probability tending to one, we have

$$\begin{aligned}\xi_{n2} &= n^{-1}\sum_{t=1}^{n}\{X_{t-1,2} - \alpha(z_t)\}^{\otimes 2} + o_p(1) \\ &= E\{X_{t-1,2} - \alpha(z_t)\}^{\otimes 2} + o_p(1). \qquad \text{(A.16)}\end{aligned}$$

Step 2. Let $\xi_{n,12} = n^{-1}\sum_{t=1}^{n}\Omega_{t,1}\tilde{X}_{t-1,1}(\Omega_{t,2}\tilde{X}_{t-1,2})^{\top}$. Then

$$\begin{aligned}\xi_{n,12} &= n^{-1}\sum_{t=1}^{n}(1 + \|\tilde{X}_{t-1,1}\|^2)^{-1/2}\tilde{X}_{t-1,1}\{X_{t-1,2} - \alpha(z_t)\}^{\top} + o_p(1) \\ &\equiv n^{-1}\sum_{t=1}^{n}U_{nt}^{*}e_t + o_p(1),\end{aligned}$$

where $U_{nt}^{*} = (1+\|\tilde{X}_{t-1,1}\|^2)^{-1/2}\tilde{X}_{t-1,1}$ and $e_t = \{X_{t-1,2} - \alpha(z_t)\}^{\top}$. Using an argument similar to that in Cai, Li and Park (2009, page 108), we obtain that

$$n^{-1}\sum_{t=1}^{n}U_{nt}^{*}e_t = o_p(1). \qquad \text{(A.17)}$$

Therefore, $\xi_{n,12} = o_p(1)$. In fact, for any $\delta \in (0,1)$, let $N = [1/\delta]$, $t_k = 1 + [nk/N]$,

$t_k^* = t_{k+1} - 1$, and $t_k^{**} = \min(t_k^*, n)$. Then, it can be rewritten that

$$\|n^{-1}\sum_{t=1}^{n} U_{nt}^* e_t\| = \|n^{-1}\sum_{k=0}^{N-1}\sum_{t=t_k}^{t_k^{**}} U_{nt}^* e_t\|. \tag{A.18}$$

By the triangle inequality, we have

$$\begin{aligned}
\|n^{-1}\sum_{k=0}^{N-1}\sum_{t=t_k}^{t_k^{**}} U_{nt}^* e_t\| &\leq \|n^{-1}\sum_{k=0}^{N-1} U_{nt_k}^* \sum_{t=t_k}^{t_k^{**}} e_t\| + \|n^{-1}\sum_{k=0}^{N-1}\sum_{t=t_k}^{t_k^{**}}(U_{nt}^* - U_{nt_k}^*)e_t\| \\
&\leq n^{-1}\sum_{k=0}^{N-1}\|U_{nt_k}^*\|\|\sum_{t=t_k}^{t_k^{**}} e_t\| + n^{-1}\sum_{k=0}^{N-1}\sum_{t=t_k}^{t_k^{**}}\|U_{nt}^* - U_{nt_k}^*\||e_t| \\
&\leq \sup_{0\leq t\leq 1}\|U_n^*(t)\| n^{-1}\sum_{k=0}^{N-1}|\sum_{t=t_k}^{t_k^*} e_t| + \sup_{|r-s|\leq\delta}\|U_n^*(r) - U_n^*(s)\| n^{-1}\sum_{t=1}^{n}|e_t|.
\end{aligned}$$

Using (A.8) and the continuous mapping theorem, we obtain that $U_n^*(r) \Rightarrow B^*(r) \equiv B(r)/\|B(r)\|$ on $[0,1]$. Hence, $\sup_{0\leq t\leq 1}\|U_n^*(t)\| = O_p(1)$. It is trivial to verify that $n^{-1}\sum_{t=1}^{n}|e_t| = O_p(1)$, since e_t is a α-mixing stationary sequence. Therefore,

$$\|n^{-1}\sum_{k=0}^{N-1}\sum_{t=t_k}^{t_k^{**}} U_{nt}^* e_t\| = O_p(1) n^{-1}\sum_{k=0}^{N-1}|\sum_{t=t_k}^{t_k^*} e_t| + O_p(1)\sup_{|r-s|\leq\delta}\|U_n^*(r) - U_n^*(s)\|. \tag{A.19}$$

Note that $\sup_{|r-s|\leq\delta}\|U_n^*(r) - U_n^*(s)\| \xrightarrow{D} \sup_{|r-s|\leq\delta}\|B^*(r) - B^*(s)\| \xrightarrow{p} 0$ as $\delta \to 0$ and

$$\begin{aligned}
E\{n^{-1}\sum_{k=0}^{N-1}|\sum_{t=t_k}^{t_k^*} e_t|\} &\leq N/n \sup_{0\leq k\leq N-1} E|\sum_{t=t_k}^{t_k^*} e_t| \\
&\leq \sup_{t\leq n} E|(\delta n)^{-1}\sum_{i=t}^{t+\delta n} e_i| \leq \sup_{t\leq n}\big[\mathrm{Var}\big\{(\delta n)^{-1}\sum_{i=t}^{t+\delta n} e_i\big\}\big]^{1/2} \\
&= O\{(\delta n)^{-1}\} \to 0.
\end{aligned}$$

as $n \to \infty$. It follows from (A.18)-(A.19) that (A.17) holds.

Combining Steps 1 and 2 leads to

$$\sigma_v^2 V_n^* = \sigma_v^2 \begin{pmatrix} \xi_{n1} & \xi_{n12} \\ \xi_{n12}^\top & \xi_{n2} \end{pmatrix} \xrightarrow{p} V = \mathrm{diag}(V_1, V_2). \tag{A.20}$$

This, combined with (A.15), completes the proof.

(ii) The result requires $2+\delta$ moment for $X_{t,2}$. Note that $Z_t(\beta_0) = \{Z_{t,1}^\top(\beta_0), Z_{t,2}^\top(\beta_0)\}^\top$, where $Z_{t,j}(\beta_0) = \Omega_{t,j}\tilde{X}_{t-1,j}\{v_t - [\hat{\theta}(z;\beta_0) - \theta(z;\beta_0)]\}$ for $j = 1, 2$. It suffices to show that

$$\eta_{nj} \equiv n^{-1/2} \max_{1\le t\le n} \|\Omega_{t,j}\tilde{X}_{t-1,j}\| = o_p(1)$$

and

$$\eta_{nj}^* \equiv n^{-1/2} \max_{1\le t\le n} \|\Omega_{t,j}\tilde{X}_{t-1,j}v_t\| = o_p(1).$$

For $j = 1$, with probability tending to one, it is obvious that $\eta_{n1} \le n^{-1/2} = o(1)$ and $\eta_{n1}^* \le n^{-1/2}\max_{1\le t\le n}|v_t| = o_p(1)$. For $j = 2$, since $\lim_{n\to\infty} P(\Omega_{t,2} = I_{k-d}) \to 1$, using (A.10) we obtain that

$$\eta_{n2} = n^{-1/2} \max_{1\le t\le n} \|\tilde{X}_{t-1,2}\| = n^{-1/2} \max_{1\le t\le n} \|X_{t-1,2} - \alpha(z_t)\| + o_p(1) = o_p(1),$$

and

$$\eta_{n2}^* = n^{-1/2} \max_{1\le t\le n} \|\tilde{X}_{t-1,2}v_t\| = n^{-1/2} \max_{1\le t\le n} \|\{X_{t-1,2} - \alpha(z_t)\}v_t\| + o_p(1) = o_p(1).$$

(iii) By (A.13), we have

$$\begin{aligned} n^{1/2}\bar{Z} &= n^{-1/2}\sum_{t=1}^n Z_t(\beta_0) \\ &= n^{-1/2}\sum_{t=1}^n \Omega_t\tilde{x}_{t-1}v_t - n^{-1/2}\sum_{t=1}^n \Omega_t\tilde{x}_{t-1}[\hat{\theta}(z_t;\beta_0) - \theta(z_t;\beta_0)]. \end{aligned} \tag{A.21}$$

By the definition of $\hat{\theta}(\cdot;\beta_0)$, we have

$$\hat{\theta}(z;\beta_0) - \theta(z) = \sum_{t=1}^{n} \xi_t(z) v_t + \sum_{t=1}^{n} \xi_t(z)\{\theta(z_t) - \theta(z)\},$$

where, by (A.6), the second term on the right hand is equal to $-\tilde{\theta}(z) = O_p(h^2)$ uniformly for $z \in supp(f)$. Therefore,

$$\hat{\theta}(z;\beta) - \theta(z) = \sum_{s=1}^{n} \xi_s(z) v_s + O_p(h^2),$$

which combined with (A.21) yields that

$$\begin{aligned} n^{1/2}\bar{Z} &= n^{-1/2}\sum_{t=1}^{n} \Omega_t \tilde{x}_{t-1} v_t - n^{-1/2}\sum_{s=1}^{n}\{\sum_{t=1}^{n} \Omega_t \tilde{x}_{t-1}\xi_s(z_t)\} v_s + O_p(\sqrt{n}h^2) \\ &\equiv M_{n1} - M_{n2} + O_p(\sqrt{nh^4}). \end{aligned} \tag{A.22}$$

Note that $E(M_{n1}) = 0$ and $\mathrm{Var}(M_{n1}) = \sigma_v^2 E\{(\Omega_t \tilde{x}_{t-1})^{\otimes 2}\} = O(1)$. Then $M_{n1} = O_p(1)$. Let $r_s = \sum_{t=1}^{n} \Omega_t \tilde{x}_{t-1}\xi_s(z_t)$. Then it can be rewritten that

$$r_s = \sum_{t=1}^{n} \Omega_t \tilde{x}_{t-1} w_s(z_t) / \sum_{i=1}^{n} w_i(z_t) = (r_{s1}^{\top}, r_{s2}^{\top})^{\top},$$

where

$$r_{s1} = \sum_{t=1}^{n} \Omega_{t,1} \tilde{X}_{t-1,1} w_s(z_t) / \sum_{i=1}^{n} w_i(z_t)$$

and

$$r_{s2} = \sum_{t=1}^{n} \Omega_{t,2} \tilde{X}_{t-1,2} w_s(z_t) / \sum_{i=1}^{n} w_i(z_t).$$

By (A.1)-(A.2) and (A.10), with probability tending to one, we have

$$\begin{aligned} r_{s1} &= \{n^{-1}\sum_{i=1}^{n} w_i(z_t)\}^{-1} n^{-1} \sum_{t=1}^{n} \tilde{X}_{t-1,1}(1+\|\tilde{X}_{t-1,1}\|^2)^{-1/2} K_h(z_s - z_t) \\ &\quad \times \{S_{n,2}(z_t) - h^{-1}(z_s - z_t) S_{n,1}(z_t)\} \\ &= O_p(1). \end{aligned}$$

and

$$\begin{aligned} r_{s2} &= \sum_{t=1}^{n} \tilde{X}_{t-1,2} w_s(z_t) / \sum_{i=1}^{n} w_i(z_t) \\ &= \{n^{-1}\sum_{i=1}^{n} w_i(z_t)\}^{-1} n^{-1} \sum_{t=1}^{n} \{X_{t-1,2} - \alpha(z_t)\} K_h(z_s - z_t) \\ &\quad \times \{S_{n,2}(z_t) - h^{-1}(z_s - z_t) S_{n,1}(z_t)\} + o_p(1) \\ &= O_p(1), \end{aligned}$$

uniformly for $s = 1, \ldots, n$, which is a random variable independent of v_s. Since $M_{n2} = n^{-1/2}\sum_{s=1}^{n} r_s v_s$ has conditional mean zero and conditional variance $r_s^2 = O_p(1)$, $M_{n2} = O_p(1)$. This, combined with (A.22), leads to $n^{1/2}\bar{Z} = O_p(1)$.

(iv) Since $\|\Omega_{t,1}\tilde{X}_{t-1,1}\| \leq \sqrt{d}$ and $\Omega_{t,2}\tilde{X}_{t-1,2} = \tilde{X}_{t-1,2}$ with probability tending to one and $E|v_t|^3 < \infty$, the result holds from (A.13).

□

APPENDIX B: SKETCH OF PROOFS

Theorem 3.1

Proof. **of Theorem 3.1.** Let $a_n = \sum_{t=1}^n (\Omega_t \tilde{x}_{t-1})^{\otimes 2}$, $b_n = \sum_{t=1}^n \Omega_t \tilde{x}_{t-1}^{\otimes 2}$, $\tilde{v}_t = v_t - \sum_{s=1}^n \xi_s(z_t) v_s$, and $\tilde{\theta}(z) = \theta(z) - \sum_{s=1}^n \xi_s(z)\theta(z_s)$. Then $\tilde{y}_t = \beta^\top \tilde{x}_{t-1} + \tilde{\theta}(z_t) + \tilde{v}_t$. By the definition of $\hat{\beta}_\omega$, we have

$$\hat{\beta}_\omega - \beta = (\sum_{t=1}^n \Omega_t \tilde{x}_{t-1}^{\otimes 2})^{-1} \sum_{t=1}^n \Omega_t \tilde{x}_{t-1} \tilde{\theta}(z_t) + (\sum_{t=1}^n \Omega_t \tilde{x}_{t-1}^{\otimes 2})^{-1} \sum_{t=1}^n \Omega_t \tilde{x}_{t-1} \tilde{v}_t.$$

Therefore,

$$a_n^{-1/2} b_n (\hat{\beta}_\omega - \beta) = a_n^{-1/2} \sum_{t=1}^n \Omega_t \tilde{x}_{t-1} \tilde{\theta}(z_t) + a_n^{-1/2} \sum_{t=1}^n \Omega_t \tilde{x}_{t-1} \tilde{v}_t = B_n + V_n \qquad \text{(A.23)}$$

Using the Hölder inequality, we obtain that

$$\| \sum_{t=1}^n \Omega_t \tilde{x}_{t-1} \tilde{\theta}(z_t) \| \le (\sum_{t=1}^n \|\Omega_t \tilde{x}_{t-1}\|^2)^{1/2} \{ \sum_{t=1}^n \tilde{\theta}(z_t)^2 \}^{1/2}, \qquad \text{(A.24)}$$

where $\|\cdot\|$ denotes the Euclidean norm. It follows from (A.24) and Lemma A.2 that

$$\| \sum_{t=1}^n \Omega_t \tilde{x}_{t-1} \tilde{\theta}(z_t) \| \le (\sum_{t=1}^n \|\Omega_t \tilde{x}_{t-1}\|^2)^{1/2} O_p(\sqrt{nh^4}) = o_p\{ (\sum_{t=1}^n \|\Omega_t \tilde{x}_{t-1}\|^2)^{1/2} \}, \quad \text{(A.25)}$$

if $nh^4 \to 0$. Further, by (A.15) and (A.20),

$$\sum_{t=1}^n \|\Omega_t \tilde{x}_{t-1}\|^2 = n \mathrm{tr}(V_n^*) = n\{\mathrm{tr}(V) + o_p(1)\} = O_p(n), \qquad \text{(A.26)}$$

Combining (A.25) and (A.26) leads to

$$\| \sum_{t=1}^n \Omega_t \tilde{x}_{t-1} \tilde{\theta}(z_t) \| = o_p(\sqrt{n}).$$

Then

$$\|B_n\| \leq \|\sum_{t=1}^{n} \Omega_t \tilde{x}_{t-1} \tilde{\theta}(z_t)\| \max eig(a_n^{-1/2}) \leq o_p[\{\min eig(n^{-1}a_n)\}^{-1/2}] = o_p[\{\min eig(V_n^*)\}^{-1/2}].$$

Applying (A.20), we obtain that

$$B_n = o_p[\{\min eig(V)\}^{-1/2}] = o_p(1). \tag{A.27}$$

Note that

$$V_n = a_n^{-1/2}(\Omega_1 \tilde{x}_0, \ldots, \Omega_n \tilde{x}_{n-1})(\tilde{v}_1, \ldots, \tilde{v}_n)^\top \equiv a_n^{-1/2} c_n \tilde{v}$$

and

$$\tilde{v} = (I_n - S)v,$$

where $v = (v_1, \ldots, v_n)^\top$, I_n is an $n \times n$ identity matrix, and S is the smoothing matrix whose components are $\mathcal{F} \equiv \sigma(u_t, t \leq n)$ measurable and independent of v. Therefore, $E\{\tilde{v}|\mathcal{F}\} = 0$ and $E(\tilde{v}\tilde{v}^\top|\mathcal{F}) = (I_n - S)(I_n - S)^\top \sigma_v^2$. Following Fan and Jiang (2005, (B.11) and (B.21)), we establish that

$$E(\tilde{v}\tilde{v}^\top|\mathcal{F}) = \sigma_v^2 I_n\{1 + o(1)\}.$$

Hence, $E(V_n|\mathcal{F}) = 0$ and

$$E(V_n^{\otimes 2}|\mathcal{F}) = \sigma_v^2 a_n^{-1/2} \sum_{t=1}^{n} (\Omega_t \tilde{x}_{t-1})^{\otimes 2} a_n^{-1/2}\{1 + o(1)\} = \sigma_v^2 I_k + o(1).$$

Since $V_n = a_n^{-1/2} c_n (I_n - S)v \equiv d_n^\top v \equiv \sum_{t=1}^{n} d_{n_i} v_i$, where $d_n = (d_{n1}, \ldots, d_{nn})^\top$. Using the martingale limit theorem (Hall and Heyde, 1980), we can show that, for any $k \times 1$ vector a,

$$a^\top V_n \xrightarrow{\mathcal{D}} N(0, \sigma_v^2 a^\top a).$$

Then, by the Wald device, we have $V_n \xrightarrow{\mathcal{D}} N(0, \sigma_v^2 I_k)$. Applying the Slutsky theorem, (A.23) and (A.27), we conclude the result of the theorem. □

APPENDIX C: SKETCH OF PROOFS

Theorem 4.1

Proof. **of Theorem 4.1.** Note that $p_t = \frac{1}{n}\frac{1}{1+\lambda^\top Z_t(\beta_0)}$, where λ satisfies that

$$g(\lambda) \equiv n^{-1}\sum_{t=1}^{n}\frac{Z_t(\beta_0)}{1+\lambda^\top Z_t(\beta_0)} = 0.$$

It follows that

$$\begin{aligned} g(\lambda) &= n^{-1}\sum_{t=1}^{n} Z_t(\beta_0)[1-\frac{\lambda^\top Z_t(\beta_0)}{1+\lambda^\top Z_t(\beta_0)}] \\ &= n^{-1}\sum_{t=1}^{n} Z_t(\beta_0) - n^{-1}\sum_{t=1}^{n}\frac{Z_t^{\otimes 2}(\beta_0)}{1+\lambda^\top Z_t(\beta_0)}\lambda \\ &\equiv \bar{Z} - \tilde{V}_n\lambda = 0. \end{aligned}$$

That is,

$$\tilde{V}_n\lambda = \bar{Z}. \tag{A.28}$$

Let $V_n = n^{-1}\sum_{t=1}^{n} Z_t^{\otimes 2}(\beta_0)$, which is non-negative definitive. Since every $p_t > 0$, we have $1+\lambda^\top Z_t(\beta_0) > 0$, and hence,

$$\begin{aligned} \lambda^\top V_n\lambda &= \lambda^\top n^{-1}\sum_{t=1}^{n}\frac{Z_t^{\otimes 2}(\beta_0)}{1+\lambda^\top Z_t(\beta_0)}\left\{1+\lambda^\top Z_t(\beta_0)\right\}\lambda \\ &\le \lambda^\top n^{-1}\sum_{t=1}^{n}\frac{Z_t^{\otimes 2}(\beta_0)}{1+\lambda^\top Z_t(\beta_0)}\left\{1+\|\lambda\|\max_{1\le t\le n}\|Z_t(\beta_0)\|\right\}\lambda \\ &= \lambda^\top\tilde{V}_n\lambda(1+\|\lambda\|Z_n^*), \end{aligned} \tag{A.29}$$

where $Z_n^* = \max_{1\le t\le n}\|Z_t(\beta_0)\|$. Let $\lambda = \rho\theta$, where $\rho \ge 0$ and $\theta \in \mathcal{R}^K$ such that $\|\theta\| = 1$. Then $\|\lambda\| = \rho$. Combining (A.28) and (A.29), we establish that

$$0 \le \rho\theta^\top V_n\theta \le \rho\theta^\top\tilde{V}_n\theta(1+\rho Z_n^*) = \theta^\top\bar{Z}(1+\rho Z_n^*).$$

Hence, by Lemma A.3(i),

$$\theta^\top \bar{Z} \geq \frac{\rho}{1+\rho Z_n^*}\theta^\top V_n \theta = \frac{\rho}{1+\rho Z_n^*}\{\theta^\top V\theta + o_p(1)\}.$$

Then

$$\rho[\min eig(V) + o_p(1) - \theta'\bar{Z}Z_n^*] \leq \theta'\bar{Z}. \tag{A.30}$$

By Lemma A.3, we have $Z_n^* = o_p(n^{1/2})$ and $\|\bar{Z}\| = O_p(n^{-\frac{1}{2}})$, Then, with probability tending to one,

$$\|\lambda\| = \rho = O_p(n^{-\frac{1}{2}}) \tag{A.31}$$

and

$$\max_{1\leq t\leq n} |\lambda^\top Z_t(\beta_0)| \leq \|\lambda\| Z_n^* = O_p(n^{-\frac{1}{2}}) \cdot o_p(n^{-\frac{1}{2}}) = o_p(1).$$

Rewrite

$$\begin{aligned} 0 = g(\lambda) &= n^{-1}\sum_{t=1}^n Z_t(\beta_0)\left\{1 - \lambda^\top Z_t(\beta_0) + \frac{\lambda^\top Z_t^{\otimes 2}(\beta_0)\lambda}{1+\lambda^\top Z_t(\beta_0)}\right\} \\ &= \bar{Z} - V_n\lambda + n^{-1}\sum_{t=1}^n \frac{\lambda^\top Z_t^{\otimes 2}(\beta_0)\lambda Z_t(\beta_0)}{1+\lambda^\top Z_t(\beta_0)}. \end{aligned} \tag{A.32}$$

By Lemma A.3 and (A.31), the last term in (A.32) is bounded by

$$\begin{aligned} \|\lambda\|^2 n^{-1}\sum_{t=1}^n \frac{\|Z_t(\beta_0)\|^3}{1-\|\lambda\|Z_n^*} &= O_p(n^{-1})\{1+o_p(1)\} n^{-1}\sum_{t=1}^n \|Z_t(\beta_0)\|^3 \\ &= O_p(n^{-1})\{1+o_p(1)\} o_p(n^{\frac{1}{2}}) = o_p(n^{-\frac{1}{2}}). \end{aligned}$$

Hence, by (A.32),

$$\lambda = V_n^{-1}\bar{Z} + o_p(n^{-\frac{1}{2}}). \tag{A.33}$$

Note that $\max_{1\le t\le n}|\lambda^\top Z_t(\beta_0)| = o_p(1)$. By Taylor's expansion, we have

$$\log\left\{1+\lambda^\top Z_t(\beta_0)\right\} = \lambda^\top Z_t(\beta_0) - \frac{1}{2}\left\{\lambda^\top Z_t(\beta_0)\right\}^2 + \eta_t, \tag{A.34}$$

where for some finite $B > 0$,

$$P\left\{|\eta_t| \le B|\lambda^\top Z_t(\beta_0)|^3,\ 1\le t\le n\right\} \to 1,\ as\ n\to\infty. \tag{A.35}$$

It follows from (A.34) that

$$\begin{aligned} l(\beta_0) &= 2\sum_{t=1}^n \log\left\{1+\lambda^\top Z_t(\beta_0)\right\} \\ &= 2\sum_{t=1}^n \lambda^\top Z_t(\beta_0) - \sum_{t=1}^n \lambda^\top Z_t^{\otimes 2}(\beta_0)\lambda + 2\sum_{t=1}^n \eta_t. \end{aligned}$$

By Lemma 1, (A.31) and (A.35), with probability trending to one, we have

$$\begin{aligned} |\sum_{t=1}^n \eta_t| &\le B\|\lambda\|^3 \sum_{t=1}^n \|Z_t(\beta_0)\|^3 \\ &= O_p(n^{-\frac{3}{2}})\cdot o_p(n^{\frac{3}{2}}) = o_p(1). \end{aligned}$$

Therefore,

$$l(\beta_0) = 2\sum_{t=1}^n \lambda^\top Z_t(\beta_0) - \sum_{t=1}^n \lambda^\top Z_t^{\otimes 2}(\beta_0)\lambda + o_p(1).$$

Then, by (A.31) and (A.33),

$$l(\beta_0) = n\bar{Z}^\top V_n^{-1}\bar{Z} + o_p(1).$$

Applying the central limit theorem, we obtain that

$$\sqrt{n}V_n^{-\frac{1}{2}}\bar{Z} = V_n^{-\frac{1}{2}} n^{-\frac{1}{2}} \sum_{t=1}^n Z_t(\beta_0) \to N(0, I_k),$$

where I_k is the $K\times K$ identity matrix. Therefore, $l(\beta_0) \xrightarrow{p} \chi^2(K)$. □