cemmap
centre for microdata methods and practice

Multi-step non- and semi-parametric predictive regressions for short and long horizon stock return prediction

Tingting Cheng
Jiti Gao
Oliver Linton

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP03/18

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

Multi-Step Non- and Semi-Parametric Predictive Regressions for Short and Long Horizon Stock Return Prediction[1]

Tingting Cheng[‡], Jiti Gao[*] and Oliver Linton[†]

Nankai University[‡], Monash University[*] and University of Cambridge[†]

Abstract

In this paper, we propose three new predictive models: the multi-step nonparametric predictive regression model and the multi-step additive predictive regression model, in which the predictive variables are locally stationary time series; and the multi-step time-varying coefficient predictive regression model, in which the predictive variables are stochastically nonstationary. We also establish the estimation theory and asymptotic properties for these models in the short horizon and long horizon case. To evaluate the effectiveness of these models, we investigate their capability of stock return prediction. The empirical results show that all of these models can substantially outperform the traditional linear predictive regression model in terms of both in-sample and out-of-sample performance. In addition, we find that these models can always beat the historical mean model in terms of in-sample fitting, and also for some cases in terms of the out-of-sample forecasting.

Keywords: Kernel estimator, locally stationary process, series estimator, stock return prediction

JEL Classification: C14, C22, G17

1 Introduction

A fundamental issue in finance is whether future stock returns are predictable using publicly available information. The seminal studies of Keim and Stambaugh (1986), Fama and French (1988) and Campbell and Shiller (1988) empirically demonstrated that variables such as dividend yield, book-to-market ratio, or interest rate spreads have significant predictive ability over future stock returns. Fama (1991) interpreted these findings as evidence of time-varying risk premia rather than as evidence against market efficiency. Although financial economists have identified variables that predict stock returns through time, the "correct" predictive regression specification has remained an open issue. Several researchers have focused on using linear models to predict stock returns (see for example, Lewellen, 2004; Campbell and Shiller, 1988). A systematic discussion on the performance of mostly linear predictive models is given by Welch and Goyal (2008). However, as pointed out by Phillips (2015), there exists a potential misbalancing problem

[1]The second author would like to thank the Australian Research Council Discovery Grants Program for its support under Grant numbers: DP150101012 and DP170104421.

in the linear predictive regression model if some of the predictors have long memory and response variable has short memory.

On the other hand, some other researchers considered nonlinear models to predict stock returns. For example, Lettau and Van Nieuwerburgh (2008) suggested that after controlling the structural shift in the mean of dividend yield, the evidence of stock return predictability is much stronger. Chen and Hong (2010) developed a nonparametric predictability test to examine whether there exists a kind of predictability for equity returns for both short and long horizons and show that the nonparametric model can outperform the linear model. Scholz, Nielsen and Sperlich (2015) used nonparametric and semiparametric techniques to investigate the prediction of stock return over one–year horizon based on yearly data. Despite the significant volume of subsequent research, the predictability debate, and many econometric issues, in terms of the size and power of the existing tests, still remain unsolved (see for example, Stambaugh, 1999; Campbell and Yogo, 2006).

In this paper, we consider nonparametric approaches that allow for both linear and nonlinear predictability. A major issue in using nonparametric methods is the curse of dimensionality (Stone, 1980), which limits the number of covariates that can be allowed for flexibly. A further issue that affects the use of nonparametric methods is nonstationarity of predictor variables. To mitigate the curse of dimensionality we propose three new predictive models: the multi-step additive predictive regression model (APR), the multi-step time-varying coefficient predictive regression model (TVCPR), and the multi-step nonparametric predictive regression model (NPR). We use rescaled time as one of our covariates, which allows for variation over time in the predictive relationships, a point emphasized by for example Pesaran and Timmermann (1995). A closely related study is done by Kasparis, Andreou and Phillips (2015), which considered nonparametric predictive regressions with the regressor being a highly persistent process. In our work, we assume that the predictive variables are locally stationary time series in the NPR and APR models and nonstationary in the TVCPR model. Note that locally stationary processes have received a lot of attention. For example, Vogt (2012) studied nonparametric models allowing for locally stationary regressors and a regression function that changes smoothly over time. Dong and Linton (2016) studied nonparametric additive models that have deterministic time trend and both stationary (or locally stationary) and integrated variables as components. Meanwhile, varying coefficient time series models have been widely applied because of its flexibility, and different theoretical results have been investigated (see for example, Cai et al., 2009; Cai, 2007; Li et al., 2002; Phillips et al., 2017). We present the theoretical properties of our estimators of the regression functions in the short horizon and long horizon case, where by long horizon we mean that the horizon increases to infinity with the size of the sample. Many empirical studies

consider the long horizon case and our results support the use of nonparametric methods in this setting. To evaluate the effectiveness of these predictive models, we investigate their capability of stock–return prediction. The empirical results show that all of these models can substantially outperform the traditional linear predictive regression model in terms of both in-sample and out-of-sample performance. In addition, we find that these models can always beat the historical mean model in terms of in-sample fitting, and also for some cases in terms of the out-of-sample forecasting. The outlook for nonparametric methods looks somewhat more promising than was presented in Diebold and Nason (1990), although we acknowledge that the magnitude of the gain provided by these methods is modest.

The rest of this paper is organized as follows. In Section 2, we describe our models (i.e. NPR, APR, and TVCPR) in detail and establish asymptotic properties for the nonparametric estimators of the predictive functions. In Section 3, we present implementation details of our proposed new models, including bandwidth selection in kernel estimation for the NPR and TVCPR models and choice of truncation parameter in sieve estimation for the APR model. In Section 4, we compare the performance of these models on the prediction of stock returns with two main competing methods. Section 5 concludes the paper. The proofs of the main results are given in an appendix.

2 Predictive models and estimation theory

In this section, we describe the NPR, TVCPR and APR models in Sections 2.1-2.3, respectively. For each model, we establish the corresponding estimation theory and asymptotic properties.

2.1 The NPR model

Consider a nonparametric predictive regression model of the form

$$y_{t+j} = g_j(\tau_t, x_t) + e_{t+j},\ t = 1, 2, \ldots, n,\ j = 1, 2, \ldots, J, \tag{1}$$

where[2] $\tau_t = \frac{t}{n}$, $x_t = (x_t^1, \cdots, x_t^d)^\top$ is a locally stationary time series, $g_j(\cdot)$ are unknown functions of τ_t and x_t, and e_{t+j} is a α-mixing error process. This model allows for variation over time in the relationship between stock returns and the covariates x_t and is completely general in the form of the relationship. Typically, y_t is (logarithmic) stock returns, but we may also be interested in predicting prices. A locally stationary process is defined as follows (see Vogt (2012)).

Definition for Locally Stationary Process: Process $\{x_t\}$ is said to be locally stationary if for each scaled time point $\tau \in [0, 1]$ there exists an associated process $\{x_t(\tau)\}$ satisfying

[2]Robinson (1989) demonstrated that this "scaled time" requirement is necessary for the asymptotic justification of the nonparametric smoothing estimators.

(i) $\{x_t(\tau)\}$ is strictly stationary with density $f_{x_t(\tau)}(x)$;

(ii) it holds that

$$\|x_t - x_t(\tau)\| \leq \left(\left| \frac{t}{n} - \tau \right| + \frac{1}{n} \right) U_{nt}(\tau) \quad a.s., \tag{2}$$

where $U_{nt}(\tau)$ is a process of positive variables such that $\mathbb{E}[(U_{nt}(\tau))^\rho] < C$ for some $\rho > 0$ and $C > 0$ independent of τ, t and n, and $\|\cdot\|$ denotes an arbitrary norm on R^d.

It follows from the definition that a stationary process is also locally stationary. From the above definition, we see that local stationarity accommodates a variety of financial datasets.

We are also interested in predicting long horizon returns $\sum_{j=1}^J y_{t+j}$ using the covariates available up to and including time t. It follows from our specification that

$$y_{t:t+J} = \sum_{j=1}^J y_{t+j} = \sum_{j=1}^J g_j(\tau_t, x_t) + \sum_{j=1}^J e_{t+j} = g(\tau_t, x_t) + e_{t:t+J}, \tag{3}$$

where $g(\tau_t, x_t) = \sum_{j=1}^J g_j(\tau_t, x_t)$, and $e_{t:t+J} = \sum_{j=1}^J e_{t+j}$. Note however that $\text{cov}(e_{t:t+J}, e_{s:s+J}) \neq 0$ when $|t-s| < J$, which must be allowed for in the distribution theory.

For each fixed j and a given point (τ, x), we use the local constant kernel method to estimate $g_j(\tau, x)$ by

$$\widehat{g}_j(\tau, x) = \sum_{t=1}^n W_{nt}(\tau, x; h_j) y_{t+j} \quad \text{with} \quad W_{nt}(\tau, x; h_j) = \frac{K\left(\frac{\tau_t - \tau}{h_j}\right) \prod_{i=1}^d K\left(\frac{x_t^i - x^i}{h_j}\right)}{\sum_{s=1}^n K\left(\frac{\tau_s - \tau}{h_j}\right) \prod_{i=1}^d K\left(\frac{x_s^i - x^i}{h_j}\right)}, \tag{4}$$

where $x = (x^1, \cdots, x^d)^\top$ for any vector $x \in R^d$, $K(\cdot)$ is a probability kernel function and h_j is a bandwidth parameter. For convenience, in this paper, we work with a product kernel and assume that the bandwidth h_j is the same for τ and x^i $(i = 1, 2, \cdots, d)$, but the results can easily be extended to the case involving non–product kernels and different bandwidths. We then define our estimator of $g(\tau, x)$ to be the sum of the one dimensional estimators

$$\widehat{g}(\tau, x) = \sum_{j=1}^J \widehat{g}_j(\tau, x). \tag{5}$$

Let $f(\tau, x) = f_{x_t(\tau)}(x)$ denote the densities of the variables $x_t(\tau)$. Define $\partial_0 f(\tau, x) = \partial f(\tau, x)/\partial \tau$, $\partial_i f(\tau, x) = \partial f(\tau, x)/\partial x^i$, $\partial_0 g_j(\tau, x) = \partial g_j(\tau, x)/\partial \tau$, $\partial_i g_j(\tau, x) = \partial g_j(\tau, x)/\partial x^i$, $\partial^2_{0,0} g_j(\tau, x) = \partial^2 g_j(\tau, x)/\partial \tau^2$ and $\partial^2_{i,i} g_j(\tau, x) = \partial^2 g_j(\tau, x)/\partial x^{i2}$, for $i = 1, 2, \cdots, d$. Then we have the following theorems; their proofs[3] are given in Appendix A.1.

[3] **Tingting**: As explained in the proof of Theorem 2.1 below, two terms may be missing from $B_{j,x,\tau}$. This is because $g_j(\tau, x)$ is a multivariate function, so you will need to decompose $g_j(\tau_t, x + h_j w) - g_j(\tau, x) = g_j(\tau_t, x + h_j w) - g_j(\tau_t, x) + g_j(\tau_t, x) - g_j(\tau, x)$, and the second term is missing in your calculation or that by Vogt (2012). Can you be more careful this time ? In addition, the notation of $\partial_i g_j(\tau, x) \partial_i f(\tau, x)$ is confusing. Please see my notation in equations (7) and (8) below.

Theorem 2.1. Assume that Assumptions A.1.1–A.1.4 hold with $\beta \geq 4$. Let $n^r h_j^{d+2} \to \infty$ with $r = \min\{\rho, 1\}$, in which ρ is defined in (2). Moreover, suppose that $f(\tau, x) > 0$ and that $\sigma_j^2(x) = \mathbb{E}[e_{t+j}^2 | x_t = x]$ is continuous. Finally, let $r > \frac{d+2}{d+5}$ to ensure that the bandwidth h_j can be chosen to satisfy $nh_j^{d+5} \to c_h$ for a constant c_h. Then for each given j and (τ, x), as $n \to \infty$,

$$\sqrt{nh_j^{d+1}}\,(\widehat{g}_j(\tau, x) - g_j(\tau, x)) \to_D N(B_{j,\tau,x}, V_{j,\tau,x}), \tag{6}$$

where $B_{j,\tau,x} = \sqrt{c_h}\kappa_2/2 \sum_{i=0}^d [2\partial_i g_j(\tau, x)\partial_i f(\tau, x) + \partial_{i,i}^2 g_j(\tau, x) f(\tau, x)]/f(\tau, x)$ and $V_{j,\tau,x} = \kappa_0^{d+1}\sigma_j^2(x)/f(\tau, x)$ with $\kappa_0 = \int K^2(u)du$ and $\kappa_2 = \int u^2 K(u)du$.

The results are very similar to those for the standard local constant estimators with strictly stationary regressors (see Page 63–64 in Chapter 2 of Li and Racine (2007)). Note however that although we include rescaled time as a covariate, the large sample variance of the nonparametric estimator depends only on the short run variance of the error term, not on its long run variance. This is because the localization by the stochastic covariate effectively shuffles much of the dependence out of the error term.

Define

$$R_j(\tau, x) = \frac{\kappa_2}{2} h_j^2 \sum_{i=0}^d \left(2\frac{\partial g_j(\tau, x)}{\partial x_i}\frac{\partial_i f(\tau, x)}{\partial x_i} + \frac{\partial^2 g_j(\tau, x)}{\partial x_i^2} f(\tau, x) \right) / f(\tau, x), \tag{7}$$

$$b_j(\tau, x) = \frac{\partial g_j(\tau, x)}{\partial \tau} h_j\, \mu(K, \tau) + \frac{1}{2}\frac{\partial^2 g_j(\tau, x)}{\partial \tau^2} h_j^2\, \sigma^2(K, \tau), \tag{8}$$

where $\mu(K, \tau) = \int_0^\infty uK(u)du\, I[\tau = 0] + \int_{-\infty}^0 uK(u)du\, I[\tau = 1] + \int_{-\infty}^\infty uK(u)du\, I[0 < \tau < 1]$, and $\sigma^2(K, \tau) = \int_0^\infty u^2 K(u)du\, I[\tau = 0] + \int_{-\infty}^0 u^2 K(u)du\, I[\tau = 1] + \int_{-\infty}^\infty u^2 K(u)du\, I[0 < \tau < 1]$.

Let $B_J(\tau, x; h) = \sum_{j=1}^J (R_j(\tau, x) + b_j(\tau, x))$, $\Sigma_J(x) = \sum_{j=1}^J \rho_j^{-(d+1)}\sigma_j^2(x)$ and $V(\tau, x) = f^{-2}(\tau, x) f(x)\, \kappa_0 \int L^2(v)dv$, where $L(v) = \prod_{i=1}^d K(v^i)$.

We then establish an asymptotic property for $\widehat{g}(\tau, x)$ in the following theorem.

Theorem 2.2 Let Assumptions A.1.1–A.1.4 hold. Suppose that $\lim_{n\to\infty} nh^{d+1}\Sigma_J^{-1}(x) = \infty$ and $\lim_{n\to\infty} nh^{d+1}\, \Sigma_J^{-1}(x)\, B_J^2(\tau, x; h) < \infty$ for each given (τ, x). Then as $n \to \infty$,

$$\sqrt{nh^{d+1}\, \Sigma_J^{-1}(x)}\,(\widehat{g}(\tau, x) - g(\tau, x) - B_J(\tau, x; h)) \to_D N\left(0, V(\tau, x)\right). \tag{9}$$

Theorems 2.1 and 2.2 show that each of $g_j(\tau, x)$ can be consistently estimated and asymptotically normally distributed. Theorem 2.2 remains valid regardless of whether J is fixed or varying.

Some details for practical implementations (in particular, the choice of bandwidth h_j) are discussed in Section 3 before an empirical application is given in Section 4. The proofs of Theorems 2.1–2.2 and Corollary 2.1 are given in Appendix A.1 below.

2.2 The TVCPR model

Consider a time-varying coefficient predictive model of the form

$$y_{t+j} = x_t^\top \beta_j(\tau_t) + e_{t+j},\ t = 1, 2, \ldots, n,\ j = 1, 2, \ldots, J, \tag{10}$$

where x_t is a d-dimensional vector of nonstationary time series, $\tau_t = t/n$, $\beta_j(\cdot)$ are unknown functions defined on [0,1], and $\{e_{t+j}\}$ is a stationary error process. This model is a special case of the NPR model. It allows only conditionally linear predictability between the covariate and response, although the parameters of that relationship are allowed to vary over time in an arbitrary way.

For each given j, we use the local constant kernel method to estimate $\beta_j(\tau)$ by

$$\widehat{\beta}_j(\tau) = \left(\sum_{t=1}^{n} x_t x_t^\top K\left(\frac{\tau_t - \tau}{h_j} \right) \right)^{-1} \sum_{t=1}^{n} x_t y_{t+j} K\left(\frac{\tau_t - \tau}{h_j} \right), \tag{11}$$

where $K(\cdot)$ is a probability kernel function and h_j is a bandwidth parameter.

To develop the limit theory, we start with some regularity conditions to characterize the nonstationary time series x_t and the stationary error process e_{t+j}. We assume that x_t is a unit root process with generating mechanism $x_t = x_{t-1} + v_t$ and the initial value $x_0 = O_P(1)$. Then (e_{t+j}, v_t) are determined according to the linear process

$$w_{t,j} = (v_t^\top, e_{t+j})^\top = \sum_{s=0}^{\infty} \Phi_{s,j} \varepsilon_{t-s}, \tag{12}$$

where $\Phi_{s,j}$ is a sequence of $(d+1) \times (d+1)$ matrices, and ε_t is a sequence of independent and identically distributed random vectors with dimension $(d+1)$. Partition $\Phi_{s,j}$ as $\Phi_{s,j} = [\Phi_{s,1}, \Phi_{s,j,2}]^\top$ so that

$$v_t = \sum_{s=0}^{\infty} \Phi_{s,1}^\top \varepsilon_{t-s},\ e_{t+j} = \sum_{s=0}^{\infty} \Phi_{s,j,2}^\top \varepsilon_{t-s}. \tag{13}$$

By functional limit theory for a standardized linear process and noting that

$$n^{-1/2} \sum_{s=1}^{\lfloor nr \rfloor} \varepsilon_s \Rightarrow B_{\varepsilon,r}(\Gamma_0)$$

with $B_{\varepsilon,r}(\Gamma_0)$ being $(d+1)$-dimensional Brownian motion (BM) with variance matrix Γ_0, $\lfloor \cdot \rfloor$ denotes the integer part and $0 < r \leq 1$, we have for $t = \lfloor nr \rfloor$,

$$\frac{x_t}{\sqrt{n}} = \frac{1}{\sqrt{n}} \sum_{s=1}^{t} v_s + \frac{1}{\sqrt{n}} x_0 = \frac{1}{\sqrt{n}} \sum_{s=1}^{\lfloor nr \rfloor} v_s + o_p(1) \Rightarrow B_{d,r}(\Omega_v), \tag{14}$$

$$n^{-1/2}\sum_{s=1}^{\lfloor nr\rfloor} w_{s,j} \Rightarrow B_{d+1,r}(\Omega_j),\ n^{-1/2}\sum_{s=1}^{\lfloor nr\rfloor} e_{s+j} \Rightarrow B_r(\Omega_{e,j})$$

where $B_{d+1,r}(\Omega_j) = (B_{d,r}(\Omega_v)^\top, B_r(\Omega_{e,j}))^\top$ is $(d+1)$-dimensional BM with variance matrix Ω_j, and

$$\Omega_j = \Phi_j(1)^\top \Gamma_0 \Phi_j(1) = \begin{bmatrix} \Phi_1(1)^\top\Gamma_0\Phi_1(1) & \Phi_1(1)^\top\Gamma_0\Phi_{2,j}(1) \\ \Phi_{2,j}(1)^\top\Gamma_0\Phi_1(1) & \Phi_{2,j}(1)^\top\Gamma_0\Phi_{2,j}(1) \end{bmatrix} \equiv \begin{bmatrix} \Omega_v & \Omega_{ve,j} \\ \Omega_{ev,j} & \Omega_{e,j} \end{bmatrix}, \tag{15}$$

with $\Phi_j(1) = \sum_{s=1}^\infty \Phi_{s,j}$, $\Phi_1(1) = \sum_{s=1}^\infty \Phi_{s,1}$, and $\Phi_{2,j}(1) = \sum_{s=1}^\infty \Phi_{s,j,2}$. Here Ω_j is the partitioned long run variance matrix of $w_{t,j} = (v_t^\top, e_{t+j})^\top$.

We define $b \equiv b_\tau = B_{d,\tau}(\Omega_v)$ and set

$$q = \frac{b}{(b^\top b)^{1/2}} = \frac{b}{\|b\|}.$$

Let $q^\perp$ be a $d \times (d-1)$ orthogonal complement matrix such that $Q = [q, q^\perp]$, $Q^\top Q = I_d$, where I_d is the $d \times d$ identity matrix. The sample version of these quantities are given by

$$q_n = \frac{b_n}{(b_n^\top b_n)^{1/2}} = \frac{b_n}{\|b_n\|},\ b_n \equiv b_{n\tau} = \frac{1}{\sqrt{n}} x_{\tau(n)},$$

where $\tau(n) = \lfloor (\tau - h_j)n \rfloor$.

Let $Q_n = [q_n, q_n^\perp]$, $Q_n^\top Q_n = I_d$. Define $D_{nj} = \mathrm{diag}(n\sqrt{h_j}, (nh_j)I_{d-1})$. Write $B_{d+1,r}(\Omega_j) = \left[B_{d,r}(\Omega_v)^\top, B_r(\Omega_{e,j})\right]^\top$ and define

$$\Delta_\tau = \begin{bmatrix} \Delta_\tau(1) & \Delta_\tau(2) \\ \Delta_\tau(2)^\top & \Delta_\tau(3) \end{bmatrix},\ \ \Gamma_{\tau,j} = \begin{bmatrix} \Gamma_{\tau,j}(1) \\ \Gamma_{\tau,j}(2) \end{bmatrix}, \tag{16}$$

where the components of the partition are

$$\begin{aligned}
\Delta_\tau(1) &= b^\top b, \\
\Delta_\tau(2) &= \sqrt{2}(b^\top b)^{1/2} \left\{ \int_{-1}^1 B^\star_{d,(r+1)/2}(\Omega_v)^\top K(r) dr \right\} q^\perp, \\
\Delta_\tau(3) &= 2(q^\perp)^\top \left\{ \int_{-1}^1 B^\star_{d,(r+1)/2}(\Omega_v) B^\star_{d,(r+1)/2}(\Omega_v)^\top K(r) dr \right\} q^\perp, \\
\Gamma_{\tau,j}(1) &= (2b^\top b)^{1/2} \int_{-1}^1 K(r) dB^\star_{(r+1)/2}(\Omega_{e,j}), \\
\Gamma_{\tau,j}(2) &= 2(q^\perp)^\top \left\{ \int_{-1}^1 K(r) B^\star_{d,(r+1)/2}(\Omega_v) dB^\star_{(r+1)/2}(\Omega_{e,j}) + \frac{1}{2}\Delta_{ve,j} \right\},
\end{aligned}$$

where $B^\star_{d+1,r}(\Omega_j) = \left[B^\star_{d,r}(\Omega_v)^\top, B^\star_r(\Omega_{e,j})\right]^\top$ is an independent copy of $B_{d+1,r}(\Omega_j) = \left[B^\top_{d,r}(\Omega_v), B_r(\Omega_{e,j})\right]^\top$.

We then establish the following theorem to state the asymptotic distribution of $\widehat{\beta}_j(\tau)$; its proof is given in Appendix A.2.

Theorem 2.3. Suppose that Assumptions A.2.1-A.2.3 are satisfied and $n^2 h_j^{1+2\gamma_1} = o(1)$. Then for each given j, as $n \to \infty$,

$$D_{nj} Q_n^\top \left\{ \widehat{\beta}_j(\tau) - \beta_j(\tau) \right\} \Rightarrow \Delta_\tau^{-1} \Gamma_{\tau,j}, \tag{17}$$

where τ is fixed $0 < \tau < 1$ such that Δ_τ is nonsingular with probability 1.

2.3 The APR model

Consider a nonparametric additive predictive regression model of the form

$$y_{t+j} = \beta_j(\tau_t) + \sum_{i=1}^{d} g_j^i(x_t^i) + e_{t+j}, \ t = 1, 2, \ldots, n, \ j = 1, 2, \ldots, J, \tag{18}$$

where $\tau_t = t/n$, $\beta_j(\cdot)$ and $g_j^i(\cdot)$, for $i = 1, \cdots, d$, are unknown smooth functions, $x_t = (x_t^1, \cdots, x_t^d)^\top$ is a locally stationary process, and e_{t+j} is an error term. Here, $\beta_j(\cdot)$ is defined on $[0, 1]$. This model allows for nonlinear predictability from the covariates to the response and it allows for time variability through the intercept functions $\beta_j(\cdot)$. It is also a special case of the NPR model but is non-nested with the TVCPR model because it precludes interaction between rescaled time and covariates, whereas the TVCPR model allows a limited form of interaction of the covariate effect with time (the cross-partial with respect to x_j and τ_t is zero for the additive model). Without loss of generality and to simplify the notation, we assume that $d = 1$. So model (18) can be simplified as

$$y_{t+j} = \beta_j(\tau_t) + g_j(x_t) + e_{t+j}, \ t = 1, 2, \ldots, n, \ j = 1, 2, \ldots, J. \tag{19}$$

In this paper, we use the series estimation method to estimate all the unknown functions in model (19). Naturally, $\beta_j(\cdot)$ and $g_j(\cdot)$ belong to different function spaces as described below.

First, we assume that $\beta_j(\cdot) \in L^2[0, 1] = \{u(\tau) : \int_0^1 u(\tau) d\tau < \infty\}$, in which the inner product is given by $\langle u_1, u_2 \rangle = \int_0^1 u_1(\tau) u_2(\tau) d\tau$ and the induced norm is $\|u\|^2 = \langle u, u \rangle$. Let $\phi_0(\tau) = 1$, and for $s \geq 1$, $\phi_s(\tau) = \sqrt{2} \cos(\pi s \tau)$. Then $\{\phi_s(\tau)\}$ is an orthonormal basis in the Hilbert space $L^2[0, 1]$, and can be used to expand the unknown continuous function $\beta_j(\tau) \in L^2[0, 1]$ into an orthogonal series of the form:

$$\beta_j(\tau) = \sum_{s=0}^{\infty} c_{s,j,1} \ \phi_s(\tau), \ \text{where } c_{s,j,1} = \langle \beta_j(\tau), \phi_s(\tau) \rangle. \tag{20}$$

Note that $\{\phi_s(\tau)\}$ can be replaced by any other orthonormal basis in $L^2[0, 1]$.

In order to expand $g_j(x_t)$, suppose that the function $g_j(\cdot)$ is in Hilbert space $L^2(V, dF(x)) = \{q(x) : \int_V q^2(x)dF(x) < \infty\}$, where $F(x)$ is a distribution on the support V that may not be compact. The sequence $\{p_s(x), s \geq 0\}$ is an orthonormal basis in $L^2(V, dF(x))$, where an inner product is given by $\langle q_1, q_2\rangle = \int_V q_1(x)q_2(x)dF(x)$ and the induced norm is $\|q\|^2 = \langle q, q\rangle$. Hence, the unknown function $g_j(x)$ has an orthogonal series expansion in terms of the basis of $\{p_s(x), s \geq 0\}$,

$$g_j(x) = \sum_{s=0}^{\infty} c_{s,j,2}\ p_s(x), \text{ where } c_{s,j,2} = \langle g_j(x), p_s(x)\rangle. \tag{21}$$

Let k_{1j} and k_{2j} be two positive integers. Let $\beta_{k_{1j}}(\tau) = \sum_{s=1}^{k_{1j}} c_{s,j,1}\ \phi_s(\tau)$ be the truncation series of $\beta_j(\tau)$ with truncation parameter k_{1j}, and $\gamma_{k_{1j}} = \sum_{s=k_{1j}+1}^{\infty} c_{s,j,1}\ \phi_s(\tau)$ be the corresponding residual after truncation. It is easy to know that $\beta_{k_{1j}}(\tau) \to \beta_j(\tau)$ as $k_{1j} \to \infty$ in pointwise sense for smooth $\beta_j(\tau)$. Similarly, let $g_{k_{2j}}(x) = \sum_{s=0}^{k_{2j}-1} c_{s,j,2}\ p_s(x)$ and $\gamma_{k_{2j}} = \sum_{s=k_{2j}}^{\infty} c_{s,j,2}\ p_s(x)$ be the truncation series and the residual of $g_j(x)$, respectively. It follows that $g_{k_{2j}}(x) \to g_j(x)$, as $k_{2j} \to \infty$ under certain conditions.

Denote $\varphi_{k_{1j}}(\tau) = (\phi_1(\tau), \cdots, \phi_{k_{1j}}(\tau))^\top$ and $c_{1j} = (c_{1,j,1}, \cdots, c_{k_{1j},j,1})^\top$. Then we have $\beta_{k_{1j}}(\tau) = \varphi_{k_{1j}}(\tau)^\top c_{1j}$. Denote also $a_{k_{2j}}(x) = (p_0(x), \cdots, p_{k_{2j}-1}(x))^\top$ and $c_{2j} = (c_{0,j,2}, \cdots, c_{k_{2j}-1,j,2})^\top$. Accordingly, $g_{k_{2j}}(x) = a_{k_{2j}}(x)^\top c_{2j}$. Thus, model (19) can be written as

$$y_{t+j} = \varphi_{k_{1j}}(\tau_t)^\top c_{1j} + a_{k_{2j}}(x_t)^\top c_{2j} + \gamma_{k_{1j}}(\tau_t) + \gamma_{k_{2j}}(x_t) + e_{t+j}, \text{ for } t = 1, \cdots, n. \tag{22}$$

Let $y_{(j)} = (y_j, \cdots, y_{n+j})^\top$, $c_{(j)} = (c_{1j}^\top, c_{2j}^\top)$, $e_{(j)} = (e_j, \cdots, e_{n+j})^\top$, $\gamma_{(j)} = (\gamma_j(1), \cdots, \gamma_j(n))^\top$ where $\gamma_j(t) = \gamma_{k_{1j}}(\tau_t) + \gamma_{k_{2j}}(x_t)$, $t = 1, \cdots, n$, and

$$B_{nk_j} = \begin{pmatrix} \varphi_{k_{1j}}(\tau_1)^\top & a_{k_{2j}}(x_1)^\top \\ \vdots & \vdots \\ \varphi_{k_{1j}}(1)^\top & a_{k_{2j}}(x_n)^\top \end{pmatrix} \tag{23}$$

be an $n \times k_j$ matrix, where $k_j = k_{1j} + k_{2j}$. Then equation (22) can be written as

$$y_{(j)} = B_{nk_j} c_{(j)} + \gamma_{(j)} + e_{(j)}. \tag{24}$$

Then the ordinary least squares (OLS) estimator of $c_{(j)}$ is given by $\widehat{c}_{(j)} = (\widehat{c}_{1j}^\top, \widehat{c}_{2j}^\top)^\top = (B_{nk_j}^\top B_{nk_j})^{-1} B_{nk_j}^\top y_{(j)}$. Therefore, for any $\tau \in [0, 1]$ and $x \in V$, we define $\widehat{\beta}_j(\tau) = \varphi_{k_{1j}}(\tau)^\top \widehat{c}_{1j}$ and $\widehat{g}_j(x) = a_{k_{2j}}(x)^\top \widehat{c}_{2j}$ as the estimators of the unknown functions $\beta_j(\tau)$ and $g_j(x)$, respectively. As a result, we can further write the above results as

$$(\widehat{\beta}_j(\tau), \widehat{g}_j(x))^\top = \Phi_j(\tau, x)^\top \widehat{c}_{(j)}, \tag{25}$$

where $\Phi_j(\tau, x)$ is a block matrix given by

$$\Phi_j(\tau, x) = \begin{pmatrix} \varphi_{k_{1j}}(\tau) & \mathbf{0} \\ \mathbf{0} & a_{k_{2j}}(x) \end{pmatrix}. \tag{26}$$

Before establishing asymptotic properties for the estimators, we need some additional notation. Define $\Delta_{nj} = \left[\Phi_j(\tau, x)^\top U_{k_j}^{-1} V_{k_j} U_{k_j}^{-1} \Phi_j(\tau, x)\right]^{1/2}$, where U_{k_j} is a symmetric 2×2 block matrix of order $k_j \times k_j$ and V_{k_j} is a 2×2 symmetric block matrix of the form:

$$U_{k_j} = \begin{pmatrix} U_{11} & U_{12} \\ U_{12}^\top & U_{22} \end{pmatrix} \quad \text{and} \quad V_{k_j} = \begin{pmatrix} V_{11} & V_{12} \\ V_{12}^\top & V_{22} \end{pmatrix}. \tag{27}$$

in which $U_{11} = I_{k_{1j}}$, $U_{12} = \int_0^1 \varphi_{k_{1j}}(\tau)\mathbb{E}[a_{k_{2j}}(x_1(\tau))^\top]d\tau$ with elements $\int_0^1 \phi_i(\tau)\mathbb{E}[p_s(x_1(\tau))]d\tau$ for $1 \le i \le k_{1j}$, $0 \le s \le k_{2j} - 1$, and $U_{22} = \int_0^1 \mathbb{E}[a_{k_{2j}}(x_1(\tau))a_{k_{2j}}(x_1(\tau))^\top]d\tau$ with elements $\int_0^1 \mathbb{E}[p_i(x_1(\tau))p_s(x_1(\tau))^\top]d\tau$ for $i, s = 0, \cdots, k_{2j} - 1$, $V_{11} = \int_0^1 \varphi_{k_{1j}}(\tau)\varphi_{k_{1j}}(\tau)^\top \sigma^2(\tau)d\tau$, $V_{12} = \int_0^1 \varphi_{k_{1j}}(\tau)\sigma^2(\tau)\mathbb{E}[a_{k_{2j}}(x_1(\tau))^\top]d\tau$ and $V_{22} = \int_0^1 \sigma^2(\tau)\mathbb{E}[a_{k_{2j}}(x_1(\tau))a_{k_{2j}}(x_1(\tau))^\top]d\tau$.

We then establish the following theorems; their proofs are given in Appendix A.3.

Theorem 2.4. Suppose that uniformly over n, all the eigenvalues of U_{k_j} and V_{k_j} are positive, and that Assumptions A.3.1-A.3.6 hold. Then, for any $\tau \in [0, 1]$ and $x \in V$, as $n \to \infty$, we have

$$\Delta_{nj}^{-1} \begin{pmatrix} \sqrt{n}[\widehat{\beta}_j(\tau) - \beta_j(\tau)] \\ \sqrt{n}[\widehat{g}_j(x) - g_j(x)] \end{pmatrix} \to_D N(\mathbf{0}, I_2), \tag{28}$$

where $\mathbf{0}$ is a 2-dimensional zero column vector.

Define $m_j(\tau, x) = \beta_j(\tau) + g_j(x)$, $\widehat{m}_j(\tau, x) = \widehat{\beta}_j(\tau) + \widehat{g}_j(x)$, $m(\tau, x) = \sum_{j=1}^J m_j(\tau, x)$ and $\widehat{m}(\tau, x) = \sum_{j=1}^J \widehat{m}_j(\tau, x)$. Define $\Omega_{nj} = \Delta_{nj}\Delta_{nj} = \Phi_j(\tau, x)^\top U_{k_j}^{-1} V_{k_j} U_{k_j}^{-1} \Phi_j(\tau, x)$. Write

$$\Omega_{nj} = \begin{pmatrix} \Omega_{11,j} & \Omega_{12,j} \\ \Omega_{21,j} & \Omega_{22,j} \end{pmatrix}.$$

and $\Sigma_{nj} = \Omega_{11,j} + \Omega_{22,j} + 2\Omega_{12,j}$.

Theorem 2.5 Let Assumptions A.3.1–A.3.6 hold. Then as $n \to \infty$,

$$\sqrt{n}\,\Gamma_{nJ}^{-1/2}\left(\widehat{m}(\tau, x) - m(\tau, x)\right) \to_D N(0, 1), \tag{29}$$

where $\Gamma_{nJ} = \sum_{j=1}^J \Sigma_{nj}$.

Remark. (i) Note that Theorems 2.4 and 2.5 show that each of $\beta_j(\tau)$ and $g_j(x)$ can be consistently estimated and asymptotically normally distributed regardless of whether j is fixed or not. Moreover, $m(\tau, x)$ and $\overline{m}(\tau, x)$ can also be consistently estimated.

(ii) Note also that Theorem 2.5 remains valid when $J \to \infty$.

Section 3 below discusses about how to choose the truncation parameters k_j. The proofs of Theorems 2.4–2.5 and Corollary 2.2 are given in Appendix A.3 below.

3 Implementation

In this section, we will discuss computational details on the implementation of the NPR, APR and TVCPR models, particularly the bandwidth selection for the NPR and TVCPR models and the truncation parameter choice for the APR model.

3.1 Bandwidth selection

As we mentioned in Section 2, we use the local constant kernel method to estimate the unknown function $g_j(\cdot)$ in the NPR model and $\beta_j(\cdot)$ in the TVCPR model. It is generally accepted that the performance of the kernel estimator is mainly determined by bandwidth. In the last thirty years, there has been a comprehensive list of studies on the bandwidth selection. This section focuses on the issue of how to choose ρ_j and h involved in $h_j = \rho_j\, h$ used in the estimation of model (1). Similar discussion may be done for model (4).

Our approach is motivated by existing studies in Härdle et al. (1988), Härdle et al. (1989), Fan and Gijbels (1995), Xia and Li (2002) and Cheng et al. (2014). Let us introduce the following notation:

$$D_j(h_j) = \frac{1}{n}\sum_{t=1}^{n} \left(\widehat{g}_j(\tau_t, x_t) - g_j(\tau_t, x_t)\right)^2 w(\tau_t, x_t), \tag{30}$$

where $w(\cdot,\cdot)$ is a probability kernel function satisfying $\int_{-\infty}^{\infty}\int_0^1 w^2(\tau,u)d\tau\, du < \infty$.

Let $\widehat{h}_j$ be chosen such that it minimizes $D_j(h_j)$ over all possible $\{h_j\}$. Let h_{j0} be chosen such that it minimizes $d_j(h_j) = E\left[D_j(h_j)\right]$. In view of both the establishment and the proofs of the results in Xia and Li (2002), it can be shown that as $n \to \infty$

$$n^{\frac{3}{10}}\left(\frac{\widehat{h}_j}{h_{j0}} - 1\right) \to_D N(0, \sigma_{j0}^2) \tag{31}$$

for each fixed j, where $0 < \min_{j\geq 1}\sigma_{j0}^2 \leq \max_{j\geq 1}\sigma_{j0}^2 < \infty$, and $h_{j0} = \rho_j\, h_0$ with $\rho_j = j^\beta$ or θ^j, in which $h_0 > 0$, $\beta > 0$ and $\theta > 1$ will all be estimated in the rest of this section.

Using equation (31), we have for large enough n

$$\log\left(\frac{\widehat{h}_j}{h_{j0}}\right) = \log\left(1 + \frac{\widehat{h}_j}{h_{j0}} - 1\right) \approx \frac{\widehat{h}_j}{h_{j0}} - 1 \equiv n^{-\frac{3}{10}}\,\varepsilon_j, \tag{32}$$

where $\varepsilon_j = n^{\frac{3}{10}} \left(\frac{\widehat{h}_j}{h_{j0}} - 1 \right) \to_D N(0, \sigma_{j0}^2)$.

This suggests an approximate regression model of the form

$$\log\left(\widehat{h}_j\right) = \log\left(h_{j0}\right) + \eta_j = \log(h_0) + \log(\rho_j) + \eta_j \tag{33}$$
$$= \begin{cases} \log(h_0) + \beta \, \log(j) + \eta_j, & \text{if } \rho_j = j^\beta, \\ \log(h_0) + j \, \log(\theta) + \eta_j, & \text{if } \rho_j = \theta^j, \end{cases}$$

where $\eta_j = n^{-\frac{3}{10}} \varepsilon_j$ can be viewed as a sequence of random errors with $E[\eta_j] = 0$ and $0 < E\left[\eta_j^2\right] = n^{-\frac{3}{5}} \sigma_{j0}^2$.

We then focus the case of either $\rho_j = j^\beta$ or $\rho_j = \theta^j$. Let $Z_j = \log(\widehat{h}_j)$. For the case of $\rho_j = j^\beta$, we can estimate β by an ordinary least squares (OLS) estimator of the form

$$\widehat{\beta} = \left(\sum_{j=1}^{J} \left(\log(j) - \overline{\log(J)} \right)^2 \right)^{-1} \sum_{j=1}^{J} \left(\log(j) - \overline{\log(J)} \right) \left(Z_j - \overline{Z} \right), \tag{34}$$

where $\overline{\log(J)} = \frac{1}{J} \sum_{j=1}^{J} \log(j)$ and $\overline{Z} = \frac{1}{J} \sum_{j=1}^{J} Z_j$.

Equations (33) and (34) imply that the following rate of convergence:

$$\widehat{\beta} - \beta = O_P \left(\left(\sqrt{J \log(J)} \right)^{-1} \cdot n^{-\frac{3}{10}} \right). \tag{35}$$

For the case of $\rho_j = \theta^j$, the OLS estimator of $\gamma = \log(\theta)$ is given by

$$\widehat{\gamma} = \left(\sum_{j=1}^{J} \left(j - \overline{J} \right)^2 \right)^{-1} \sum_{j=1}^{J} \left(j - \overline{J} \right) \left(Z_j - \overline{Z} \right), \tag{36}$$

where $\overline{J} = \frac{1}{J} \sum_{j=1}^{J} j = \frac{(J+1)}{2}$.

Meanwhile, equations (33) and (36) imply a rate of convergence of the form:

$$\widehat{\gamma} - \gamma = O_P \left(J^{-\frac{3}{2}} \cdot n^{-\frac{3}{10}} \right). \tag{37}$$

We finally estimate h_0 by $\widehat{h}_0 = \frac{1}{J} \sum_{j=1}^{J} \widehat{h}_j \, \widehat{\rho}_j^{-1}$, where $\widehat{\rho}_j = j^{\widehat{\beta}}$ or $\widehat{\theta}^j$, in which $\widehat{\theta} = e^{\widehat{\gamma}}$.

Equations (35) and (37) imply that the OLS estimators may have fast rates. If we do choose $h_0 = n^{-\frac{1}{5}}$ and assume that $h_j \to 0$ as $(n, j) \to (\infty, \infty)$, there will be some restrictions on (J, n) such that either $J^{\widehat{\beta}} \cdot n^{-\frac{1}{5}} \to 0$ or $\widehat{\theta}^J \cdot n^{-\frac{1}{5}} \to 0$ as $(n, J) \to (\infty, \infty)$.

3.2 Truncation parameter choice

We use the series expansion method to estimate unknown functions $\beta_j(\cdot)$ and $g_j(\cdot)$ in the APR model. A key issue in using the series method in practice is the choice of truncation parameters

k_j $(k_{1j} + k_{2j})$ in the orthogonal expansions. Since there is no universal guide for the choice of such parameters, in this study, we choose the truncation parameters for the APR model through the out-of-sample mean squared errors. The procedure is given as follows.

- We divide the sample into two sets, the initialization set with sample size n_1 and validation set with sample size $n - n_1$.
- The initialization set is used to estimate the model for a given value of (k_{1j}, k_{2j}), then the estimated model is used to forecast the response variable in the validation set, based on which we compute the out-of-sample mean squared errors.
- We repeat the above procedure for all feasible values of (k_{1j}, k_{2j}).
- We then pick the optimal value of (k_{1j}, k_{2j}) which results in the smallest out-of-sample mean squared errors.

In the following section, we will evaluate the effectiveness of these models by investigating their capability of stock return prediction.

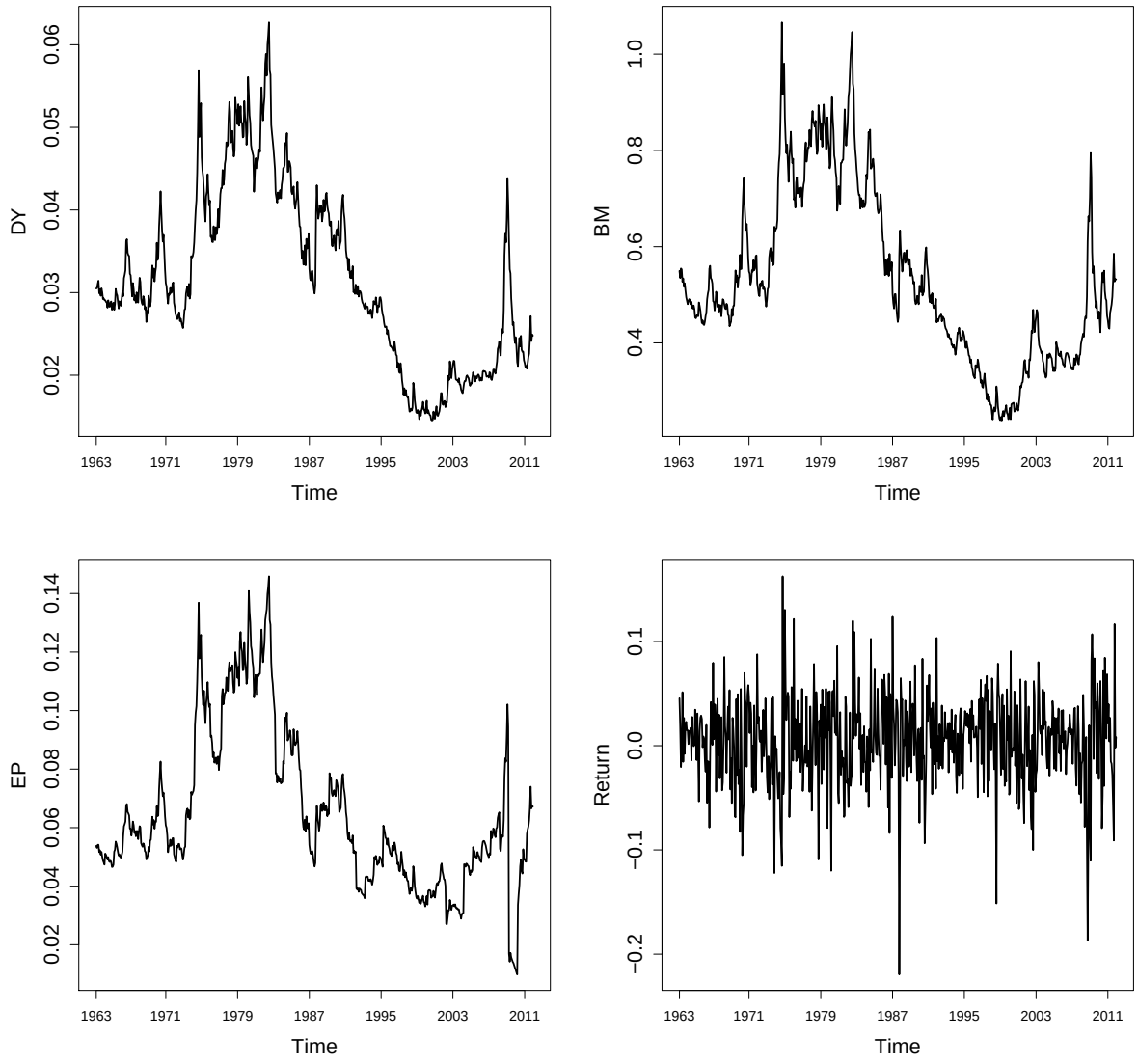
4 Stock return prediction using NPR, APR and TVCPR

In this section, we implement the NPR, APR and TVCPR models proposed in Section 2 to predict stock return using dividend yield, book-to-market ratio and earning-price ratio. The price and dividends data are from Center for Research in Security Prices (CRSP) data set, and we focus on the value-weighted NYSE index so as to be consistent with existing research. Dividend yield is calculated monthly on the value-weighted NYSE index, and it is defined as dividends paid over the prior year divided by the current level of index. The returns data are from April 1963 to December 2011 with a total number of 585 data points. We investigated the prediction for the excess value-weighted stock return (real return or excess return) which is defined by the value-weighted return minus t-bill rate. Let x_t^1, x_t^2 and x_t^3 denote the dividend yield, the book-to-market ratio and the earning-price ratio at time t, respectively. The time series plots of the dividend yield, book-to-market ratio, earning-price ratio and excess value-weighted stock returns are given in Figure 1.

In the following, we will examine the performance of the NPR, APR and TVCPR models for predicting the stock return. For comparison purposes, we also considered the commonly used historical mean model and the traditional linear predictive regression model. Therefore, we predict stock returns using the following five models:

- Mean: $y_{t+j} = \mu + e_{t+j}$;

Figure 1: Plot of dividend yield, book-to-market ratio, earning-price ratio and excess value-weighted stock returns.



- Linear: $y_{t+j} = \alpha_j + \beta_{1j}x_t^1 + \beta_{2j}x_t^2 + \beta_{3j}x_t^3 + e_{t+j}$;
- NPR: $y_{t+j} = g_j(\tau_t, x_t^1, x_t^2, x_t^3) + e_{t+j}$;
- APR: $y_{t+j} = g_j^0(\tau_t) + \sum_{i=1}^3 g_j^i(x_t^i) + e_{t+j}$;
- TVCPR: $y_{t+j} = \beta_{0j}(\tau_t) + \beta_{1j}(\tau_t)x_{1t} + \beta_{2j}(\tau_t)x_{2t} + \beta_{3j}(\tau_t)x_{3t} + e_{t+j}$.

Note that we use kernel method to estimate the unknown function $g_j(\cdot)$ in the NPR model and $\beta_{ij}(\cdot)$ in the TVCPR model, for $i = 0, 1, 2, 3$. As we know, the performance of the kernel estimator is mainly determined by the choice of bandwidth.

We use the series expansion method to estimate unknown functions $g_j^i(\cdot)$, for $i = 0, 1, 2, 3$, in the APR model. We define the truncation series with truncation parameter k_{ij} for $g_j^i(\tau)$ as $g_j^i(\tau, k_{ij}) = \sum_{s=1}^{k_{ij}} c_{s,j,i}\ \phi_s(\tau)$, for $i = 0, 1, 2, 3$, and let $c_{ij} = (c_{1,j,i}, \cdots, c_{k_{ij},j,i})^\top$ and $\phi_s(\tau)$ denote an orthonormal basis. Here we choose $\phi_s(\tau) = \sqrt{2}\cos(\pi s\tau)$ for $s \geq 1$. Then we estimate c_{ij}, for $i = 0, 1, 2, 3$, by the ordinary least squares method. As discussed in Section 3, in this study, we choose the truncation parameters for the APR model through the out-of-sample mean squared errors. For different prediction steps, we may obtain different truncation parameters. For example, we have $c_{(1)} = (3, 3, 1, 1)^\top$ and $c_{(36)} = (1, 1, 1, 1)^\top$.

In what follows, we will evaluate the performance of all of these models from both in-sample and out-of-sample performance.

4.1 Full sample estimation

In this section, we use the whole sample from April 1963 to December 2011 to evaluate the in-sample performance of all of these models in terms of the coefficient of determination. For a given predictive step j, the coefficient of determination can be calculated by

$$R_{IS,j}^2 = 1 - \frac{\sum_{t=1}^{n}(y_{t+j} - \widehat{y}_{t+j})^2}{\sum_{t=1}^{n}(y_{t+j} - \overline{y}_j)^2}, \tag{38}$$

where y_{t+j} is the observed stock return, $\widehat{y}_{t+j}$ is the corresponding predicted stock return and $\overline{y}_j = \frac{1}{n}\sum_{t=1}^{n} y_{t+j}$, which is also the predicted return from historical mean model. Thus for the historical mean model, $R_{IS,j}^2$ takes value of zero for all given values of j. From (38), it is easy to see that $R_{IS,j}^2$ can be written as

$$R_{IS,j}^2 = 1 - \frac{\text{MSE}_A}{\text{MSE}_M}, \tag{39}$$

where $\text{MSE}_M = 1/n\sum_{t=1}^{n}(y_{t+j} - \overline{y}_j)^2$ is the mean squared error of the historical mean model and $\text{MSE}_A = \sum_{t=1}^{n}(y_{t+j} - \widehat{y}_{t+j})^2$ is the mean squared error of an alternative model which produces the predicted value $\widehat{y}_{t+j}$. Therefore, $R_{IS,j}^2$ can also indicate the relative ratio of the mean squared errors between the historical mean model and the other models. If $R_{IS,j}^2$ for a certain model is positive, then this model performs better than the historical mean model. Simply speaking, the larger the $R_{IS,j}^2$ is, the better the corresponding model performs.

The results of $R_{IS,j}^2$ for different models with $j = 1, 6, 12, 18, 24, 36$ are presented in Table 1. To see the behavior of $R_{IS,j}^2$ for different prediction steps, we also produce the plot of $R_{IS,j}^2$ for these models with $j = 1, \cdots, 36$ in Figure 2. From Table 1 and Figure 2, we find the following facts.

- The historical mean model has the smallest $R^2_{IS,j}$ among all the competing models. This implies that in terms of in-sample fitting, the historical mean model has no advantage and can be easily beat by other models.

- The NPR, APR and TVCPR models have larger $R^2_{IS,j}$ than the traditional historical mean model and linear model, for $j = 1, 2, \cdots, 36$. This means that the NPR, APR and TVCPR models have better in-sample performance than the traditional parametric model.

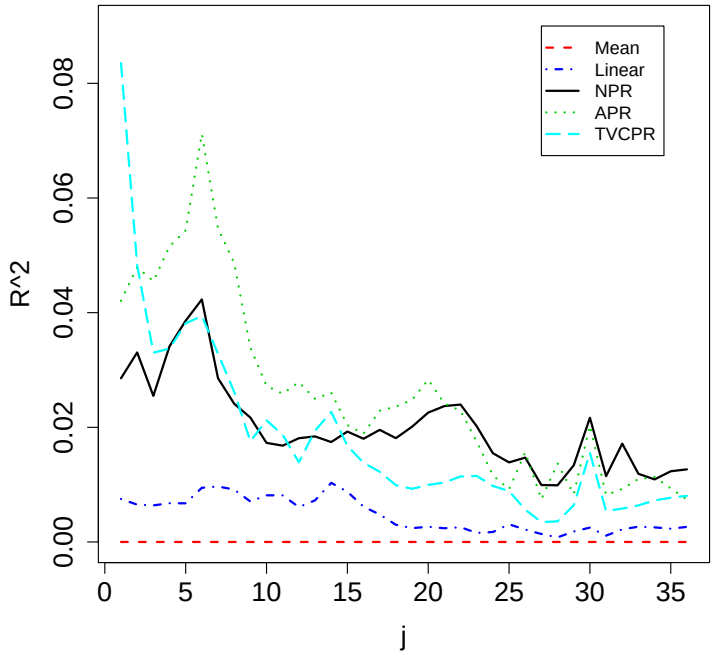
- When the prediction step is smaller than 22, the APR model has better performance than the NPR model, but when prediction step becomes large, the NPR and APR models have similar performance.

- When $j = 1$, the TVCPR model has the largest $R^2_{IS,j}$, which is 0.08355. Then with the increase of j, $R^2_{IS,j}$ of the TVCPR model decreases rapidly and is smaller than that of the APR model.

Table 1: Results of $R^2_{IS,j}$ for all the models.

Models	$j = 1$	$j = 6$	$j = 12$	$j = 18$	$j = 24$	$j = 36$
Mean	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
Linear	0.00751	0.00945	0.00609	0.00300	0.00173	0.00263
NPR	0.02855	0.04230	0.01810	0.01811	0.01548	0.01267
APR	0.04208	0.07118	0.02788	0.02360	0.01161	0.00740
TVCPR	0.08355	0.03941	0.01391	0.00991	0.00977	0.00804

Figure 2: Plot of $R^2_{IS,j}$ with $j = 1, 2, \cdots, 36$ for all the models.



From the results in Table 1 and Figure 2, we observe that the NPR, APR and TVCPR models have more advantages in terms of $R^2_{IS,j}$. We also plot the pictures of estimated functions and their 95% confidence intervals in Figure 3, including $\widehat{g}_j(\tau_t, x_t^1, x_t^2, x_t^3)$ in the NPR, $\widehat{\beta}_{ij}(\tau_t)$, for $i = 0, 1, 2, 3$ in the TVCPR and $\widehat{g}_j^0(\tau_t)$ and $\widehat{g}_j^i(x_t^i)$, for $i = 1, 2, 3$ in the APR model.

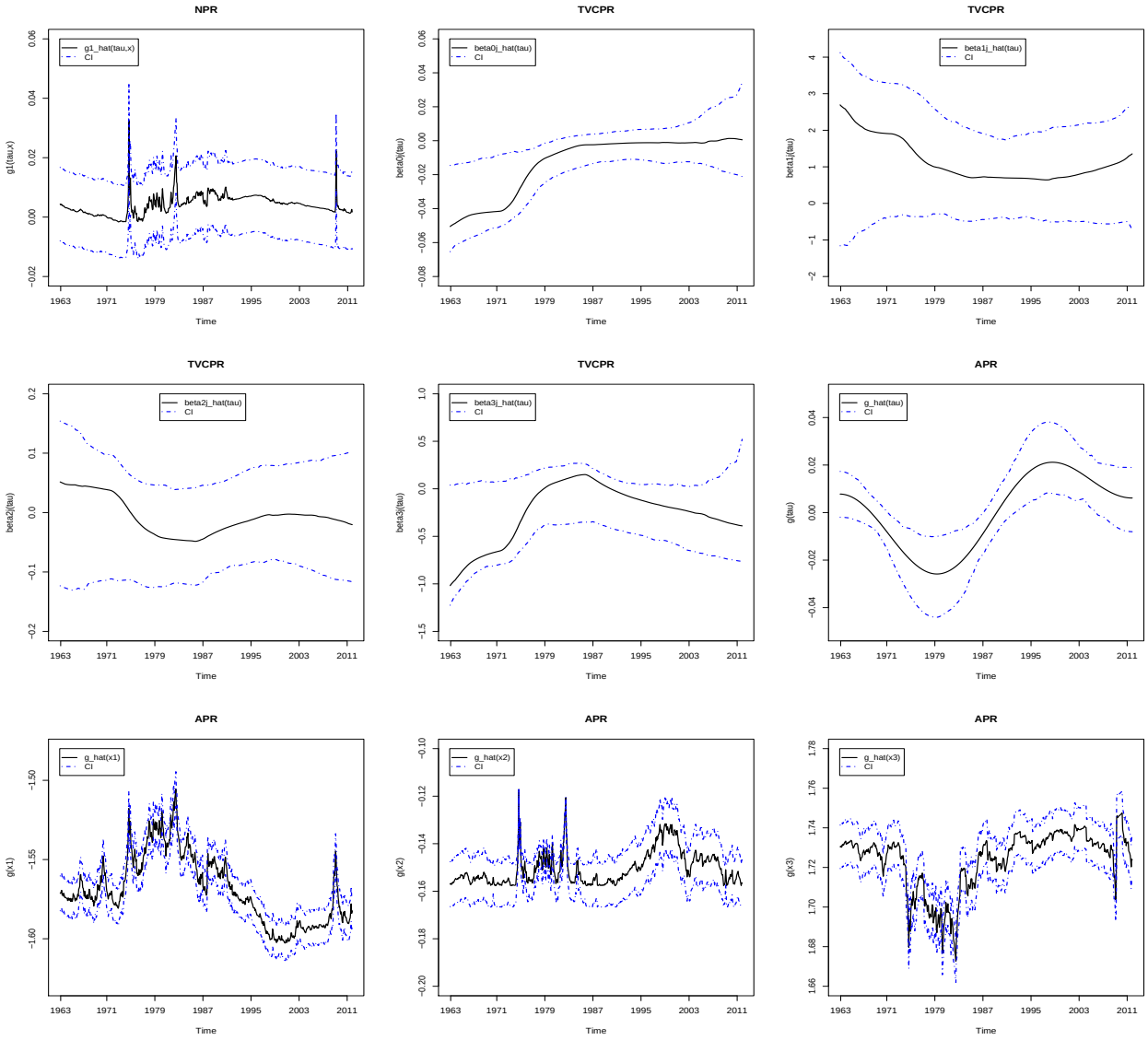
As we are more interested in the predicted returns of the models, in Figure 4, we plot the corresponding values produced by these models when $j = 1$ and $j = 3$. From Figure 4, we can see that the predicted returns by the NPR, APR and TVCPR models, in particular the APR model, are more volatile and are much closer to the true value of return than estimates generated by both the linear model and the historical mean model.

4.2 Out-of-sample evaluation

In the existing literature, the general conclusion is that the evidence for stock return predictability is predominantly in-sample while out-of-sample stock return forecast fails to beat the simple historical mean forecast (see for example Welch and Goyal (2008)). To check whether it is still true with the NPR, APR and TVCPR models, in this section, we evaluate the out-of-sample performance of these models using the following expansive window scheme. The details are described as follows.

- For the first window, we conduct the multi-step prediction based on $n-1$ observations. At the point x_n, we predict y_{n+1} using these $n-1$ pairs of observations $\{(x_1, y_2), (x_2, y_3), \cdots, (x_{n-1}, y_n)\}$.

Figure 3: Plot of estimated functions and 95% confidence intervals.



The estimated value of y_{n+1} is denoted as $\widehat{y}_{n+1}$. Then we use the observations

$$\{(x_1, y_3), (x_2, y_4), \cdots, (x_{n-2}, y_n), (x_{n-1}, \widehat{y}_{n+1})\}$$

to predict y_{n+2} at the point x_n. Similarly, we predict y_{n+3} at the point x_n using observations

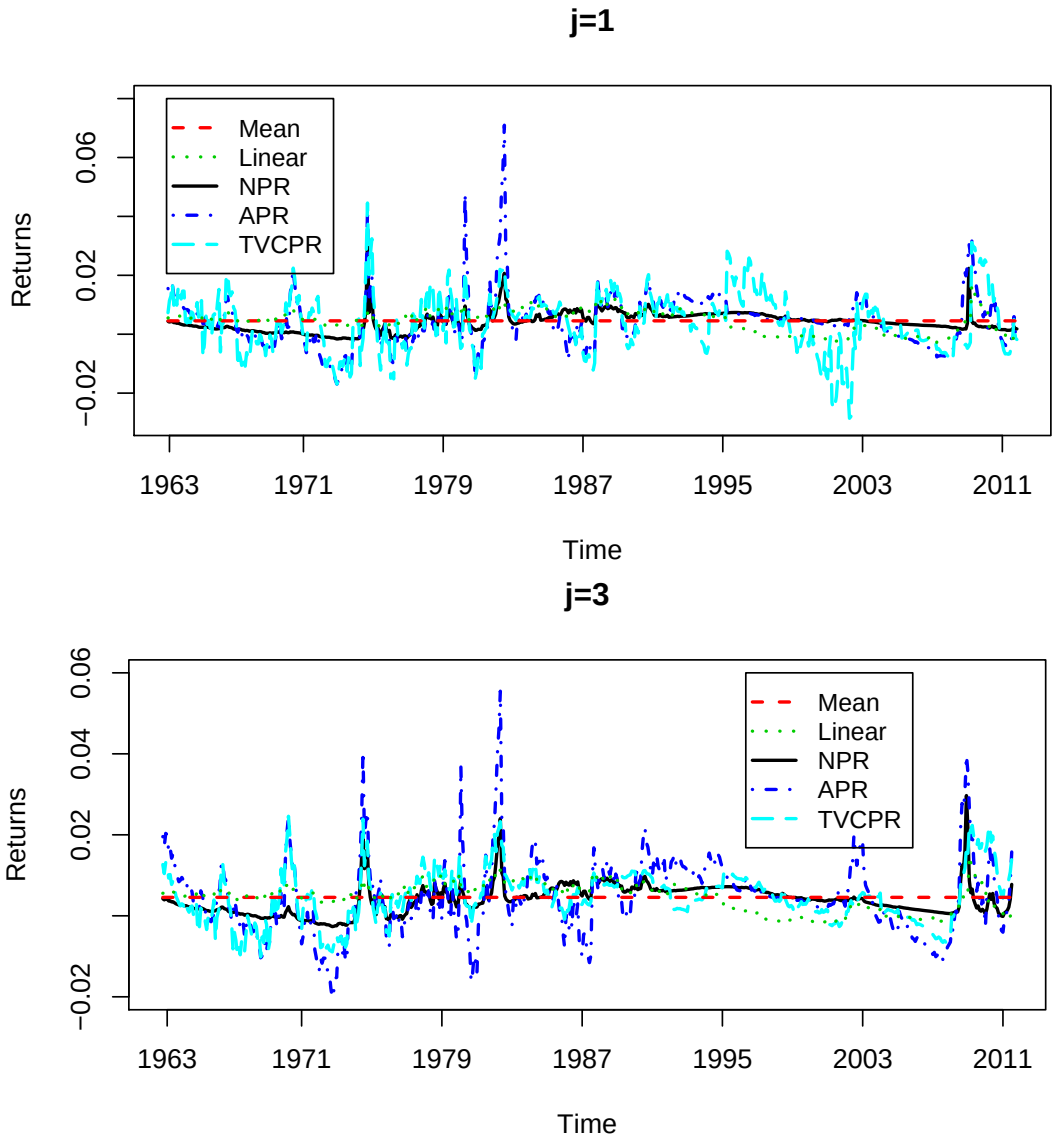
$$\{(x_1, y_4), (x_2, y_5), \cdots, (x_{n-2}, \widehat{y}_{n+1}), (x_{n-1}, \widehat{y}_{n+2})\}.$$

Repeating such procedure, we obtain the predicted return series for $y_{n+1}, y_{n+2}, \cdots, y_{n+J}$ denoted as

$$\widehat{y}_{n+1,1}, \widehat{y}_{n+1,2}, \cdots, \widehat{y}_{n+1,J}.$$

- The second window is obtained by expanding the first window to include x_n. At the point x_{n+1}, we conduct the multi-step prediction to predict $y_{n+2}, y_{n+3}, \cdots, y_{n+J+1}$ with the

Figure 4: Plots of predicted returns by all the models when $j = 1$ (top panel) and $j = 3$ (bottom panel).



predicted values denoted as

$$\widehat{y}_{n+2,1}, \widehat{y}_{n+2,2}, \cdots, \widehat{y}_{n+2,J}.$$

- The procedure continues until we obtain the Rth window. At the point x_{n+R-1}, we conduct the multi-step prediction for $y_{n+R}, y_{n+R+1}, \cdots, y_{n+R+J-1}$ and the predicted values are denoted as

$$\widehat{y}_{n+R,1}, \widehat{y}_{n+R,2}, \cdots, \widehat{y}_{n+R,J}.$$

We know that the out-of-sample forecast uses only the data available up to the time at which the forecast is made. Therefore, for a given predictive step j, following the work by Campbell

and Thompson (2008), we compute the out-of-sample R^2, which is defined as

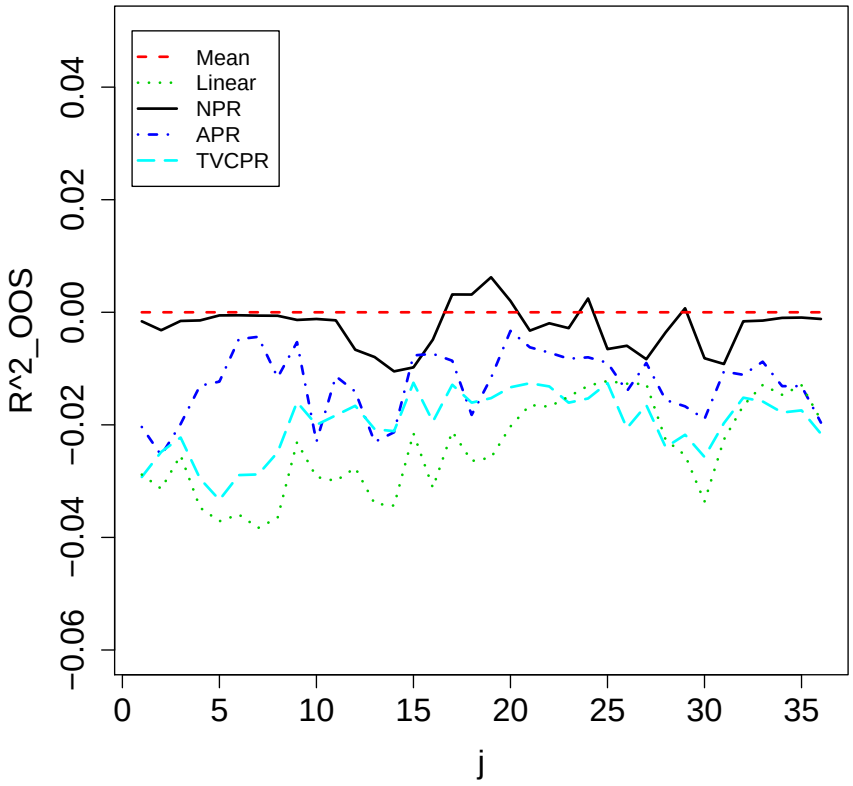
$$R^2_{OOS,j,n,R} = 1 - \frac{\sum_{r=1}^{R}(y_{n+r,j} - \widehat{y}_{n+r,j})^2}{\sum_{r=1}^{R}(y_{n+r,j} - \overline{y}_{n+r,j})^2},$$

where $\widehat{y}_{n+r,j}$ is the j-th step predicted return in the r-th window, $y_{n+r,j}$ is the corresponding observed return, $\overline{y}_{n+r,j}$ is the sample mean of observations using the information up to $n+r-1$, n is the sample size of the initial data to get a regression estimate at the start of evaluation period, and R is the total number of expansive windows. Here we choose $n = 241$, that is, we start the prediction of stock return in June 1983 and $R = 308$. The results of $R^2_{OOS,j,n,R}$ with $j = 1, 6, 12, 18, 24, 36$ are presented in Table 2. We also plot $R^2_{OOS,j,n,R}$ with j taking values from 1 to 36 in Figure 5. From Table 2 and Figure 5, we can find that (1) overall, linear regression model has the lowest $R^2_{OOS,j,n,R}$ and has no advantage compared with other competing models; (2) the NPR model performs better than the APR model and the APR model outperforms the TVCPR model for most of the predictive steps; (3) when the prediction step is between 17 and 20, the NPR model outperforms the historical mean model, but when the prediction step is small, they have similar performance.

Table 2: Results of $R^2_{OOS,j,n,R}$ for all the models.

Models	j=1	j=6	j=12	j=18	j=24	j=36
Mean	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
Linear	-0.02884	-0.03592	-0.02763	-0.02643	-0.01306	-0.01915
NPR	-0.00160	-0.00053	-0.00665	**0.00315**	**0.00245**	-0.00119
APR	-0.02037	-0.00478	-0.01409	-0.01824	-0.00800	-0.01960
TVCPR	-0.02926	-0.02893	-0.01661	-0.01606	-0.01531	-0.02158

Figure 5: Plot of $R^2_{OOS,j,n,R}$ with $j = 1, 2, \cdots, 36$ for all the models.



Apart from looking at behavior of $R^2_{OOS,j,n,R}$ of all of these models with the increase of predictive steps, we also looked at the cumulative out-of-sample R^2 for one particular given value of j, that is, we look at the performance of $R^2_{OOS,j,n,R}$ with the increase of R. We produce the plot for the cases of $j = 1$, $j = 12$ and $j = 24$ in Figure 6. Note that in Figure 6, we start the plot for $R \geq 12$ as it cannot tell much information when R is too small. From Figure 6, we can see that in the cases of $j = 1$ and $j = 12$, when R increases, the historical mean model beat other models, since the other four models have smaller cumulative out-of-sample R^2 than that of the historical mean model. However, when $j = 24$, we find that the NPR model has an absolute advantage compared with the other four models.

We also plot the out-of-sample predicted return when $j = 1$ and $j = 12$ in Figure 7, from which we can find that the NPR model generate more volatile predicted returns than the historical mean model.

Figure 6: Plots of cumulative $R^2_{OOS,j,n,R}$ with R ranging from 12 to 308 for all the models (top panel: j=1; middle panel: j=12; bottom panel: j=24).

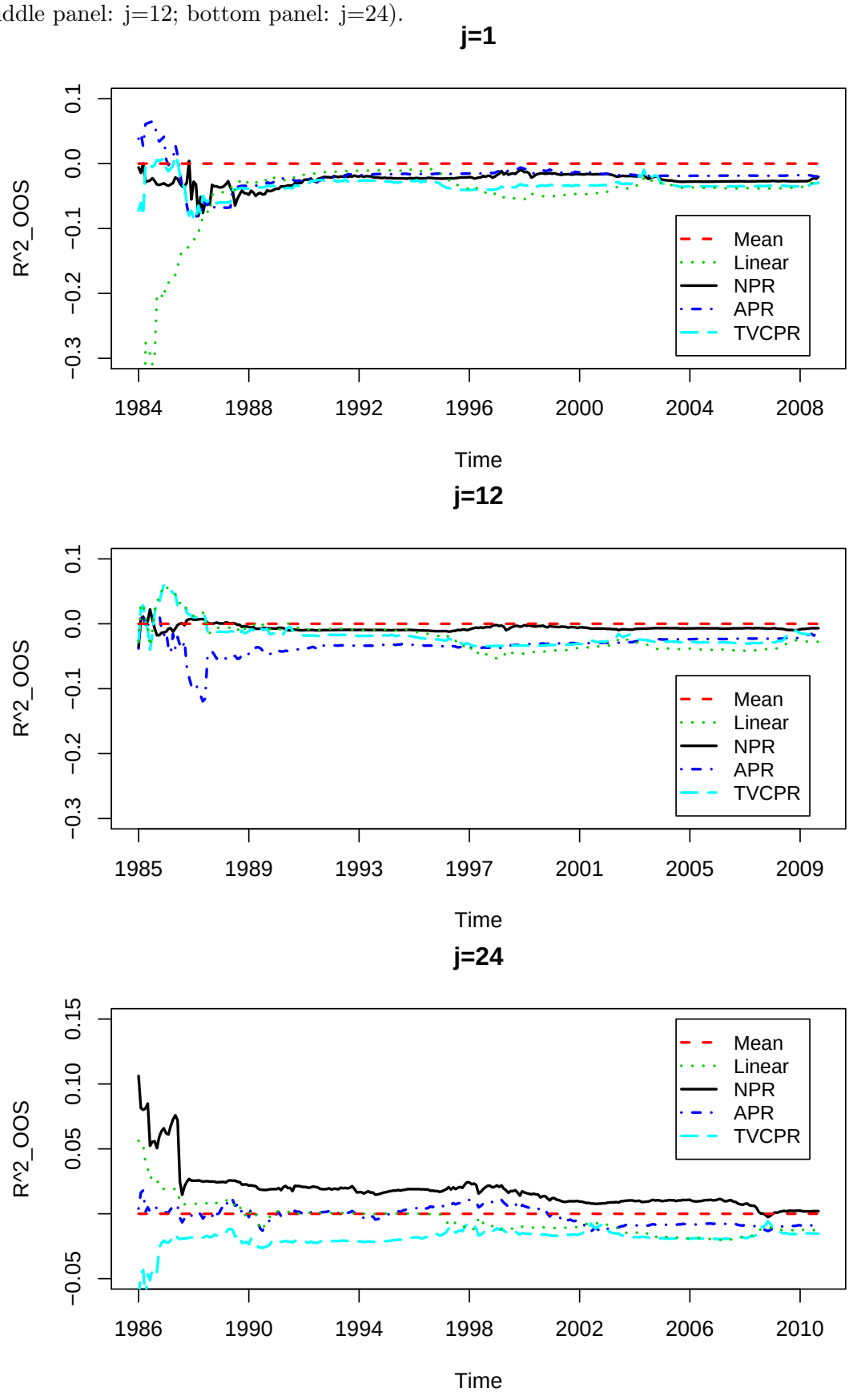
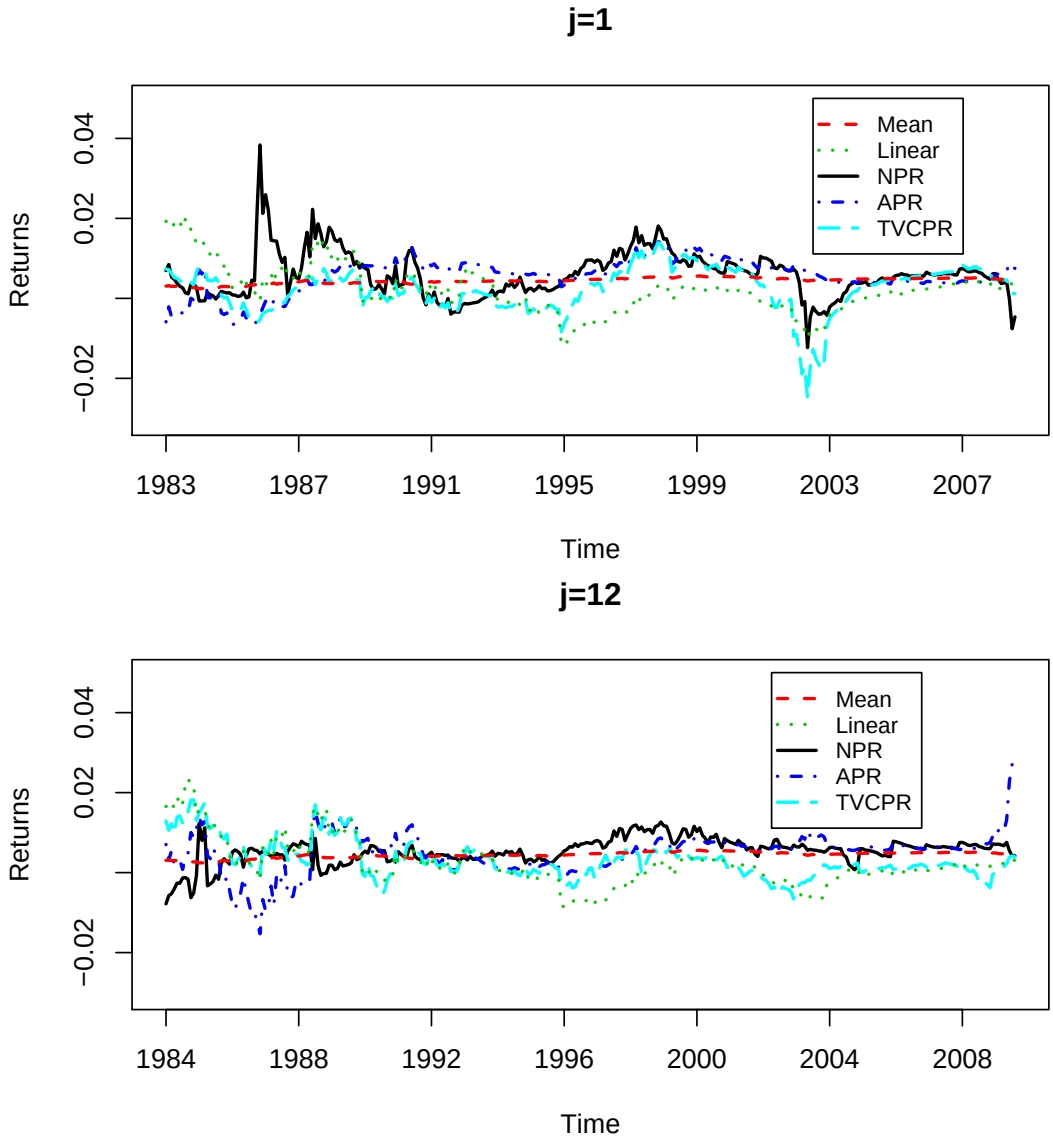


Figure 7: Plots of out-of-sample predicted returns for all the models (top panel: j=1; bottom panel: j=12).



4.2.1 Long Horizon Return Prediction

We also examined the out-of-sample prediction for long horizon returns $y_{n:n+J} = \sum_{j=1}^{J} y_{n+j}$. We define the out-of-sample R^2 as follows.

$$R^2_{OOS,J,n,R} = 1 - \frac{\sum_{r=1}^{R}(y^{(r)}_{n:n+J} - \widehat{y}^{(r)}_{n:n+J})^2}{\sum_{r=1}^{R}(y^{(r)}_{n:n+J} - \sum_{j=1}^{J}\overline{y}_{n+r,j})^2},$$

where $\widehat{y}^{(r)}_{n:n+J}$ denotes the estimated value of $y^{(r)}_{n:n+J}$ from the r-th expansive window. With $J = 2, 3, 4, 6, 12$, we present the results of $R^2_{OOS,J,n,R}$ in Table 3, from which we can find that when J is reasonably small, the NPR model performs best. When J takes values of 6 and 12, historical mean model performs best. Among all the cases, the linear regression model may be the last choice.

Table 3: Results of $R^2_{OS,Jn,R}$ for all the models.

Models	J=2	J=3	J=4	J=6	J=12
Mean	0.00000	0.00000	0.00000	0.00000	0.00000
Linear	-0.05407	-0.07740	-0.10983	-0.17632	-0.34089
NPR	**0.00151**	**0.01835**	**0.01446**	-0.01150	-0.02483
APR	-0.03876	-0.05722	-0.07078	-0.08493	-0.12825
TVCPR	-0.02599	-0.02783	-0.03857	-0.06367	-0.09306

We also computed the out-of-sample mean squared prediction errors for long horizon returns $y_{n:n+J} = \sum_{j=1}^{J} y_{n+j}$ given by

$$\text{MSE} = \frac{1}{R}\sum_{r=1}^{R}(y^{(r)}_{n:n+J} - \widehat{y}^{(r)}_{n:n+J})^2,$$

where $\widehat{y}^{(r)}_{n:n+J}$ is from the r-th expansive window.

With $J = 2, 3, 4, 6, 12$, we present the results of MSE in Table 4. From Table 4, we can see the effect of different horizon J on the prediction accuracy measured by the mean squared errors–MSEs. We find that when J is smaller than 4, the NPR model results in the smallest value of MSE. In other cases, the historical mean model performs best in predicting $y_{n:n+J}$.

Table 4: Results of MSE for all the models.

Models	J=2	J=3	J=4	J=6	J=12
Mean	0.00385	0.00579	0.00773	0.01179	0.02403
Linear	0.00406	0.00624	0.00858	0.01387	0.03223
NPR	**0.00385**	**0.00568**	**0.00762**	0.01192	0.02463
APR	0.00400	0.00612	0.00828	0.01279	0.02712
TVCPR	0.00395	0.00595	0.00803	0.01254	0.02627

4.3 Trading strategy

In this section, we propose an explicit trading strategy that switches between stocks and bonds based on whether predicted stock returns are greater than a threshold. We also compare this strategy with the buy and hold strategy that just holds stocks for the duration.

We first employ our proposed NPR, TVCPR and APR models to predict stock returns respectively, and obtain their corresponding one-step-ahead forecasts, then we compare these values with a chosen threshold. If the corresponding value is greater than the given threshold, we put money in stock market; Otherwise we buy a risk free bond with rate $r_0 = 0.02/12$ per month. In this study, we consider six different thresholds to examine the performance of our trading strategy with the buy and hold strategy in terms of profit. To check the robustness of our proposed trading strategy, we consider three investment starting dates, i.e., May 1983, May 1993, and May 2003. We also assume that the cost such as transaction fee during the trading could be ignored.

Tables 5– 7 show the results of stock return predictions that with NPR, TVCPR and APR models respectively. From these results, we can see that there always exists some thresholds under which our proposed strategies can outperform the buy and hold strategy in terms of profit. For example, for the NPR model, the thresholds are 0.001,0.002 and 0.003; for the TVCPR model, the thresholds are 0.001,0.005 and 0.006; and for the APR model, the thresholds are 0.001 and 0.002. As a result, we see that our propsoed trading strategies could be a better alternative of the buy and hold strategy in reality.

Table 5: Profit of trading strategy with the use of NPR model.

Starting date	Our trading strategy with different threshold						Buy and hold
	0.001	0.002	0.003	0.004	0.005	0.006	
1983 May	7.9413	7.9413	7.7871	8.5917	1.6772	2.6091	7.4383
1993 May	2.9617	2.9617	2.9617	2.6624	1.1923	1.9554	2.7188
2003 May	0.7895	0.7895	0.7895	0.6543	0.4523	0.1871	0.6324

Table 6: Profit of trading strategy with the use of TVCPR model.

Starting date	Our trading strategy with different threshold						Buy and hold
	0.001	0.002	0.003	0.004	0.005	0.006	
1983 May	11.9404	5.2708	6.2691	7.1274	8.9720	8.0030	7.4383
1993 May	3.7347	3.0305	3.0749	2.9902	3.0763	5.1143	2.7188
2003 May	1.0161	0.7895	0.7933	0.7501	0.7879	1.0190	0.6324

Table 7: Profit of trading strategy with the use of APR model.

Starting date	Our trading strategy with different threshold						Buy and hold
	0.001	0.002	0.003	0.004	0.005	0.006	
1983 May	10.4255	12.0673	9.1203	10.7768	9.4310	6.1093	7.4383
1993 May	2.8739	2.8739	2.5552	0.8061	0.2221	0.3799	2.7188
2003 May	0.7498	0.7498	0.6557	-0.0728	-0.2609	-0.1192	0.6324

5 Conclusions

In this paper, we have introduced the multi-step NPR and the APR models, in which the predictive variables are locally stationary time series; and the TVCPR, in which the predictive variables are nonstationary. Estimation theory and asymptotic properties have been established for all of these models in both the short horizon and long horizon case. Moreover, we have employed these models to investigate monthly stock return predictability over the period 1963-2011. The empirical results show that all of these models can substantially outperform the traditional

linear predictive regression model in terms of both in-sample and out-of-sample performance. In addition, we find that these models can always beat the historical mean model in terms of in-sample fitting, and also for some cases in terms of the out-of-sample forecasting. In particular, we find that the NPR model performs relatively well, especially at predicting two, three, and four month returns out of sample, where it beats all the alternative methods we have considered. We also showed how our methods can be used to deliver a trading strategy that beats the buy and hold strategy over our sample period.

Appendix

In this appendix, we provide the proofs of Theorem 2.1–Theorem 2.5. Section A.1–Section A.3 below provide the necessary assumptions and the proofs of the main results for the estimators in the NPR, TVCPR and APR models, respectively.

A.1. The NPR model

First, we present some assumptions for the establishment of asymptotic properties for $\widehat{g}_j(\tau, x)$, $g(\tau, x)$ for the NPR model.

Assumption A.1.1 (i) The process $\{x_t\}$ is locally stationary according to the definition in Section 2.1. (ii) It holds that $\max_{j\geq 1}\mathbb{E}|e_{t+j}|^s \leq C$ for some $s \geq 2$ and $C < \infty$. (iii) The array $\{(x_t, e_{t+1}, \cdots, e_{t+J}\}$ is α–mixing with mixing coefficient α satisfying $\alpha(k) \leq Ak^{-\beta}$ for some $A < \infty$ and $\beta > \frac{2s-2}{s-2}$.

Assumption A.1.2 (i) $g_j(\tau, x)$ is twice continuously partially differentiable with the first derivatives $\partial_i g_j(\tau, x)$ and second derivatives being denoted by $\partial^2_{is} g_j(\tau, x)$ for $i, s = 0, \cdots, d$. (ii) The densities $f(\tau, x) := f_{x_t(\tau)}(x)$ of the variables $x_t(\tau)$ are smooth in τ for each time point $\tau \in [0, 1]$. In particular, $f(\tau, x)$ is differentiable with respect to τ for each $x \in \mathbb{R}^d$, and the derivative $\partial_0 f(\tau, x) := \partial f(\tau, x)/\partial\tau$ is continuous. (iii) $f(\tau, x)$ is partially differentiable with respect to x for each $\tau \in [0, 1]$. The derivatives $\partial_i f(\tau, x) := \partial f(\tau, x)/\partial x^i$ are continuous for $i = 1, \cdots, d$.

Assumption A.1.3 Let f_{x_t} and $f_{x_t, x_{t+l}}$ be the densities of x_t and (x_t, x_{t+l}), respectively. For any compact set $S \subseteq \mathbb{R}^d$, there exists a constant $C = C(S)$ such that $\sup_t \sup_{x\in S} f_{x_t}(x) \leq C$ and $\sup_t \sup_{x\in S} \mathbb{E}[|e_{t+j}|^s | x_t = x] f_{x_t}(x) \leq C$. Moreover, there exists a natural number $l^\star < \infty$ such that for all $l \geq l^\star$, $\sup_t \sup_{x,x'\in S} \mathbb{E}[|e_{t+j}||e_{t+j+l}||x_t = x, x_{t+l} = x']f_{x_t, x_{t+l}}(x, x') \leq C$.

Assumption A.1.4 (i) The kernel function $K(\cdot)$ is bounded and has compact support, that is, $K(v) = 0$ for all $|v| > C_1$ with some $C_1 < \infty$. Also, the first moment is zero, that is, $\int vK(v)dv = 0$. Furthermore, K is Lipschitz continuous, that is, $|K(v) - K(v')| \leq L|v - v'|$ for some $L < \infty$ and all $v, v' \in \mathbb{R}$. (ii) Let $h_j = \rho_j h$, where each ρ_j is a positive constant and $\rho_j \to \infty$ as $j \to \infty$ and $h \to 0$ as $n \to \infty$. In addition, $nh_j \to \infty$ as $n \to \infty$.

Assumption A.1.1 allows us to approximate the locally stationary variable x_t by stationary variable $x_t(\tau)$ when τ_t is in a small neighborhood of τ. Assumption A.1.2(i) imposes smoothness condition on the unknown functions, which is to guarantee a certain rate of the convergence. Assumption A.1.4 is a standard assumption for kernel function $K(\cdot)$ and bandwidth h_j.

Proof of Theorem 2.1.

Observe that

$$\widehat{g}_j(\tau, x) - g_j(\tau, x) = \frac{1}{\widehat{f}(\tau, x)} \left(\widehat{g}_j^E(\tau, x) + \widehat{g}_j^B(\tau, x) - g_j(\tau, x)\widehat{f}(\tau, x) \right), \tag{40}$$

where we let $L(x) = \prod_{i=1}^d K(x^i)$ and then write

$$\begin{aligned}
\widehat{f}(\tau, x) &= \frac{1}{nh_j^{d+1}} \sum_{t=1}^n K\left(\frac{\tau_t - \tau}{h_j}\right) L\left(\frac{x_t - x}{h_j}\right), \\
\widehat{g}_j^E(\tau, x) &= \frac{1}{nh_j^{d+1}} \sum_{t=1}^n K\left(\frac{\tau_t - \tau}{h_j}\right) L\left(\frac{x_t - x}{h_j}\right) e_{t+j}, \\
\widehat{g}_j^B(\tau, x) &= \frac{1}{nh_j^{d+1}} \sum_{t=1}^n K\left(\frac{\tau_t - \tau}{h_j}\right) L\left(\frac{x_t - x}{h_j}\right) g_j(\tau_t, x_t).
\end{aligned}$$

Let $B_j(\tau, x) = \sqrt{nh_j^{d+1}} \left(\widehat{g}_j^B(\tau, x) - g_j(\tau, x)\widehat{f}(\tau, x) \right)$ denote the bias part and $V_j(\tau, x) = \sqrt{nh_j^{d+1}} \widehat{g}_j^E(\tau, x)$ denote the stochastic part.

Then we have $(\widehat{g}_j(\tau, x) - g_j(\tau, x)) = \left(nh_j^{d+1}\right)^{-1/2} \widehat{f}(\tau, x)^{-1} \left(V_j(\tau, x) + B_j(\tau, x)\right)$.

We then proceed with the following three steps to show the asymptotic normality of the estimator $\widehat{g}_j(\tau, x)$. The steps are similar to the proof[4] of Theorem 4.3 in Vogt (2012).

- Recall that $B_{j,\tau,x} = \sqrt{c_h}\kappa_2/2 \sum_{i=0}^d [2\partial_i g_j(u, x)\partial_i f(\tau, x) + \partial_{i,i}^2 g_j(\tau, x) f(\tau, x)]/f(\tau, x)$. Then from (iii) in the proof of Theorem 4.2 in Vogt (2012), we can show that $B_j(\tau, x)$ converges in probability to $B_{j,\tau,x}$.

- By using the block argument, that is, decomposing $V_j(\tau, x)$ alternately into big blocks and small blocks, we can neglect the small blocks and exploit the mixing conditions to replace

[4] **Tingting**: Please check the notation of $B_j(\tau, x)$ and $B_{j,\tau,x}$ to avoid any notational inconsistency between the notation used in the proofs and that introduced in Theorems 2.1 and 2.2.

the big blocks by independent random variables. Then apply a Lindeberg theorem, we can get that $V_j(\tau,x) \to_D N(0, \kappa_0^{d+1}\sigma^2(\tau,x)f(\tau,x))$, where $\kappa_0 = \int K^2(u)du$. The proof is in the same spirit as that for the standard strictly stationary setting. The variance of $V_j(\tau,x)$ can be calculated by the same steps as in Theorem 1 of Hansen (2008).

- It is easy to show that $\widehat{f}(\tau,x) - f(\tau,x) = o_P(1)$ and $\widehat{f}(\tau,x)^{-1} = O_P(1)$.

We then combine the above three points to complete the proof[5] of Theorem 2.1.

Proof of Theorem 2.2.

Observe that

$$
\begin{aligned}
(\widehat{g}_j(\tau,x) - g_j(\tau,x)) &= \left(nh_j^{d+1}\right)^{-1/2} \widehat{f}(\tau,x)^{-1} \left(V_j(\tau,x) + B_j(\tau,x)\right) \\
&= \left(nh^{d+1}\right)^{-1/2} f(\tau,x)^{-1}(1+o_P(1))\, \rho_j^{-(d+1)/2}\, V_j(\tau,x) \\
&+ \left(nh^{d+1}\right)^{-1/2} f(\tau,x)^{-1}(1+o_P(1))\, \rho_j^{-(d+1)/2}\, B_j(\tau,x),
\end{aligned} \tag{41}
$$

which gives

$$
\begin{aligned}
&\left(\sum_{j=1}^{J} \widehat{g}_j(\tau,x) - \sum_{j=1}^{J} g_j(\tau,x)\right) \\
&= \left(nh^{d+1}\right)^{-1/2} f(\tau,x)^{-1}(1+o_P(1)) \sum_{j=1}^{J} \rho_j^{-(d+1)/2}\, V_j(\tau,x) \\
&+ \left(nh^{d+1}\right)^{-1/2} f(\tau,x)^{-1}(1+o_P(1)) \sum_{j=1}^{J} \rho_j^{-(d+1)/2}\, B_j(\tau,x) \\
&\equiv \left(nh^{d+1}\right)^{-1/2} f(\tau,x)^{-1}(1+o_P(1))\, S_{nJ}(\tau,x) + \left(nh^{d+1}\right)^{-1/2} f(\tau,x)^{-1}(1+o_P(1))\, R_{nJ}(\tau,x),
\end{aligned} \tag{42}
$$

where

$$
\begin{aligned}
S_{nJ}(\tau,x) &= \sum_{j=1}^{J} \rho_j^{-(d+1)/2}\, V_j(\tau,x) \\
&= \left(nh^{d+1}\right)^{-1/2} \sum_{t=1}^{n} \left(\sum_{j=1}^{J} \rho_j^{-(d+1)} K\left(\frac{\tau_t - \tau}{h_j}\right) L\left(\frac{x_t - x}{h_j}\right) e_{t+j} \right),
\end{aligned} \tag{43}
$$

$$
\begin{aligned}
R_{nJ}(\tau,x) &= \sum_{j=1}^{J} \rho_j^{-(d+1)/2}\, B_j(\tau,x) \\
&= \left(nh^{d+1}\right)^{-1/2} \sum_{t=1}^{n} \left(\sum_{j=1}^{J} \rho_j^{-(d+1)} K\left(\frac{\tau_t - \tau}{h_j}\right) L\left(\frac{x_t - x}{h_j}\right) \left(g_j(\tau_t,x_t) - g_j(\tau,x)\right) \right).
\end{aligned} \tag{44}
$$

[5] **Tingting**: As mentioned briefly just above Theorem 2.1, you will need to provide a full proof. It looks to me that there is something missing from the bias term. You may like to have a look at the proof of Theorem 2.2 to see whether my derivations are correct or not.

It is obvious that $E[S_{nJ}(\tau,x)]=0$. It can also shown that

$$E\left[S_{nJ}^2(\tau,x)\right]=(1+o(1))\,f(x)\int L^2(u)du\int K^2(v)dv\sum_{j=1}^{J}\rho_j^{-(d+1)}\sigma_j^2(x). \tag{45}$$

In view of the α–mixing condition, using the big– and small– blocks approach, we can show that as $n\to\infty$

$$\left(\sum_{j=1}^{J}\rho_j^{-(d+1)}\sigma_j^2(x)\right)^{-1/2}S_{nJ}(\tau,x)\to_D N\left(0,f(x)\int L^2(u)du\int K^2(v)dv\right). \tag{46}$$

Meanwhile, we have a look at the bias term involving

$$\begin{aligned}
&\sum_{t=1}^{n}\left(\sum_{j=1}^{J}\rho_j^{-(d+1)}K\left(\frac{\tau_t-\tau}{h_j}\right)L\left(\frac{x_t-x}{h_j}\right)\left(g_j(\tau_t,x_t)-g_j(\tau,x)\right)\right)\\
&=\sum_{t=1}^{n}\left(\sum_{j=1}^{J}\rho_j^{-(d+1)}K\left(\frac{\tau_t-\tau}{h_j}\right)L\left(\frac{x_t-x}{h_j}\right)\left(g_j(\tau_t,x_t)-g_j(\tau_t,x)\right)\right)\\
&+\sum_{t=1}^{n}\left(\sum_{j=1}^{J}\rho_j^{-(d+1)}K\left(\frac{\tau_t-\tau}{h_j}\right)L\left(\frac{x_t-x}{h_j}\right)\left(g_j(\tau_t,x)-g_j(\tau,x)\right)\right)\\
&\equiv R_{1nJ}(\tau,x)+R_{2nJ}(\tau,x),
\end{aligned} \tag{47}$$

where

$$\begin{aligned}
&E\left[R_{1nJ}(\tau,x)\right]\\
&=\sum_{t=1}^{n}\left(\sum_{j=1}^{J}\rho_j^{-(d+1)}K\left(\frac{\tau_t-\tau}{h_j}\right)E\left[L\left(\frac{x_t-x}{h_j}\right)\left(g_j(\tau_t,x_t)-g_j(\tau_t,x)\right)\right]\right)\\
&=\sum_{t=1}^{n}\left(\sum_{j=1}^{J}\rho_j^{-(d+1)}K\left(\frac{\tau_t-\tau}{h_j}\right)\int L\left(\frac{u-x}{h_j}\right)\left(g_j(\tau_t,u)-g_j(\tau_t,x)\right)f(u)du\right)\\
&=\sum_{t=1}^{n}\sum_{j=1}^{J}\rho_j^{-(d+1)}h_j^dK\left(\frac{\tau_t-\tau}{h_j}\right)\left(\int L(w)\left(g_j(\tau_t,x+h_j\,w)-g_j(\tau_t,x)\right)f(x+h_j\,w)dw\right),\\
&=n(1+o(1))\sum_{j=1}^{J}\rho_j^{-(d+1)}h_j^d\\
&\times\left(\int_0^1 K\left(\frac{u-\tau}{h_j}\right)\left(\int L(w)\left(g_j(u,x+h_j\,w)-g_j(u,x)\right)f(x+h_j\,w)dw\right)du\right),
\end{aligned} \tag{48}$$

and[6] similarly,

$$E\left[R_{2nJ}(\tau,x)\right] \tag{49}$$

[6] **Tingting**: Can you calculate the bias terms in (48) and (49) to see whether the bias expression in Theorem 2.2 is correct or not ?

$$
\begin{aligned}
&= \sum_{j=1}^{J} \rho_j^{-(d+1)} \sum_{t=1}^{n} K\left(\frac{\tau_t - \tau}{h_j}\right) E\left[L\left(\frac{x_t - x}{h_j}\right)\right] \left(g_j(\tau_t, x) - g_j(\tau, x)\right) \\
&= \sum_{j=1}^{J} \rho_j^{-(d+1)} h_j^d \sum_{t=1}^{n} K\left(\frac{\tau_t - \tau}{h_j}\right) \left(g_j(\tau_t, x) - g_j(\tau, x)\right) \int L(w) f(x + h_j\, w) dw \\
&= n(1 + o(1)) \sum_{j=1}^{J} \rho_j^{-(d+1)} h_j^d \int L(w) f(x + h_j\, w) dw \\
&\times \left(\int_0^1 K\left(\frac{v - \tau}{h_j}\right) \left(g_j(v, x) - g_j(\tau, x)\right) dv\right).
\end{aligned}
$$

Therefore, equations (41)–(49) complete the proof of Theorem 2.2.

A.2. The TVCPR model

In order to establish asymptotic properties for $\widehat{\beta}_j(\cdot)$, we impose the following assumptions.

Assumptions A.2.1 Let ε_t be a $d+1$–dimensional vector of independent and identically distributed random variables with $\mathbb{E}[\varepsilon_t] = 0$, $\Gamma_0 \equiv \mathbb{E}[\varepsilon_t \varepsilon_t^\top] > 0$, and $\mathbb{E}[\|\varepsilon_t\|^{4+\gamma_0}] < \infty$ for $\gamma_0 > 0$. The linear process coefficient matrices satisfy $\sum_{s=0}^{\infty} s\|\Phi_{s,j}\| < \infty$.

Assumptions A.2.2 $\beta_j(\cdot)$ is continuous with $|\beta_j(\tau + z) - \beta_j(\tau)| = O(|z|^{\gamma_1})$ as $z \to 0$ for some $\frac{1}{2} < \gamma_1 \leq 1$.

Assumptions A.2.3 (i) The kernel function $K(\cdot)$ is continuous, positive and has compact support $[-1, 1]$ with $\int_{-1}^{1} K(v)dv = 1$, and the first moment is zero, that is, $\int_{-1}^{1} vK(v)dv = 0$. (ii) The bandwidth h_j satisfies $h_j \to 0$ and $nh_j \to \infty$ as $n \to \infty$.

Assumption A.2.1 is a standard assumption for linear process. Assumptions A.2.2 and A.2.3 impose a smoothness condition on the functional coefficient $\beta_j(\cdot)$ and some commonly-used conditions on the kernel function and bandwidth.

Proof of Theorem 2.3. The proof follows that of Theorem 3.1 of Phillips et al. (2017).

A.3. The APR model

In order to establish asymptotic properties for $\widehat{\beta}_j(\tau)$ and $\widehat{g}_j(x)$, we introduce the following assumptions.

Assumption A.3.1 (i) $\{x_t\}$ is locally stationary with associated process $\{x_t(\tau)\}$, and all x_t $(1 \leq t \leq n)$ have the same compact support $V = [a_{\min}, a_{\max}]$. Moreover, the density $f(\tau, x)$ of $x_t(\tau)$ is smooth in τ. (ii) For each $\tau \in [0, 1]$, $x_t(\tau)$ is a strictly stationary and α-mixing

process with mixing coefficient $\alpha(i)$ such that $\sum_{i=1}^{\infty}\alpha^{\delta/(2+\delta)}(i)<\infty$ for some $\delta>0$. For $u\neq\tau\in[0,1]$, $x_t(\tau)$ and $x_s(u)$ are uncorrelated for any t and s.

Assumption A.3.2 There exists an orthogonal function sequence $\{p_i(x), i\geq 0\}$ on the support $[a_{\min}, a_{\max}]$ with respect to $dF(x)$ such that $\sup_{\tau\in[0,1]}\sup_{j\geq 0}\mathbb{E}|p_j(x_1(\tau))|<\infty$.

Assumption A.3.3 For all t and any $\tau\in[0,1]$, $x_t(\tau)$ is independent of $\{e_s, -\infty<s<\infty\}$.

Assumption A.3.4 Suppose that there is a filtration sequence $\mathcal{F}_{nt}$ such that $(e_t, \mathcal{F}_{n,t})$ form a martingale difference sequence. Meanwhile, $\mathbb{E}(e_t^2|\mathcal{F}_{n,t-1})=\sigma^2(\tau_t)$ almost surely with continuous and nonzero function $\sigma(\cdot)$ and for some $q\geq 4$, $\max_{1\leq t\leq n}\mathbb{E}(|e_t|^q|\mathcal{F}_{n,t-1})<\infty$.

Assumption A.3.5 (i)The functions $\beta_j(\cdot)$ and $g_j(\cdot)$ are continuously differentiable up to s_1 and s_2, respectively. (ii)For $\beta_j(\cdot)$ function, let $\int_0^1\beta_j(r)dr=0$.

Assumption A.3.6 Suppose that as $n\to\infty$, (i) $nk_{1j}^{-(2s_1-1)}=o(1)$ and $nk_{2j}^{-(2s_2-1)}=o(1)$ and (ii) $nk_{2j}k_{1j}^{-2s_1}=o(1)$, $nk_{1j}k_{2j}^{-s_2}=o(1)$.

Assumptions A.3.1–A.3.4 allow us to approximate the locally stationary variable x_t by stationary variable $x_t(\tau)$ when τ_t is in a small neighborhood of τ. In this paper, we require the support of the locally stationary process to be compact. Assumption A.3.5 (i) imposes a smoothness condition on the unknown functions, which is to guarantee a certain rate of the convergence. Assumption A.3.5(ii) is an identification condition since in both the expansions of $\beta_j(\cdot)$ and $g_j(\cdot)$, there is a constant term that could not be distinguished one from another in the regression. Assumption A.3.6 imposes the rates of divergence on k_{1j} and k_{2j}, which guarantee the convergence of the proposed estimators.

Proof of Theorem 2.4.

Let $D_{nj}=\mathrm{diag}(\sqrt{n}I_{k_{1j}}, \sqrt{n}I_{k_{2j}})$ denote a diagonal matrix of $k_j\times k_j$ with $k_j=k_{1j}+k_{2j}$. From Lemma A.3 of Dong and Linton (2016), we have that $\|D_{nj}^{-1}B_{nk_j}^{\top}B_{nk_j}D_{nj}^{-1}-U_{k_j}\|=o_P(1)$, then we have

$$\begin{aligned}\widehat{c}_{(j)}&=(B_{nk_j}^{\top}B_{nk_j})^{-1}B_{nk_j}^{\top}y_{(j)}=(B_{nk_j}^{\top}B_{nk_j})^{-1}B_{nk_j}^{\top}(B_{nk_j}c_{(j)}+\gamma_{(j)}+e_{(j)})\\&=c_{(j)}+(B_{nk_j}^{\top}B_{nk_j})^{-1}B_{nk_j}^{\top}(\gamma_{(j)}+e_{(j)}).\end{aligned}$$

Thus

$$\begin{aligned}\widehat{c}_{(j)}-c_{(j)}&=(B_{nk_j}^{\top}B_{nk_j})^{-1}B_{nk_j}^{\top}(\gamma_{(j)}+e_{(j)})=D_{nj}^{-1}(D_{nj}^{-1}B_{nk_j}^{\top}B_{nk_j}D_{nj}^{-1})^{-1}D_{nj}^{-1}B_{nk_j}^{\top}(\gamma_{(j)}+e_{(j)})\\&=D_{nj}^{-1}(U_{kj}+o_P(1))^{-1}D_{nj}^{-1}B_{nk_j}^{\top}(\gamma_{(j)}+e_{(j)})=D_{nj}^{-1}(U_{kj}^{-1}+o_P(1))D_{nj}^{-1}B_{nk_j}^{\top}(\gamma_{(j)}+e_{(j)}).\end{aligned}$$

Then we have

$$D_{nj}(\widehat{c}_{(j)} - c_{(j)}) = (U_{k_j}^{-1} + o_P(1))D_{nj}^{-1}B_{nk_j}^{\top}(\gamma_{(j)} + e_{(j)}).$$

Then, for any $\tau \in [0,1]$ and $x \in V$,

$$\begin{aligned}\begin{pmatrix} \sqrt{n}[\widehat{\beta}_j(\tau) - \beta_j(\tau)] \\ \sqrt{n}[\widehat{g}_j(x)] - g_j(x)] \end{pmatrix} &= \Phi_j(\tau,x)^{\top}D_{nj}(\widehat{c}_{(j)} - c_{(j)}) + \begin{pmatrix} \sqrt{n}\gamma_{k_{1j}}(\tau) \\ \sqrt{n}\gamma_{k_{2j}}(x) \end{pmatrix} \\ &= \Phi_j(\tau,x)^{\top}U_{k_j}^{-1}D_{nj}^{-1}B_{nk_j}^{\top}(\gamma_{(j)} + e_{(j)}) + \begin{pmatrix} \sqrt{n}\gamma_{k_{1j}}(\tau) \\ \sqrt{n}\gamma_{k_{2j}}(x) \end{pmatrix}.\end{aligned}$$

We then proceed with two main steps as follows.

- First, we can establish the asymptotic normality from $\Phi_j(\tau,x)^{\top}U_{k_j}^{-1}D_{nj}^{-1}B_{nk_j}^{\top}e_{(j)}$ by Cramér-Wold theorem.
- Second, we can show that the remainder terms are asymptotically negligible.

For the proof of normality, we can write that $\Phi_j(\tau,x)^{\top}U_{k_j}^{-1}D_{nj}^{-1}B_{nk_j}^{\top}e_{(j)} = \sum_{t=1}^{n}\eta_{nt}e_{t+j}$, where

$$\eta_{nt} = \Phi_j(\tau,x)^{\top}U_{k_j}^{-1}D_{nj}^{-1}\begin{pmatrix} \phi_{k_{1j}}(\tau_t) \\ a_{k_{2j}}(x_t) \end{pmatrix}.$$

Recall that $\Delta_{nj} = \left[\Phi_j(\tau,x)^{\top}U_{k_j}^{-1}V_{k_j}U_{k_j}^{-1}\Phi_j(\tau,x)\right]^{1/2}$. By Cramér-Wold theorem and Corollary 3.1 of Hall and Heyde (1980), we can prove that $\Delta_{nj}^{-1}\sum_{t=1}^{n}\eta_{nt}e_{t+j} \to_D N(0, I_{k_j})$. The details are similar to the proofs of Theorem 3.1 and 3.2 in Dong and Linton (2016).

Proof of Theorem 2.5.

Define $\Omega_{nj} = \Delta_{nj}\Delta_{nj} = \Phi_j(\tau,x)^{\top}U_{k_j}^{-1}V_{k_j}U_{k_j}^{-1}\Phi_j(\tau,x)$.

Theorem 2.4 implies that for large enough n, we have

$$\begin{pmatrix} \sqrt{n}[\widehat{\beta}_j(\tau) - \beta_j(\tau)] \\ \sqrt{n}[\widehat{g}_j(x)] - g_j(x)] \end{pmatrix} \approx_D N(\mathbf{0}, \Omega_{nj}).$$

Let

$$\Omega_{nj} = \begin{pmatrix} \Omega_{11,j} & \Omega_{12,j} \\ \Omega_{21,j} & \Omega_{22,j} \end{pmatrix}.$$

Then we have

$$\sqrt{n}\left(\widehat{\beta}_j(\tau) + \widehat{g}_j(x) - \beta_j(\tau) - g_j(x)\right) \approx_D N(\mathbf{0}, \Sigma_{nj}),$$

where $\Sigma_{nj} = \Omega_{11,j} + \Omega_{22,j} + 2\Omega_{12,j}$.

Define $m_j(\tau, x) = \beta_j(\tau) + g_j(x)$ and $\widehat{m}_j(\tau, x) = \widehat{\beta}_j(\tau) + \widehat{g}_j(x)$.

$$\sqrt{n}\left(\widehat{m}_j(\tau, x) - m_j(x, \tau)\right) \approx_D N(\mathbf{0}, \Sigma_{nj}),$$

By the following definitions:

$$\widehat{m}(\tau, x) = \sum_{j=1}^{J} \widehat{m}_j(\tau, x) \quad \text{and} \quad m(\tau, x) = \sum_{j=1}^{J} m_j(\tau, x),$$

We then have as $n \to \infty$

$$\sqrt{n}\Sigma_{nJ}^{-1/2}\left(\widehat{m}(\tau, x) - m(\tau, x)\right) \to_D N(0, 1), \tag{50}$$

where $\Sigma_{nJ} = \sum_{j=1}^{J} \Sigma_{nj}$.

References

Cai, Z. (2007), 'Trending time–varying coefficient time series models with serially correlated errors', *Journal of Econometrics* **136**(1), 163–188.

Cai, Z., Li, Q. and Park, J. (2009), 'Functional–coefficient models for nonstationary time series data', *Journal of Econometrics* **148**(2), 101–113.

Campbell, J. Y. and Shiller, R. J. (1988), 'The dividend–price ratio and expectations of future dividends and discount factors', *Review of Financial Studies* **1**(3), 195–228.

Campbell, J. Y. and Thompson, S. B. (2008), 'Predicting excess stock returns out of sample: Can anything beat the historical average?', *Review of Financial Studies* **21**(4), 1509–1531.

Campbell, J. Y. and Yogo, M. (2006), 'Efficient tests of stock return predictability', *Journal of Financial Economics* **81**(1), 27–60.

Chen, Q. and Hong, Y. (2010), 'Predictability of equity returns over different time horizons: a nonparametric approach', *Manuscript, Cornell University* .

Cheng, T., Gao, J. and Zhang, X. (2014), Semiparametric localized bandwidth selection in kernel density estimation, Working paper, Monash University.
URL: *https://ideas.repec.org/p/msh/ebswps/2014-27.html*

Diebold, F. X. and Nason, J. A. (1990), 'Nonparametric exchange rate prediction?', *Journal of international Economics* **28**(3-4), 315–332.

Dong, C. and Linton, O. (2016), Additive nonparametric models with time variable and both stationary and nonstationary regressors. Working paper.

Fama, E. F. (1991), 'Efficient capital markets: II', *The Journal of Finance* **46**(5), 1575–1617.

Fama, E. F. and French, K. R. (1988), 'Dividend yields and expected stock returns', *Journal of Financial Economics* **22**(1), 3–25.

Fan, J. and Gijbels, I. (1995), 'Data–driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation', *Journal of the Royal Statistical Society. Series B (Methodological)* **57**(2), 371–394.

Hall, P. and Heyde, C. C. (1980), *Martingale limit theory and its application*, Academic press, New York.

Hansen, B. E. (2008), 'Uniform convergence rates for kernel estimation with dependent data', *Econometric Theory* **24**(03), 726–748.

Härdle, W., Hall, P. and Marron, J. (1988), 'How far are automatically chosen regression smoothing parameters from their optimum ?', *Journal of the American Statistical Association* **83**(401), 86–95.

Härdle, W., Hall, P. and Marron, J. (1989), 'Regression smoothing parameters that are not far from their optimum', *Journal of the American Statistical Association* **83**, 227–233.

Kasparis, I., Andreou, E. and Phillips, P. C. B. (2015), 'Nonparametric predictive regression', *Journal of Econometrics* **185**(2), 468–494.

Keim, D. B. and Stambaugh, R. F. (1986), 'Predicting returns in the stock and bond markets', *Journal of Financial Economics* **17**(2), 357–390.

Lettau, M. and Van Nieuwerburgh, S. (2008), 'Reconciling the return predictability evidence', *Review of Financial Studies* **21**(4), 1607–1652.

Lewellen, J. (2004), 'Predicting returns with financial ratios', *Journal of Financial Economics* **74**(2), 209–235.

Li, Q., Huang, C., Li, D. and Fu, T. (2002), 'Semiparametric smooth coefficient models', *Journal of Business and Economic Statistics* **20**(3), 412–422.

Li, Q. and Racine, J. S. (2007), *Nonparametric Econometrics: Theory and Practice*, Princeton University Press.

Pesaran, M. H. and Timmermann, A. (1995), 'Predictability of stock returns: Robustness and economic significance', *The Journal of Finance* **50**(4), 1201–1228.

Phillips, P. C. B. (2015), 'Pitfalls and possibilities in predictive regression', *Cowles Foundation Discussion Paper* .

Phillips, P. C. B., Li, D. and Gao, J. (2017), 'Estimating smooth structural change in cointegration models', *Journal of Econometrics* **196**(1), 180–195.

Robinson, P. M. (1989), Nonparametric estimation of time–varying parameters, *in* P. Hackl, ed., 'Statistical Analysis and Forecasting of Economis Structural Change', Springer, Berlin, pp. 253–264.

Scholz, M., Nielsen, J. P. and Sperlich, S. (2015), 'Nonparametric prediction of stock returns based on yearly data: The long-term view', *Insurance: Mathematics and Economics* **65**, 143–155.

Stambaugh, R. F. (1999), 'Predictive regressions', *Journal of Financial Economics* **54**(3), 375–421.

Stone, C. J. (1980), 'Optimal rates of convergence for nonparametric estimators', *The Annals of Statistics* **8**(6), 1348–1360.

Vogt, M. (2012), 'Nonparametric regression for locally stationary time series', *The Annals of Statistics* **40**(5), 2601–2633.

Welch, I. and Goyal, A. (2008), 'A comprehensive look at the empirical performance of equity premium prediction', *Review of Financial Studies* **21**(4), 1455–1508.

Xia, Y. and Li, W. (2002), 'Asymptotic behaviour of bandwidth selected by the cross–validation method for local polynomial fitting', *Journal of Multivariate Analysis* **83**(2), 265–287.