UNIVERSITY OF CAMBRIDGE
Faculty of Economics

Cambridge Working Papers in Economics

Cambridge Working Papers in Economics: 1932

# NONPARAMETRIC PREDICTIVE REGRESSIONS FOR STOCK RETURN PREDICTION

Tingting Cheng Jiti Gao Oliver Linton

25 March 2019

We propose two new nonparametric 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. We define estimation methods and establish the large sample properties of these methods in the short horizon and the long horizon case. We apply our methods to stock return prediction using a number of standard predictors such as dividend yield. 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 _nd 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. We also compare our methods with the linear regression and historical mean methods according to an economic metric. In particular, we show how our methods can be used to deliver a trading strategy that beats the buy and hold strategy (and linear regression based alternatives) over our sample period.

# Nonparametric Predictive Regressions for Stock Return Prediction[1]

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

School of Finance, Nankai University[‡],

Department of Econometrics and Business Statistics, Monash University[*]

Faculty of Economics, University of Cambridge[†]

March 25, 2019

**Abstract**

We propose two new nonparametric 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. We define estimation methods and establish the large sample properties of these methods in the short horizon and the long horizon case. We apply our methods to stock return prediction using a number of standard predictors such as dividend yield. 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. We also compare our methods with the linear regression and historical mean methods according to an economic metric. In particular, we show how our methods can be used to deliver a trading strategy that beats the buy and hold strategy (and linear regression based alternatives) over our sample period.

**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 for future stock returns using data upto the early 1980's. Fama (1991) interpreted these findings as evidence

[1]We would like to thank Jens Perch Nielsen for helpful comments. The first author is supported by the National Natural Science Foundation of China (Project No. 71803091) and by the MOE (Ministry of Education in China) Project of Humanities and Social Sciences (Project No. 18YJC790015). The second author would like to thank the Australian Research Council Discovery Grants Program for its support under Grant number: DP170104421.

of time-varying risk premium 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 in the linear predictive regression model if some of the predictors have long memory and the response variable has short memory. This suggests including multiple persistent variables (or their lags) or in the regression so as to allow balancing.

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 for a possible 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 returns over a one–year horizon based on yearly data. Nielsen and Sperlich (2003) also looked at one year predictions into the future based on the nonparametric technique; they worked on the data from the Danish stock market. Scholz, Sperlich and Nielsen (2016) further employed a two-step nonparametric regression to show that bond returns could improve stock prediction. Despite the significant amount of subsequent research, the predictability debate remains unresolved (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 in practice. A further issue that affects the use of nonparametric methods is nonstationarity of the predictor variables, since this slows down the convergence rates in contrast to the linear case where nonstationarity can speed up convergence rates. To mitigate the curse of dimensionality we propose two new predictive models: the multi-step additive predictive regression model (APR) 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 relationship, 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. 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 (2018) studied nonparametric additive models that have deterministic time trend and both stationary (or locally stationary) and integrated variables as components. 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 monthly stock–return prediction 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. 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. To quantify the economic benefits of our methodology we define a trading strategy based on our fitting methods. We show that with appropriate choice of tuning parameter our strategy outperforms the buy and hold method (which corresponds to histroical mean predictor).

The rest of this paper is organized as follows. In Section 2, we describe our models (i.e. NPR and APR) 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 model 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

We describe the NPR and APR models in Sections 2.1 and 2.2, 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 $\tau_t = \frac{t}{n}$ (Robinson (1989) demonstrated that this "scaled time" requirement is necessary for the asymptotic justification of the nonparametric kernel smoothing.), $x_t = (x_t^1, \cdots, x_t^d)^\top$ is a

vector of locally stationary time series, $g_j(\cdot)$ are unknown functions of $\tau_t$ and $x_t$, and $e_{t+j}$ follows a $\alpha$-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 or volatility. A locally stationary process is defined as follows (see Vogt (2012)).

**Definition of 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*

*(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 $\tau, t$ 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 stochastic processes commonly used to model 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 $(\tau, x)$, 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 $\tau$ 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 $\kappa_0=\int K^2(u)du$, $\kappa_2=\int u^2K(u)du$ and

$$R_j(\tau,x)=\frac{\kappa_2}{2}\sum_{i=1}^{d}\left(2\frac{\partial g_j(\tau,x)}{\partial x^i}\frac{\partial f(\tau,x)}{\partial x^i}+\frac{\partial^2 g_j(\tau,x)}{\partial {x^i}^2}f(\tau,x)\right)/f(\tau,x), \tag{6}$$

$$b_j(\tau,x)=\frac{\kappa_2}{2}\left(2\frac{\partial g_j(\tau,x)}{\partial \tau}\frac{\partial f(\tau,x)}{\partial \tau}+\frac{\partial^2 g_j(\tau,x)}{\partial \tau^2}f(\tau,x)\right)/f(\tau,x), \tag{7}$$

Then we have the following theorems; their proofs are given in Appendix A.1.

**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 $\rho$ 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. Then for each given $j$ and $(\tau,x)$, as $n\to\infty$,*

$$\sqrt{nh_j^{d+1}}\left(\widehat{g}_j(\tau,x)-g_j(\tau,x)-B_{j,\tau,x}\right)\to_D N(0,V_{j,\tau,x}), \tag{8}$$

*where $B_{j,\tau,x}=h_j^2(R_j(\tau,x)+b_j(\tau,x))$ and $V_{j,\tau,x}=\kappa_0^{d+1}\sigma_j^2(x)/f(\tau,x)$.*

It can be shown that the bias of $\widehat{g}_j(\tau,x)$ includes a standard component of order $O_P(h_j^2)$ and a nonstandard component of order $O_P(n^{-r}h_j^{-d})$ (see Appendix A.1), however, given the assumption $n^r h_j^{d+2}\to\infty$, the estimation bias resulted from the nonstationarity of regressors (i.e., the part of order $O_P(n^{-r}h_j^{-d})$) is asymptotically negligible. As a result, we find that the asymptotic properties of $\widehat{g}_j(\tau,x)$ 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.

Let $h_j=\rho_j\,h$, $B_J(\tau,x;h)=h^2\sum_{j=1}^{J}\rho_j^2\left(R_j(\tau,x)+b_j(\tau,x)\right)$, $\Sigma_J(x)=\sum_{j=1}^{J}\rho_j^{-(d+1)}\sigma_j^2(x)$ and $V(\tau,x)=\kappa_0^{d+1}/f(\tau,x)$, in which $h>0$ is a bandwidth parameter, $\rho_j$ is a sequence of positive numbers, and $h\to 0$ as $n\to\infty$ and $\rho_j\to\infty$ as $j\to\infty$. 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 $(\tau,x)$. Then as $n\to\infty$,*

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

Theorems 2.1 and 2.2 show that $\widehat{g}_j(\tau, x)$ is a consistent estimator of $g_j(\tau, x)$ and is asymptotically normally distributed. Theorem 2.2 remains valid regardless of whether $J$ is fixed or varying. In the case $J \to \infty$, the rate of $J \to \infty$ is linked implicitly in the conditions of $\frac{nh^{d+1}}{\Sigma_J(x)} \to \infty$ as $(n, J) \to (\infty, \infty)$ and $\lim_{n\to\infty} nh^{d+1}\, \Sigma_J^{-1}(x)\, B_J^2(\tau, x; h) < \infty$. For example, when $\sigma_j^2(x) = 1$ and $\rho_j = j^\gamma$ for $0 < \gamma < \frac{1}{d+1}$, the condition of $\frac{nh^{d+1}}{\Sigma_J(x)} \to \infty$ reduces to requiring $\frac{nh^{d+1}}{J^{1-\gamma(d+1)}} \to \infty$ as $(n, J) \to (\infty, \infty)$. This can be satisfied when $J = \left[n^{\delta_1}\right]$ for a certain choice of $0 < \delta_1 < 1$ such that $n^{1-\delta_1(1-\gamma(d+1))}h^{d+1} \to \infty$ as $n \to \infty$. In the parametric case, the large sample variance reflects the use of overlapping data and standard errors need to be adjusted, Hansen and Hodrick (1980) and Hodrick (1992); in the nonparametric case, we have the "whitening by smoothing" phenomenon, which has been commented on by many authors. Our results show that this continues to hold when the degree of overlap increases with sample size. Consequently, standard error construction is straightforward.

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 and 2.2 are given in Appendix A.1 below.

## 2.2 The APR model

Consider a nonparametric additive predictive regression model of the form

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

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. Without loss of generality and to simplify the notation, we assume that $d = 1$. So model (10) can be simplified as

$$y_{t+j} = \beta_j(\tau_t) + g_j(x_t) + e_{t+j},\ t = 1, 2, \ldots, n,\ j = 1, 2, \ldots, J. \tag{11}$$

In this paper, we use the series estimation method to estimate all the unknown functions in model (11). Naturally, $\beta_j(\cdot)$ and $g_j(\cdot)$ belong to different function spaces as described below.

First, we assume that $\beta_j(\cdot) \in L^2[0, 1] = \{u(\tau) : \int_0^1 u^2(\tau)d\tau < \infty\}$, 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:

$$\beta_j(\tau) = \sum_{s=0}^{\infty} c_{s,j,1}\ \phi_s(\tau),\ \text{where } c_{s,j,1} = \langle \beta_j(\tau), \phi_s(\tau)\rangle. \tag{12}$$

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 $\{p_s(x), s \geq 0\}$,

$$g_j(x) = \sum_{s=0}^{\infty} c_{s,j,2}\ p_s(x),\ \text{where } c_{s,j,2} = \langle g_j(x), p_s(x)\rangle. \tag{13}$$

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 (11) can be written as

$$y_{t+j} = \varphi_{k_{1j}}(\tau_t)^\top c_{1j} + a_{k_{2j}}(x_t)^\top c_{2j} + \gamma_{k_{1j}}(\tau_t) + \gamma_{k_{2j}}(x_t) + e_{t+j},\ \text{for } t = 1, \cdots, n. \tag{14}$$

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

$$B_{nk_j} = \begin{pmatrix} \varphi_{k_{1j}}(\tau_1)^\top & a_{k_{2j}}(x_1)^\top \\ \vdots & \vdots \\ \varphi_{k_{1j}}(1)^\top & a_{k_{2j}}(x_n)^\top \end{pmatrix} \tag{15}$$

be an $n \times k_j$ matrix, where $k_j = k_{1j} + k_{2j}$. Then equation (14) can be written as

$$y_{(j)} = B_{nk_j} c_{(j)} + \gamma_{(j)} + e_{(j)}. \tag{16}$$

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

$$(\widehat{\beta}_j(\tau), \widehat{g}_j(x))^\top = \Phi_j(\tau, x)^\top \widehat{c}_{(j)}, \tag{17}$$

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

$$\Phi_j(\tau, x) = \begin{pmatrix} \varphi_{k_{1j}}(\tau) & \mathbf{0} \\ \mathbf{0} & a_{k_{2j}}(x) \end{pmatrix}. \tag{18}$$

Before establishing asymptotic properties for the estimators, we need some additional notations. 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 $2 \times 2$ block matrix of order $k_j \times k_j$ and $V_{k_j}$ is a $2 \times 2$ symmetric block matrix of the form:

$$U_{k_j} = \begin{pmatrix} U_{11} & U_{12} \\ U_{12}^\top & U_{22} \end{pmatrix} \quad \text{and} \quad V_{k_j} = \begin{pmatrix} V_{11} & V_{12} \\ V_{12}^\top & V_{22} \end{pmatrix}. \tag{19}$$

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.2.

**Theorem 2.3** *Suppose that uniformly over $n$, all the eigenvalues of $U_{k_j}$ and $V_{k_j}$ are positive, and that Assumptions A.2.1-A.2.6 hold. Then, for any $\tau \in [0, 1]$ and $x \in V$, as $n \to \infty$, we have*

$$\Delta_{nj}^{-1} \begin{pmatrix} \sqrt{n}[\widehat{\beta}_j(\tau) - \beta_j(\tau)] \\ \sqrt{n}[\widehat{g}_j(x) - g_j(x)] \end{pmatrix} \to_D N(\mathbf{0}, I_2), \tag{20}$$

*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 2.4** *Let Assumptions A.2.1–A.2.6 hold. Then as* $n \to \infty$,

$$\sqrt{n}\,\Gamma_{nJ}^{-1/2}\left(\widehat{m}(\tau,x) - m(\tau,x)\right) \to_D N(0,1), \tag{21}$$

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

**Remark.** (i) Note that Theorems 2.3 and 2.4 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, $m(\tau,x)$ can also be consistently estimated. (ii) Note also that Theorem 2.4 remains valid when $J \to \infty$. In the case $J \to \infty$, the rate of $J \to \infty$ is linked implicitly in the condition of $\frac{n}{\Gamma_{nJ}} \to \infty$ as $(n,J) \to (\infty,\infty)$.

Section 3 below discusses about how to choose the truncation parameters $k_j$. The proofs of Theorems 2.3 and 2.4 are given in Appendix A.2 below.

# 3 Implementation

In this section, we will discuss computational details on the implementation of the NPR and APR models, particularly the bandwidth selection for the NPR model 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. 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 $\rho_j$ 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. (2018). Let us introduce the following notation:

$$D_j(h_j) = \frac{1}{n}\sum_{t=1}^{n}\left(\widehat{g}_j(\tau_t,x_t) - g_j(\tau_t,x_t)\right)^2 w(\tau_t,x_t), \tag{22}$$

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 $n \to \infty$

$$n^{\frac{3}{10}} \left( \frac{\widehat{h}_j}{h_{j0}} - 1 \right) \to_D N(0, \sigma_{j0}^2) \tag{23}$$

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 $\theta^j$, in which $h_0 > 0$, $\beta > 0$ and $\theta > 1$ will all be estimated in the rest of this section.

Using equation (23), we have for large enough $n$

$$\log \left( \frac{\widehat{h}_j}{h_{j0}} \right) = \log \left( 1 + \frac{\widehat{h}_j}{h_{j0}} - 1 \right) \approx \frac{\widehat{h}_j}{h_{j0}} - 1 \equiv n^{-\frac{3}{10}} \, \varepsilon_j, \tag{24}$$

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

$$\begin{aligned} \log \left( \widehat{h}_j \right) &= \log \left( h_{j0} \right) + \eta_j = \log(h_0) + \log(\rho_j) + \eta_j \\ &= \begin{cases} \log(h_0) + \beta \, \log(j) + \eta_j, & \text{if } \rho_j = j^\beta, \\ \log(h_0) + j \, \log(\theta) + \eta_j, & \text{if } \rho_j = \theta^j, \end{cases} \end{aligned} \tag{25}$$

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 $\beta$ by an ordinary least squares (OLS) estimator of the form

$$\widehat{\beta} = \left( \sum_{j=1}^{J} \left( \log(j) - \overline{\log(J)} \right)^2 \right)^{-1} \sum_{j=1}^{J} \left( \log(j) - \overline{\log(J)} \right) \left( Z_j - \overline{Z} \right), \tag{26}$$

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 (25) and (26) imply that the following rate of convergence:

$$\widehat{\beta} - \beta = O_P \left( \left( \sqrt{J \log(J)} \right)^{-1} \cdot n^{-\frac{3}{10}} \right). \tag{27}$$

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

$$\widehat{\gamma} = \left( \sum_{j=1}^{J} \left( j - \overline{J} \right)^2 \right)^{-1} \sum_{j=1}^{J} \left( j - \overline{J} \right) \left( Z_j - \overline{Z} \right), \tag{28}$$

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

Meanwhile, equations (25) and (28) imply a rate of convergence of the form:

$$\widehat{\gamma} - \gamma = O_P \left( J^{-\frac{3}{2}} \cdot n^{-\frac{3}{10}} \right). \tag{29}$$

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 (27) and (29) imply that the OLS estimators may have fast convergence 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.

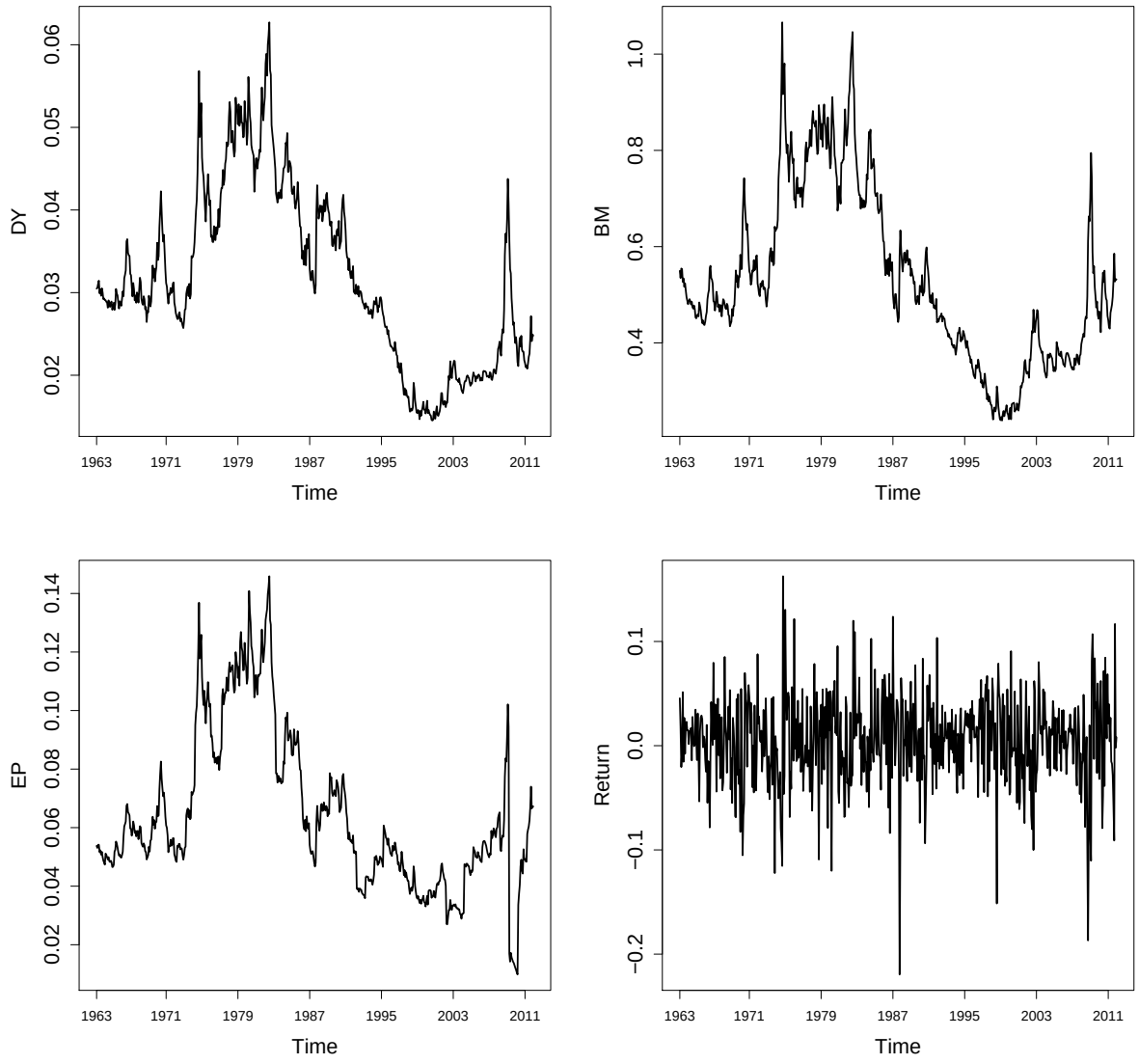
- We divide the sample into two sets, the initialization set with sample size $n_1$ and validation set with sample size $n - n_1$.
- The initialization set is used to estimate the model for a given value of $(k_{1j}, k_{2j})$, then the estimated model is used to forecast the response variable in the 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 and APR models

In this section, we implement the NPR and APR 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. We also compute the correlation coefficient between the three predictors and find that they are highly correlated. In particular, the correlation is 0.9493 between $x_t^1$ and $x_t^2$, 0.8921 between $x_t^1$ and $x_t^3$, and 0.9027 between $x_t^2$ and $x_t^3$.

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




In the following, we will examine the performance of the NPR and APR 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 four models:

- Mean: $y_{t+j} = \mu + e_{t+j}$;
- 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: $y_{t+j} = g_j^0(\tau_t) + \sum_{i=1}^3 g_j^i(x_t^i) + e_{t+j}$.

Note that we use kernel method to estimate the unknown function $g_j(\cdot)$ in the NPR model. 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

$$R^2_{IS,j} = 1 - \frac{\sum_{t=1}^{n}(y_{t+j} - \widehat{y}_{t+j})^2}{\sum_{t=1}^{n}(y_{t+j} - \overline{y}_j)^2}, \tag{30}$$

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, $R^2_{IS,j}$ takes value of zero for all given values of $j$. From (30), it is easy to see that $R^2_{IS,j}$ can be written as

$$R^2_{IS,j} = 1 - \frac{\text{MSE}_A}{\text{MSE}_M}, \tag{31}$$

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, $R^2_{IS,j}$ can also indicate the relative ratio of the mean squared errors between the historical mean model and the other models. If $R^2_{IS,j}$ for a certain model is positive, then this model performs better than the historical mean model. Simply speaking, the larger the $R^2_{IS,j}$ is, the better the corresponding model performs.

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

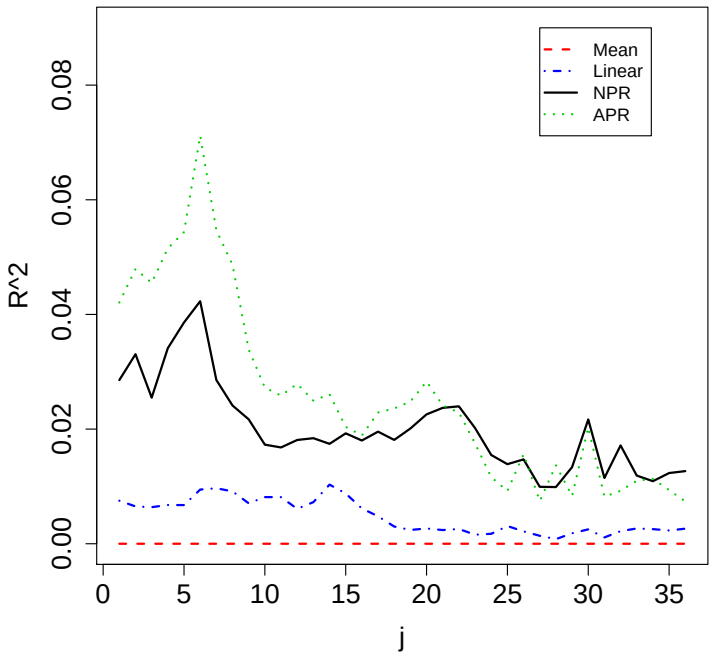
- The NPR and APR models have larger $R^2_{IS,j}$ than the traditional historical mean model and linear model, for $j = 1, 2, \cdots, 36$. This means that the NPR and APR models have better in–sample performance than the traditional mean and linear model.

- 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.

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 |

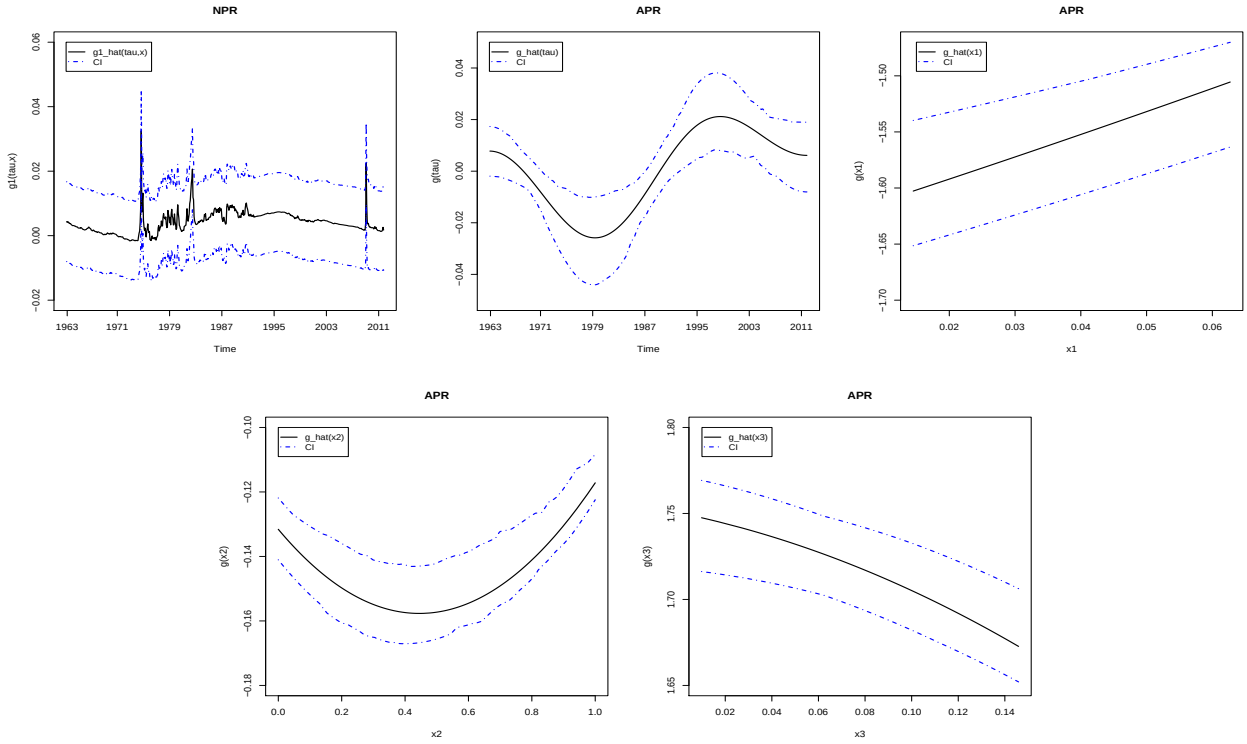
Figure 2: Plot of $R^2_{IS,j}$ with $j = 1, 2, \cdots, 36$ for all the models.




From the results in Table 1 and Figure 2, we observe that the NPR and APR 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, 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 and APR 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.

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




It is pointed out that the literature on parametric model specifications, such as Hansen and Hodrick (1980) and Hodrick (1992), discusses the standard errors and then about how to correct standard errors for overlappingness. In our nonparametric framework, we are mostly looking at out–of–sample evaluation as discussed in the following section.

## 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 and APR 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.

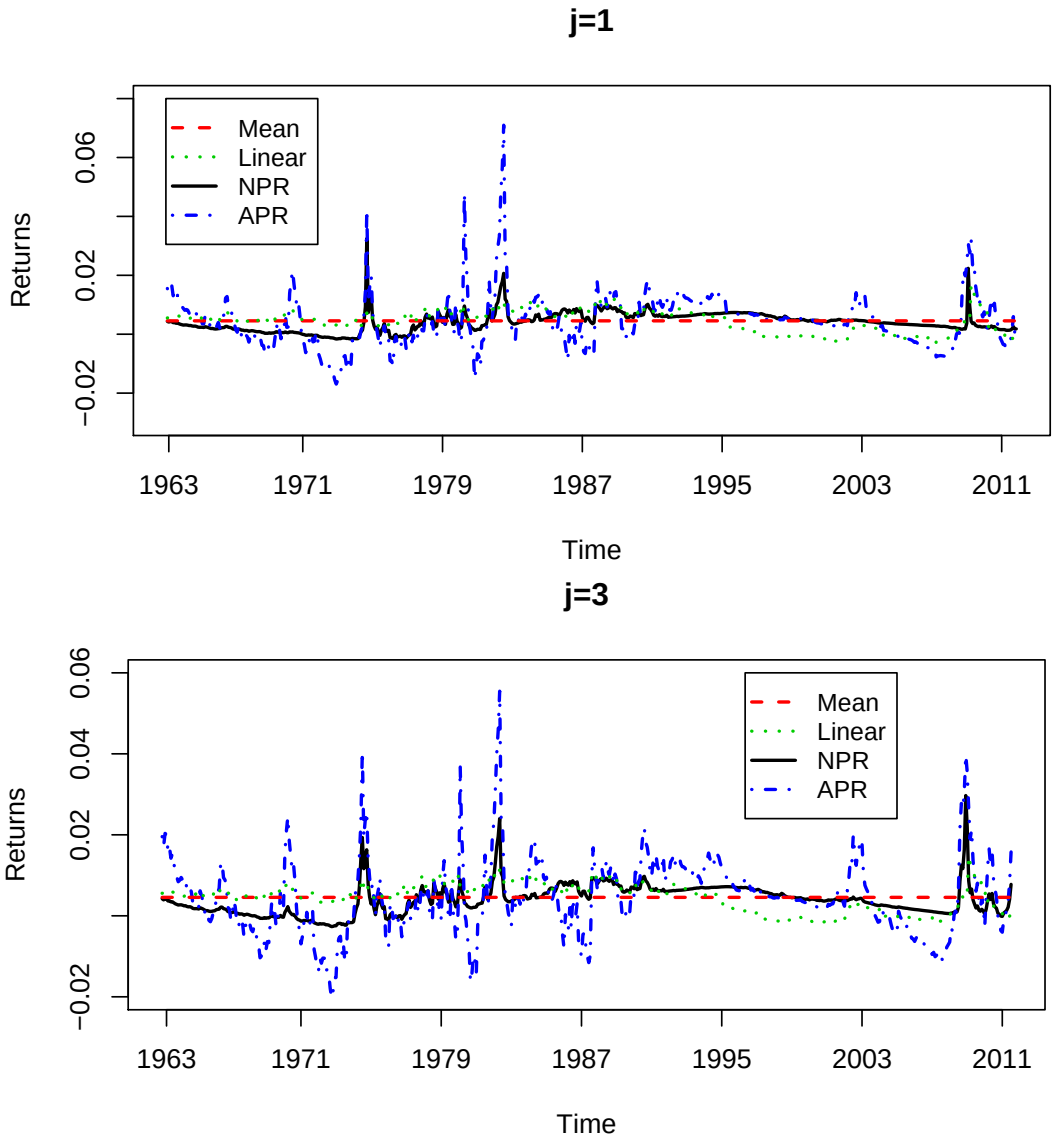
- For the first window, we conduct the multi-step prediction based on $n-1$ observations. At the point $x_n$, we predict $y_{n+1}$ using these $n-1$ pairs of observations $\{(x_1, y_2), (x_2, y_3), \cdots, (x_{n-1}, y_n)\}$. The estimated value of $y_{n+1}$ is denoted as $\widehat{y}_{n+1}$. Then we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

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

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




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 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 $R$th 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

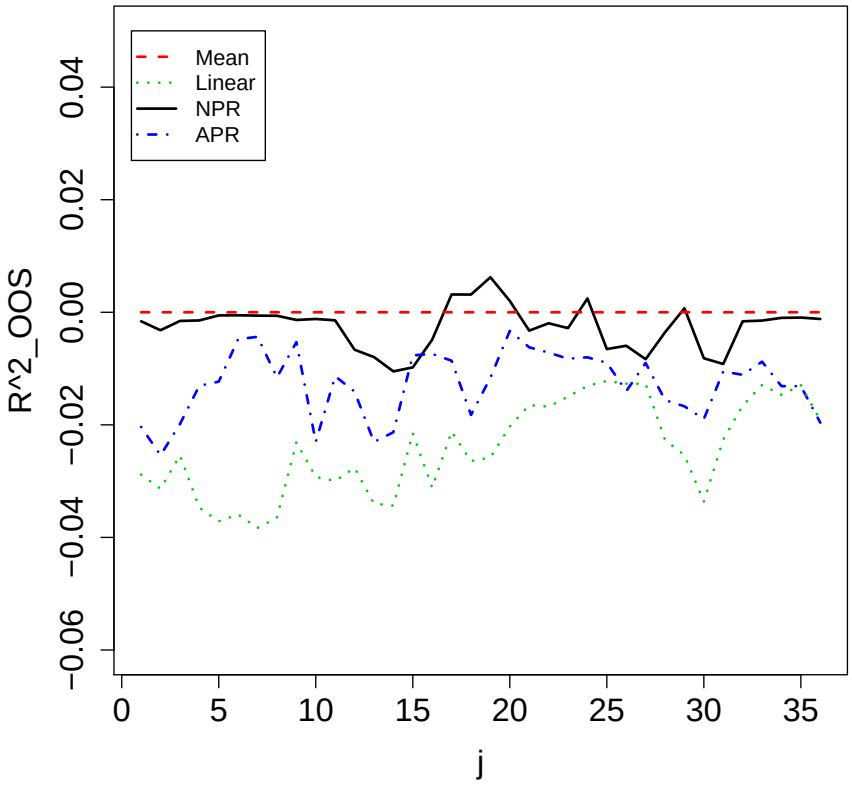
$$R^2_{OOS,j,n,R} = 1 - \frac{\sum_{r=1}^{R}(y_{n+r,j} - \widehat{y}_{n+r,j})^2}{\sum_{r=1}^{R}(y_{n+r,j} - \overline{y}_{n+r,j})^2},$$

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 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 |

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




Apart from looking at behaviour 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 three 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 three 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).

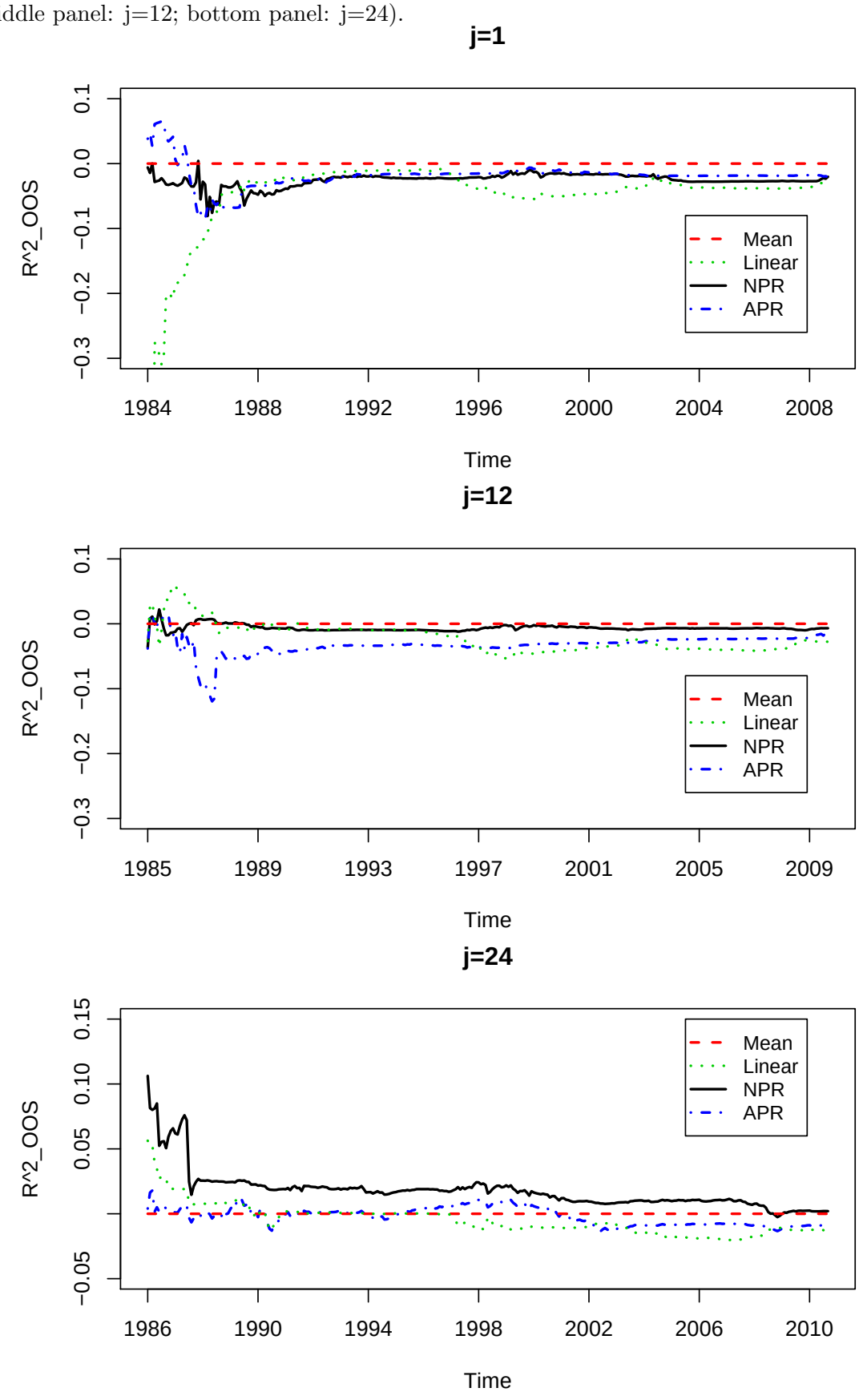
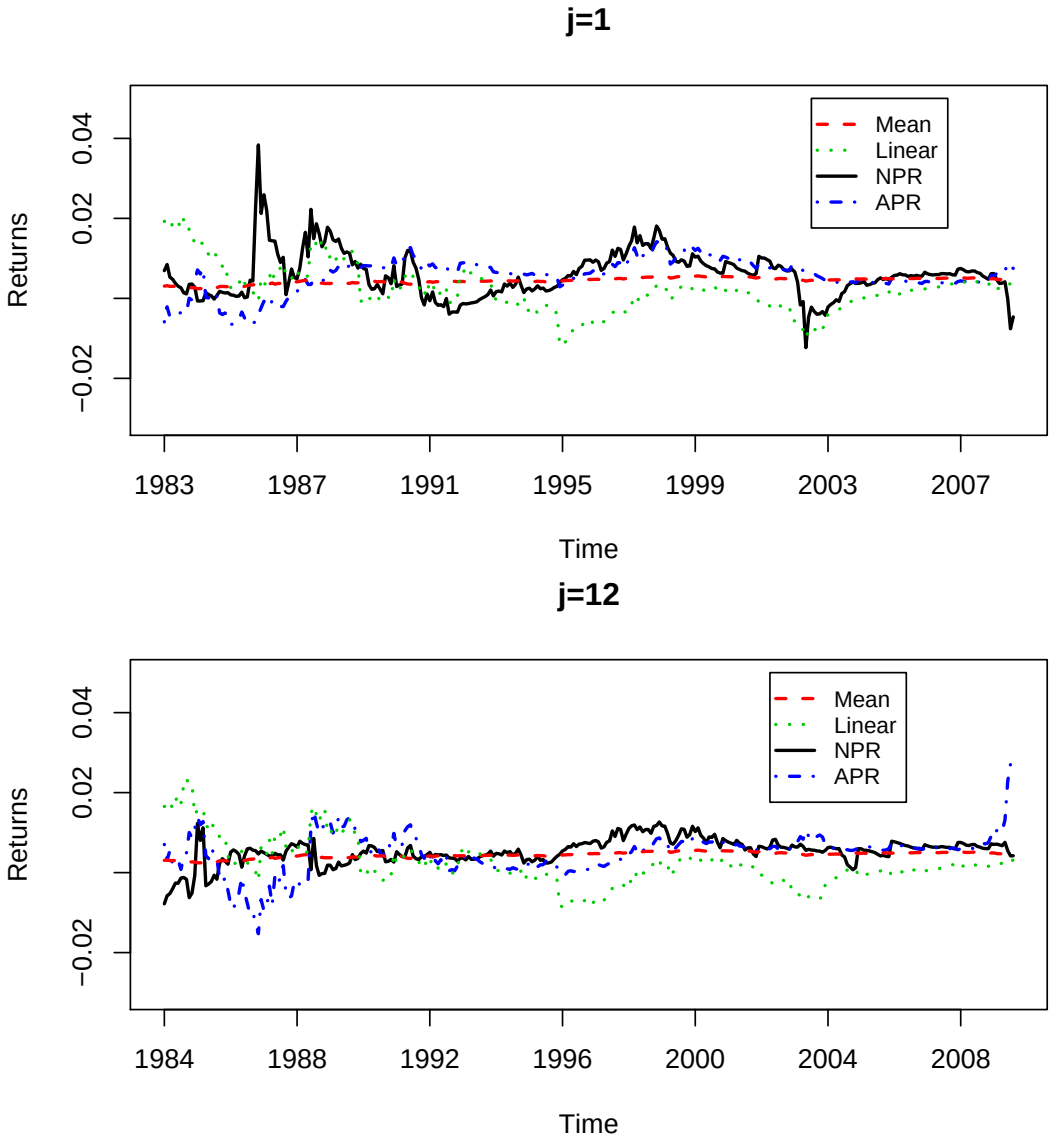


Figure 7: Plots of out-of-sample predicted returns for all the models (top panel: j=1; bottom panel: j=12).



### 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 |

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 |

## 4.3 Trading strategy

It is hard to beat the historical mean model according to out of sample prediction measured by squared error loss. We now turn to an economic metric for comparing our methods with the historical mean model. 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 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. So our trading strategy earns in one period is $w_t r_{t+1} + (1 - w_t) r_0$, where $r_0$ is the bond rate and $r_{t+1}$ is the outturn on the stock market next period and our weights are $w_t = I(\widehat{r}_{t+1} > c)$, in which $c$ is a selected threshold. In this study, we consider six different thresholds $(0.001, 0.002, \cdots, 0.006$, which are corresponding to quantiles between 30%-50% of historical distribution of returns) to examine the performance of our trading strategy with the buy and hold strategy in terms of profit. We compute the profit of both strategies based on the NASDAQ Composite Index. For example, In May 1983, NASDAQ index was at its closing price of 308.73 and until December 2011, the closing price was then 2605.15. Assume that the initial investment is 1 unit, then using a buy-and-hold strategy will result in a return of 7.4383 (7.4383= 2605.15/308.73-1).

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 assume that the cost such as transaction fee during the trading could be ignored.

Tables 5–6 show the results of stock return predictions that with NPR and APR models respectively. For comparison, we also present the corresponding results using the linear model in Section 4 in Table 7. To see whether our trading strategy involves lots of buying and selling, we present the number of transactions that would be required in each case in Table 8. 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; and for the APR model, the thresholds are 0.001 and 0.002. Moreover, using the same trading strategy, our proposed models can make more profit than the linear model in almost all cases. As a result, we see that our proposed trading strategies with the use of NPR and APR could be better alternatives 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 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 |

Table 7: Profit of trading strategy with the use of linear 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 | 2.7176 | 3.8320 | 0.7033 | 0.5771 | 0.3669 | 0.2441 | 7.4383 |
| 1993 May | 1.6131 | 2.8377 | 0.5896 | 0.7162 | 0.9250 | 0.8333 | 2.7188 |
| 2003 May | 0.2214 | 0.0959 | 0.1022 | 0.4053 | 0.5763 | 0.5012 | 0.6324 |

Table 8: Number of transactions required using our trading strategy.

| Starting date | Our trading strategy with different threshold | | | | | | Duration |
|---|---|---|---|---|---|---|---|
| | 0.001 | 0.002 | 0.003 | 0.004 | 0.005 | 0.006 | |
| NPR | | | | | | | |
| 1983 May | 0 | 0 | 31 | 70 | 199 | 297 | 343 |
| 1993 May | 0 | 0 | 0 | 4 | 79 | 177 | 223 |
| 2003 May | 0 | 0 | 0 | 4 | 39 | 103 | 103 |
| APR | | | | | | | |
| 1983 May | 61 | 73 | 93 | 117 | 131 | 166 | 343 |
| 1993 May | 13 | 13 | 15 | 45 | 88 | 114 | 223 |
| 2003 May | 13 | 13 | 14 | 42 | 80 | 90 | 103 |
| Linear | | | | | | | |
| 1983 May | 148 | 178 | 222 | 250 | 267 | 281 | 343 |
| 1993 May | 126 | 151 | 187 | 204 | 213 | 218 | 223 |
| 2003 May | 41 | 51 | 69 | 84 | 93 | 98 | 103 |

In this trading strategy, we could also use the historical t-bill rate instead of the risk free bond rate 0.02/12. The results are presented in Tables 9-11. It is easy to see that the results are similar to those obtained by using the risk free bond rate.

Table 9: 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 | 9.3190 | 11.5301 | 3.0527 | 4.9841 | 7.4383 |
| 1993 May | 2.9617 | 2.9617 | 2.9617 | 2.6397 | 1.1664 | 2.1989 | 2.7188 |
| 2003 May | 0.7895 | 0.7895 | 0.7895 | 0.6440 | 0.3708 | 0.1653 | 0.6324 |

Table 10: 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 | 12.7081 | 15.1186 | 12.2220 | 15.1435 | 13.7191 | 9.6557 | 7.4383 |
| 1993 May | 2.7932 | 2.7932 | 2.4863 | 0.8163 | 0.2515 | 0.4402 | 2.7188 |
| 2003 May | 0.7134 | 0.7134 | 0.6187 | -0.0757 | -0.2546 | -0.1215 | 0.6324 |

Table 11: Profit of trading strategy with the use of linear 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 | 3.3436 | 4.9845 | 1.3156 | 1.2402 | 0.9961 | 0.8564 | 7.4383 |
| 1993 May | 1.8263 | 3.3252 | 0.9093 | 1.0871 | 1.3594 | 1.2289 | 2.7188 |
| 2003 May | 0.1701 | 0.0520 | 0.0824 | 0.3909 | 0.5724 | 0.4855 | 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. 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 have found 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 have found 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 compared our methods with the linear regression and historical mean methods according to an economic metric. In particular, we showed how our methods can be used to deliver a trading strategy that beats the buy and hold strategy (and linear regression based alternatives) over our sample period.

# Appendix

In this appendix, we provide the proofs of Theorem 2.1–Theorem 2.4. Sections A.1 and A.2 below provide the necessary assumptions and the proofs of the main results for the estimators in the NPR and APR models, respectively.

## A.1. The NPR model

First, we present some assumptions for the establishment of asymptotic properties for $\widehat{g}_j(\tau,x)$ and $g(\tau,x)$ 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 $\alpha$–mixing with mixing coefficient $\alpha$ 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. (ii) The densities $f(\tau,x):=f_{x_t(\tau)}(x)$ of the variables $x_t(\tau)$ are smooth in $\tau$ for each time point $\tau\in[0,1]$. In particular, $f(\tau,x)$ is differentiable with respect to $\tau$ for each $x\in\mathbb{R}^d$, and the derivative $\partial f(\tau,x)/\partial\tau$ is continuous. (iii) $f(\tau,x)$ is partially differentiable with respect to $x$ for each $\tau\in[0,1]$. The derivatives $\partial f(\tau,x)/\partial x^i$ 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 $\rho_j$ is a positive constant and $\rho_j\to\infty$ as $j\to\infty$ ; $h\to 0$ as $n\to\infty$. In addition, $nh_j^{d+1}\to\infty$ 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 $\tau_t$ is in a small neighborhood of $\tau$. Assumption A.1.2 imposes smoothness condition on the unknown functions and the density of $x_t(\tau_t)$. Assumption A.1.3 is required to guarantee a certain rate of the convergence rate, which is also used in Vogt (2012). Assumption A.1.4 is a standard assumption for kernel function $K(\cdot)$ and bandwidth $h_j$.

**Proof of Theorem 2.1.**

Observe that

$$\widehat{g}_j(\tau,x)-g_j(\tau,x)=\frac{1}{\widehat{f}(\tau,x)}\left(\widehat{g}_j^E(\tau,x)+\widehat{g}_j^B(\tau,x)-g_j(\tau,x)\widehat{f}(\tau,x)\right), \tag{32}$$

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

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

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 of Theorem 4.3 in Vogt (2012).

- We first show that $B_j(\tau, x) = \sqrt{nh_j^{d+1}} f(\tau, x) \left(B_{j,\tau,x} + o_P(h_j^2)\right)$, where $B_{j,\tau,x} = h_j^2(R_j(\tau, x) + b_j(\tau, x))$.
- We establish the asymptotic normality $V_j(\tau, x) \to_D N(0, \kappa_0^{d+1}\sigma_j^2(x)f(\tau, x))$, where $\kappa_0 = \int K^2(u)du$.
- We then show that $\widehat{f}(\tau, x) - f(\tau, x) = o_P(1)$ and $\widehat{f}(\tau, x)^{-1} = O_P(1)$.

Following the spirit of Vogt (2012) that approximate the locally stationary time series $x_t$ by its stationary counterpart $x_t(\tau_t)$, we write

$$\mathbb{E}[\widehat{g}_j^B(\tau, x) - g_j(\tau, x)\widehat{f}(\tau, x)] = Q_1(\tau, x) + \cdots + Q_4(\tau, x),$$

where $Q_i(\tau, x) = \frac{1}{nh_j^{d+1}} \sum_{t=1}^{n} K_h(\tau - \tau_t) q_i(\tau, x)$, $K_h(x) = K(x/h)$ and

$$\begin{aligned}
q_1(\tau, x) =& \mathbb{E}\left[\prod_{i=1}^{d} \bar{K}_h(x^i - x_t^i)\left\{\prod_{i=1}^{d} K_h(x^i - x_t^i) - \prod_{i=1}^{d} K_h(x^i - x_t^i(\tau_t))\right\} \times \left\{g_j(\tau_t, x_t) - g_j(\tau, x)\right\}\right] \\
q_2(\tau, x) =& \mathbb{E}\left[\prod_{i=1}^{d} \bar{K}_h(x^i - x_t^i) \prod_{i=1}^{d} K_h(x^i - x_t^i(\tau_t)) \times \left\{g_j(\tau_t, x_t) - g_j(\tau_t, x_t(\tau_t))\right\}\right], \\
q_3(\tau, x) =& \mathbb{E}\left[\left\{\prod_{i=1}^{d} \bar{K}_h(x^i - x_t^i) - \prod_{i=1}^{d} \bar{K}_h(x^i - x_t^i(\tau_t))\right\}\right. \\
& \left. \times \prod_{i=1}^{d} K_h(x^i - x_t^i(\tau_t))\left\{g_j(\tau_t, x_t(\tau_t)) - g_j(\tau, x)\right\}\right], \\
q_4(\tau, x) =& \mathbb{E}\left[\prod_{i=1}^{d} K_h(x^i - x_t^i(\tau_t))\left\{g_j(\tau_t, x_t(\tau_t)) - g_j(\tau, x)\right\}\right],
\end{aligned}$$

in which $\bar{K}$ is a Lipschitz continuous function with support $[-qC_1, qC_1]$ for some $q > 1$ and $\bar{K}(x) = 1$ for all $x \in [-C_1, C_1]$ and write $\bar{K}_h(x) = \bar{K}(x/h)$.

As the kernel function is bounded, we have

$$\left|\prod_{i=1}^{d} K_h\big(x^i - x_t^i\big) - \prod_{i=1}^{d} K_h\big(x^i - x_t^i(\tau_t)\big)\right| \leq C \sum_{k=1}^{d} \left|K_h\big(x^k - x_t^k\big) - K_h\big(x^k - x_t^k(\tau_t)\big)\right|^r,$$

where $C$ is a finite constant and $r = \min\{\rho, 1\}$. Then we have

$$\begin{aligned}|Q_1(\tau,x)| \leq & \frac{C}{nh_j^{d+1}} \sum_{t=1}^{n} K_h(\tau - \tau_t) \times \mathbb{E}\left[\sum_{k=1}^{d} \left|K_h\big(x^k - x_t^k\big) - K_h\big(x^k - x_t^k(\tau_t)\big)\right|^r \right. \\ & \left. \times \prod_{i=1}^{d} \bar{K}_h\big(x^i - x_t^i\big) \left|g_j\big(\tau_t, x_t\big) - g_j(\tau, x)\right|\right].\end{aligned}$$

Under Assumptions A.1.1(i) and A.1.4(i), we further have

$$\begin{aligned}|Q_1(\tau,x)| &\leq \frac{C}{nh_j^d} \sum_{t=1}^{n} K_h(\tau - \tau_t) \mathbb{E}\left[\sum_{k=1}^{d} \left|K_h\big(x^k - x_t^k\big) - K_h\big(x^k - x_t^k(\tau_t)\big)\right|^r\right] \\ &\leq \frac{C}{nh_j^d} \sum_{t=1}^{n} K_h(\tau - \tau_t) \mathbb{E}\left[\sum_{k=1}^{d} \left|\frac{1}{nh_j} U_{nt}(\tau_t)\right|^r\right] \leq \frac{C}{n^r h_j^{d-1+r}}.\end{aligned}$$

Similarly, we can show that $|Q_2(\tau,x)| \leq \frac{C}{n^r h_j^d}$ and $|Q_3(\tau,x)| \leq \frac{C}{n^r h_j^{d-1+r}}$. These results are uniformly in $\tau$ and $x$.

Define

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

Then we can write $Q_4(\tau,x) = \mathbb{E}\left[\widehat{g}_j^{B*}(\tau,x) - g_j(\tau,x)\widehat{f}^*(\tau,x)\right]$.

Under Assumptions A.1.1(i), A1.2(ii)(iii), A.1.4(i) and by change of variables, Taylor expansion, we can show that

$$\begin{aligned}(33)\qquad \mathbb{E}\widehat{f}^*(\tau,x) &= \frac{1}{nh_j^{d+1}} \sum_{t=1}^{n} K\left(\frac{\tau_t - \tau}{h_j}\right) \mathbb{E}\left[L\left(\frac{x_t(\tau_t) - x}{h_j}\right)\right] \\ &= \frac{1}{nh_j^{d+1}} \sum_{t=1}^{n} K\left(\frac{\tau_t - \tau}{h_j}\right) \left[\int L\left(\frac{x_y - x}{h_j}\right) f(\tau_t, x_y)\, dx_y\right] \\ &= \frac{1 + o_P(1)}{h_j^{d+1}} \iint K\left(\frac{\tau_y - \tau}{h_j}\right) L\left(\frac{x_y - x}{h_j}\right) f(\tau_y, x_y)\, d\tau_y\, dx_y \\ &= \frac{1 + o_P(1)}{h_j^{d+1}} \int \cdots \int K\left(\frac{\tau_y - \tau}{h_j}\right) \prod_{i=1}^{d} K\left(\frac{x_y^i - x}{h_j}\right) f(\tau_y, x_y^1, \cdots, x_y^d)\, d\tau_y\, dx_y^1 \cdots dx_y^d \\ &= \int \cdots \int K(p) \prod_{i=1}^{d} K(q_i) f(\tau + ph_j, x^1 + q_1 h_j, \cdots, x^d + q_d h_j)\, dp\, dq^1 \cdots dq^d (1 + o_P(1)) \\ &= \int \cdots \int K(p) \prod_{i=1}^{d} K(q_i) f(\tau, x)\, dp\, dq^1 \cdots dq^d\end{aligned}$$

$$+\int\cdots\int K(p)\prod_{i=1}^{d}K(q_i)\left(ph_j\frac{\partial f(\tau,x)}{\partial\tau}+\sum_{i=1}^{d}q_ih_j\frac{\partial f(\tau,x)}{\partial x^i}\right)\,dp\,dq^1\cdots dq^d$$
$$+\int\cdots\int K(p)\prod_{i=1}^{d}K(q_i)\frac{1}{2}\left(p^2h_j^2\frac{\partial^2 f(\tau,x)}{\partial\tau^2}+\sum_{i=1}^{d}q_i^2h_j^2\frac{\partial^2 f(\tau,x)}{\partial {x^i}^2}\right)\,dp\,dq^1\cdots dq^d$$
$$+\int\cdots\int K(p)\prod_{i=1}^{d}K(q_i)\frac{1}{2}\left(\sum_{i=1}^{d}pq_ih_j^2\frac{\partial^2 f(\tau,x)}{\partial\tau\partial x^i}+2\sum_{i=2}^{d}\sum_{s=1}^{i-1}q_iq_sh_j^2\frac{\partial^2 f(\tau,x)}{\partial x^i\partial x^s}\right)\,dp\,dq^1\cdots dq^d$$
$$+o_P(h_j^2)=f(\tau,x)+\frac{\kappa_2}{2}h_j^2\left(\frac{\partial^2 f(\tau,x)}{\partial\tau^2}+\sum_{i=1}^{d}\frac{\partial^2 f(\tau,x)}{\partial {x^i}^2}\right)+o_P(h_j^2).$$

Similarly, we can show that

$$\begin{aligned}
\mathbb{E}[\widehat{g}_j^{B*}(\tau,x)] &= \frac{1}{nh_j^{d+1}}\sum_{t=1}^{n}K\left(\frac{\tau_t-\tau}{h_j}\right)\mathbb{E}\left[L\left(\frac{x_t(\tau_t)-x}{h_j}\right)g_j(\tau_t,x_t(\tau_t))\right]\\
&= \frac{1}{nh_j^{d+1}}\sum_{t=1}^{n}K\left(\frac{\tau_t-\tau}{h_j}\right)\left[\int L\left(\frac{x_y-x}{h_j}\right)g_j(\tau_t,x_y)f(\tau_t,x_y)\,dx_y\right]\\
&= \frac{1+o_P(1)}{h_j^{d+1}}\iint K\left(\frac{\tau_y-\tau}{h_j}\right)L\left(\frac{x_y-x}{h_j}\right)g_j(\tau_y,x_y)f(\tau_y,x_y)\,d\tau_y\,dx_y\\
&= \frac{1+o_P(1)}{h_j^{d+1}}\int\cdots\int K\left(\frac{\tau_y-\tau}{h_j}\right)\prod_{i=1}^{d}K\left(\frac{x_y^i-x}{h_j}\right)g_j(\tau_y,x_y^1,\cdots,x_y^d)f(\tau_y,x_y^1,\cdots,x_y^d)\,d\tau_y\,dx_y^1\cdots dx_y^d.
\end{aligned}$$

Taking the second-order Taylor expansion for $g_j(\tau_y,x_y^1,\cdots,x_y^d)$ and $f(\tau_y,x_y^1,\cdots,x_y^d)$ and keeping the terms up to $O_P(h_j^2)$, we obtain that

$$\begin{aligned}
\mathbb{E}\widehat{g}_j^{B*}(\tau,x) = g_j(\tau,x)f(\tau,x)&+\frac{\kappa_2}{2}h_j^2\left(2\frac{\partial g_j(\tau,x)}{\partial\tau}\frac{\partial f(\tau,x)}{\partial\tau}+\frac{\partial^2 g_j(\tau,x)}{\partial\tau^2}f(\tau,x)\right)\\
&+\frac{\kappa_2}{2}h_j^2\sum_{i=1}^{d}\left(2\frac{\partial g_j(\tau,x)}{\partial x^i}\frac{\partial f(\tau,x)}{\partial x^i}+\frac{\partial^2 g_j(\tau,x)}{\partial {x^i}^2}f(\tau,x)\right)\\
&+\frac{\kappa_2}{2}h_j^2\left(\frac{\partial^2 f(\tau,x)}{\partial\tau^2}g_j(\tau,x)+\sum_{i=1}^{d}\frac{\partial^2 f(\tau,x)}{\partial {x^i}^2}g_j(\tau,x)\right)+o_P(h_j^2).
\end{aligned}$$

Then we have

$$\begin{aligned}
Q_4(\tau,x) &= \mathbb{E}[\widehat{g}_j^{B*}(\tau,x)]-g_j(\tau,x)\mathbb{E}[\widehat{f}^*(\tau,x)]\\
&= \frac{\kappa_2}{2}h_j^2\sum_{i=1}^{d}\left(2\frac{\partial g_j(\tau,x)}{\partial x^i}\frac{\partial f(\tau,x)}{\partial x^i}+\frac{\partial^2 g_j(\tau,x)}{\partial {x^i}^2}f(\tau,x)\right)\\
&\quad+\frac{\kappa_2}{2}h_j^2\left(2\frac{\partial g_j(\tau,x)}{\partial\tau}\frac{\partial f(\tau,x)}{\partial\tau}+\frac{\partial^2 g_j(\tau,x)}{\partial\tau^2}f(\tau,x)+o_P(h_j^2)\right)\\
&= f(\tau,x)\left(h_j^2R_j(\tau,x)+h_j^2b_j(\tau,x)+o_P(h_j^2)\right)\\
&= f(\tau,x)\left(B_{j,\tau,x}+o_P(h_j^2)\right), \qquad (34)
\end{aligned}$$

where $B_{j,\tau,x}=h_j^2R_j(\tau,x)+h_j^2b_j(\tau,x)$ and

$$R_j(\tau,x)=\frac{\kappa_2}{2}\sum_{i=1}^{d}\left(2\frac{\partial g_j(\tau,x)}{\partial x^i}\frac{\partial f(\tau,x)}{\partial x^i}+\frac{\partial^2 g_j(\tau,x)}{\partial {x^i}^2}f(\tau,x)\right)/f(\tau,x),$$

$$b_j(\tau,x)=\frac{\kappa_2}{2}\left(2\frac{\partial g_j(\tau,x)}{\partial\tau}\frac{\partial f(\tau,x)}{\partial\tau}+\frac{\partial^2 g_j(\tau,x)}{\partial\tau^2}f(\tau,x)\right)/f(\tau,x).$$

Define $B_j^*(\tau,x)=\frac{1}{\sqrt{nh_j^{d+1}}}B_j(\tau,x)$. Then we have

$$\mathbb{E}[B_j^*(\tau,x)]=Q_1(\tau,x)+\cdots+Q_4(\tau,x).$$

Combining the results of $Q_1(\tau,x),\cdots,Q_4(\tau,x)$, we have that

$$\mathbb{E}[B_j^*(\tau,x)]=f(\tau,x)B_{j,\tau,x}+o_P(h_j^2)+O_P\left(\frac{1}{n^r h_j^d}\right). \tag{35}$$

As we assume that $n^r h_j^{d+2}\to\infty$, we have

$$\mathbb{E}[B_j^*(\tau,x)]-f(\tau,x)B_{j,\tau,x}=o_P(h_j^2)=o_P(1).$$

Similar to the derivation of equation (35), we can show that

$$\mathrm{Var}[B_j^*(\tau,x)]=\mathbb{E}[B_j^{*2}(\tau,x)]-\left(\mathbb{E}[B_j^*(\tau,x)]\right)^2=o_P(1). \tag{36}$$

Hence, we have $B_j^*(\tau,x)-\mathbb{E}[B_j^*(\tau,x)]=o_P(1)$.

Therefore, we have

$$B_j^*(\tau,x)-f(\tau,x)B_{j,\tau,x}=B_j^*(\tau,x)-\mathbb{E}[B_j^*(\tau,x)]+\mathbb{E}[B_j^*(\tau,x)]-f(\tau,x)B_{j,\tau,x}=o_P(1),$$

which is equivalent as

$$\frac{B_j(\tau,x)}{\sqrt{nh_j^{d+1}}}f(\tau,x)^{-1}-B_{j,\tau,x}=o_P(1). \tag{37}$$

On the other hand, we have that

$$V_j(\tau,x)=\sqrt{nh_j^{d+1}}\widehat{g}_j^E(\tau,x)=\frac{1}{\sqrt{nh_j^{d+1}}}\sum_{t=1}^n K\left(\frac{\tau_t-\tau}{h_j}\right)L\left(\frac{x_t-x}{h_j}\right)e_{t+j}.$$

It is obvious that $\mathbb{E}[V_j(\tau,x)]=0$ and

$$\begin{aligned}V_j^2(\tau,x)=&\frac{1}{nh_j^{d+1}}\sum_{t=1}^n K^2\left(\frac{\tau_t-\tau}{h_j}\right)L^2\left(\frac{x_t-x}{h_j}\right)e_{t+j}^2\\&+\frac{2}{nh_j^{d+1}}\sum_{t=2}^n\sum_{s=1}^{t-1}K\left(\frac{\tau_t-\tau}{h_j}\right)K\left(\frac{\tau_s-\tau}{h_j}\right)L\left(\frac{x_t-x}{h_j}\right)L\left(\frac{x_s-x}{h_j}\right)e_{t+j}e_{s+j},\\&\equiv A_1+A_2,\end{aligned}$$

where $A_1=\frac{1}{nh_j^{d+1}}\sum_{t=1}^n K^2\left(\frac{\tau_t-\tau}{h_j}\right)L^2\left(\frac{x_t-x}{h_j}\right)e_{t+j}^2$ and
$A_2=\frac{2}{nh_j^{d+1}}\sum_{t=2}^n\sum_{s=1}^{t-1}K\left(\frac{\tau_t-\tau}{h_j}\right)K\left(\frac{\tau_s-\tau}{h_j}\right)L\left(\frac{x_t-x}{h_j}\right)L\left(\frac{x_s-x}{h_j}\right)e_{t+j}e_{s+j}$.

By iterated expectations and change of variables, we can show that

$$\mathbb{E}[A_1]=\frac{1}{nh_j^{d+1}}\sum_{t=1}^n K^2\left(\frac{\tau_t-\tau}{h_j}\right)\mathbb{E}\left[L^2\left(\frac{x_t-x}{h_j}\right)e_{t+j}^2\right] \tag{38}$$

$$
\begin{aligned}
&= \frac{1}{h_j^{d+1}} K^2\left(\frac{\tau_t-\tau}{h_j}\right) \mathbb{E}\left[L^2\left(\frac{x_t-x}{h_j}\right)\mathbb{E}[e_{t+j}^2|x_t=x]\right] \\
&= \frac{\sigma_j^2(x)}{h_j^{d+1}} K^2\left(\frac{\tau_t-\tau}{h_j}\right) \mathbb{E}\left[L^2\left(\frac{x_t-x}{h_j}\right)\right] \\
&= \frac{\sigma_j^2(x)}{h_j^{d+1}} K^2\left(\frac{\tau_t-\tau}{h_j}\right) \mathbb{E}\left[L^2\left(\frac{x_t(\tau_t)-x}{h_j}\right)\right](1+o_P(1)) \\
&= \frac{\sigma_j^2(x)}{h_j^{d+1}} K^2\left(\frac{\tau_t-\tau}{h_j}\right) \int L^2\left(\frac{x_y-x}{h_j}\right) f(\tau_t,x_y)\, dx_y (1+o_P(1)) \\
&= \frac{\sigma_j^2(x)}{h_j^{d+1}} \iint K^2\left(\frac{\tau_y-\tau}{h_j}\right) L^2\left(\frac{x_y-x}{h_j}\right) f(\tau_y,x_y)\, d\tau_y\, dx_y (1+o_P(1)) \\
&= \sigma_j^2(x) \int\cdots\int K^2(p)\prod_{i=1}^d K^2(q_i) f(\tau,x)\, dp\, dq^1\cdots dq^d \\
&+\sigma_j^2(x) \int\cdots\int K^2(p)\prod_{i=1}^d K^2(q_i)\left(ph_j\frac{\partial f(\tau,x)}{\partial\tau}+\sum_{i=1}^d q_i h_j \frac{\partial f(\tau,x)}{\partial x^i}\right) dp\, dq^1\cdots dq^d \\
&+\sigma_j^2(x) \int\cdots\int K^2(p)\prod_{i=1}^d K^2(q_i)\frac{1}{2}\left(p^2h_j^2\frac{\partial^2 f(\tau,x)}{\partial\tau^2}+\sum_{i=1}^d q_i^2 h_j^2 \frac{\partial^2 f(\tau,x)}{\partial {x^i}^2}\right) dp\, dq^1\cdots dq^d \\
&+\sigma_j^2(x) \int\cdots\int K^2(p)\prod_{i=1}^d K^2(q_i)\frac{1}{2}\Bigg(\sum_{i=1}^d pq_i h_j^2 \frac{\partial^2 f(\tau,x)}{\partial\tau\partial x^i} \\
&+ 2\sum_{i=2}^d\sum_{s=1}^{i-1} q_i q_s h_j^2 \frac{\partial^2 f(\tau,x)}{\partial x^i \partial x^s}\Bigg) dp\, dq^1\cdots dq^d + o_P(h_j^2) = \sigma_j^2(x) f(\tau,x)\kappa_0^{d+1} + O_P(h_j).
\end{aligned}
$$

Meanwhile, by the same steps as in Theorem 1 of Hansen (2008), we have that $\mathbb{E}[A_2] = o_P(1)$. Therefore, we can obtain that $\mathrm{Var}[V_j(\tau,x)] = \sigma_j^2(x) f(\tau,x)\kappa_0^{d+1} + o_P(1)$.

We then use the small-block and big-block arguments (refer to Fan and Yao (2003)), that is, decompose $V_j(\tau,x)$ alternately into big blocks and small blocks, we can neglect the small blocks and use the mixing conditions to replace the big blocks by independent random variables. Then apply a Lindeberg theorem, we can get that $V_j(\tau,x) \to_D N(0, \kappa_0^{d+1}\sigma_j^2(x) f(\tau,x))$. The proof is in the same spirit as that for the standard strictly stationary setting.

By similar argument as the (35), we have that

$$
\mathbb{E}\widehat{f}(\tau,x) = \mathbb{E}\widehat{f}^*(\tau,x)(1+o_P(1)).
$$

Then the bias of $\widehat{f}(\tau,x)$ will be

$$
\mathbb{E}\widehat{f}(\tau,x) - f(\tau,x) = \frac{\kappa_2}{2} h_j^2 \left(\frac{\partial^2 f(\tau,x)}{\partial\tau^2} + \sum_{i=1}^d \frac{\partial^2 f(\tau,x)}{\partial {x^i}^2}\right) + o_P(h_j^2). \tag{39}
$$

Following the similar steps of the proof of Theorem 1.1 in Li and Racine (2007), we can obtain the variance of $\widehat{f}(\tau,x)$:

$$
\mathrm{Var}\left(\widehat{f}(\tau,x)\right) = \frac{1}{nh_j^{d+1}}\left(\kappa_0^{d+1} f(\tau,x) + O_P(h_j)\right). \tag{40}
$$

Based on equations (39) and (40) and Assumption A.1.4(ii), we can obtain that $\widehat{f}(\tau,x)-f(\tau,x)=o_P(1)$. It is also straightforward to see that $\widehat{f}(\tau,x)^{-1}=O_P(1)$.

Then $V_j(\tau,x)/\widehat{f}(\tau,x)\to_D N(0,V_{j,\tau,x})$, where $V_{j,\tau,x}=\kappa_0^{d+1}\sigma_j^2(x)/f(\tau,x)$. Combining with equation (37), we have

$$\sqrt{nh_j^{d+1}}\left(\widehat{g}_j(\tau,x)-g_j(\tau,x)-B_{j,\tau,x}\right)\to_D N(0,V_{j,\tau,x}), \tag{41}$$

Therefore, we have completed the proof of Theorem 2.1.

**Proof of Theorem 2.2.**

Observe that

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

which gives

$$\begin{aligned}&\left(\sum_{j=1}^J\widehat{g}_j(\tau,x)-\sum_{j=1}^J g_j(\tau,x)\right)\\ &= \left(nh^{d+1}\right)^{-1/2} f(\tau,x)^{-1}(1+o_P(1))\sum_{j=1}^J\rho_j^{-(d+1)/2}\,V_j(\tau,x)\\ &+ \left(nh^{d+1}\right)^{-1/2} f(\tau,x)^{-1}(1+o_P(1))\sum_{j=1}^J\rho_j^{-(d+1)/2}\,B_j(\tau,x)\\ &\equiv \left(nh^{d+1}\right)^{-1/2} f(\tau,x)^{-1}(1+o_P(1))\,S_{nJ}(\tau,x)+\left(nh^{d+1}\right)^{-1/2} f(\tau,x)^{-1}(1+o_P(1))\,R_{nJ}(\tau,x),\end{aligned} \tag{43}$$

where

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

$$\begin{aligned}R_{nJ}(\tau,x) &= \sum_{j=1}^J\rho_j^{-(d+1)/2}\,B_j(\tau,x)\\ &= \left(nh^{d+1}\right)^{-1/2}\sum_{t=1}^n\left(\sum_{j=1}^J\rho_j^{-(d+1)}K\left(\frac{\tau_t-\tau}{h_j}\right)L\left(\frac{x_t-x}{h_j}\right)\left(g_j(\tau_t,x_t)-g_j(\tau,x)\right)\right).\end{aligned} \tag{45}$$

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

$$S_{nJ}(\tau,x)=\sum_{j=1}^J\rho_j^{-\frac{d+1}{2}}V_j(\tau,x)$$

$$= (nh^{d+1})^{-\frac{1}{2}} \sum_{j=1}^{J} \rho_j^{-(d+1)} \sum_{t=1}^{n} K\left(\frac{\tau_t - \tau}{h_j}\right) L\left(\frac{x_t - x}{h_j}\right) e_{t+j}$$
$$= (nh^{d+1})^{-\frac{1}{2}} T_{nJ}(\tau, x),$$

where $T_{nJ}(\tau, x) = \sum_{t=1}^{n} \left( \sum_{j=1}^{J} \rho_j^{-(d+1)} K\left(\frac{\tau_t - \tau}{h_j}\right) L\left(\frac{x_t - x}{h_j}\right) e_{t+j} \right) = \sum_{t=1}^{n} U_t(J)$, in which

$$U_t(J) = \sum_{j=1}^{J} \rho_j^{-(d+1)} K\left(\frac{\tau_t - \tau}{h_j}\right) L\left(\frac{x_t - x}{h_j}\right) e_{t+j}.$$

It is easy to see that

$$T_{nJ}^2(\tau, x) = \sum_{t=1}^{n} U_t^2(J) + 2 \sum_{t=2}^{n} \sum_{s=1}^{t-1} U_t(J) U_s(J),$$

and

$$U_t^2(J) = \sum_{j=1}^{J} \rho_j^{-2(d+1)} K^2\left(\frac{\tau_t - \tau}{h_j}\right) L^2\left(\frac{x_t - x}{h_j}\right) e_{t+j}^2$$
$$+ 2 \sum_{i=2}^{J} \sum_{j=1}^{i-1} \rho_i^{-(d+1)} \rho_j^{-(d+1)} K\left(\frac{\tau_t - \tau}{h_i}\right) K\left(\frac{\tau_t - \tau}{h_j}\right) L\left(\frac{x_t - x}{h_i}\right) L\left(\frac{x_t - x}{h_j}\right) e_{t+i} e_{t+j}.$$

According to equation (38), we can show that

$$\mathbb{E}[U_t^2(J)] = \sum_{j=1}^{J} \rho_j^{-2(d+1)} K^2\left(\frac{\tau_t - \tau}{h_j}\right) \mathbb{E}\left[\mathbb{E}[e_{t+j}^2 \mid x_t = x] L^2\left(\frac{x_t - x}{h_j}\right)\right]$$
$$= \sum_{j=1}^{J} \rho_j^{-2(d+1)} \sigma_j^2(x) h_j^{d+1} f(\tau, x) \kappa_0^{d+1} (1 + o_P(1))$$
$$= f(\tau, x) \kappa_0^{d+1} h^{d+1} \sum_{j=1}^{J} \sigma_j^2(x) \rho_j^{-(d+1)} (1 + o_P(1)).$$

Similar to the derivation of variance of $V_j(\tau, x)$ in Theorem 2.1, we have

$$\mathrm{Var}(T_{nJ}^2(\tau, x)) = \mathbb{E}[T_{nJ}^2(\tau, x)] = \mathbb{E}\left[\sum_{t=1}^{n} U_t^2(J)\right] + o_P(1)$$
$$= n f(\tau, x) \kappa_0^{d+1} h^{d+1} \sum_{j=1}^{J} \sigma_j^2(x) \rho_j^{-(d+1)} + o_P(1).$$

Hence, we have

$$\mathrm{Var}\left(S_{nJ}(\tau, x)\right) = \mathbb{E}\left[S_{nJ}^2(\tau, x)\right] = (nh^{d+1})^{-1} \mathbb{E}\left[T_{nJ}^2(\tau, x)\right]$$
$$= n^{-1} \frac{1}{h^{d+1}} n f(\tau, x) \kappa_0^{d+1} h^{d+1} \sum_{j=1}^{J} \sigma_j^2(x) \rho_j^{-(d+1)} + o_P(1)$$
$$= f(\tau, x) \kappa_0^{d+1} \sum_{j=1}^{J} \sigma_j^2(x) \rho_j^{-(d+1)} + o_P(1).$$

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

$$\left(\sum_{j=1}^{J} \rho_j^{-(d+1)} \sigma_j^2(x)\right)^{-1/2} S_{nJ}(\tau, x) \to_D N\left(0, f(\tau,x)\kappa_0^{d+1}\right). \tag{46}$$

From equation (43), we have that

$$f(\tau,x)\sqrt{nh^{d+1}}\left(\sum_{j=1}^{J}\widehat{g}_j(\tau,x) - \sum_{j=1}^{J} g_j(\tau,x)\right) = (1+o_P(1))S_{nJ}(\tau,x) + (1+o_P(1))R_{nJ}(\tau,x)$$

Let $B_J(\tau,x) = \frac{R_{nJ}(\tau,x)}{f(\tau,x)\sqrt{nh^{d+1}}}$. Then based on equation (37) and under Assumption A.1.4(ii), it is easy to show that

$$\begin{aligned} B_J(\tau,x) &= \sum_{j=1}^{J}(h_j^2 R_j(\tau,x) + h_j^2 b_j(\tau,x)) \\ &= \sum_{j=1}^{J}\rho_j^2 h^2 \kappa_2 \left[\frac{1}{2}\frac{\partial^2 g_j(\tau,x)}{\partial \tau^2} + \frac{1}{2}\sum_{i=1}^{d}\frac{\partial^2 g_j(\tau,x)}{\partial x^{i2}} + f^{-1}(\tau,x)\frac{\partial f(\tau,x)}{\partial \tau}\frac{\partial g_j(\tau,x)}{\partial \tau} + f^{-1}(\tau,x)\sum_{i=1}^{d}\frac{\partial f(\tau,x)}{\partial x^i}\frac{\partial g_i(\tau,x)}{\partial x^i}\right]. \end{aligned}$$

Then we have

$$f(\tau,x)\sqrt{nh^{d+1}}\left(\sum_{j=1}^{J}\widehat{g}_j(\tau,x) - \sum_{j=1}^{J} g_j(\tau,x) - B_J(\tau,x)\right) = (1+o_P(1))S_{nJ}(\tau,x),$$

which shows that

$$f(\tau,x)\sqrt{nh^{d+1}\Sigma_J^{-1}(x)}\left(\sum_{j=1}^{J}\widehat{g}_j(\tau,x) - \sum_{j=1}^{J}\widehat{g}_j(\tau,x) - B_J(\tau,x)\right) \to_D N\left(0, f(\tau,x)\kappa_0^{d+1}\right),$$

where $\Sigma_J(x) = \sum_{j=1}^{J}\rho_j^{-(d+1)}\sigma_j^2(x)$.

Therefore, we have

$$\sqrt{nh^{d+1}\Sigma_J^{-1}(x)}\left(\sum_{j=1}^{J}\widehat{g}_j(\tau,x) - \sum_{j=1}^{J} g_j(\tau,x) - B_J(\tau,x)\right) \to_D N\left(0, V(\tau,x)\right),$$

where $V(\tau,x) = \kappa_0^{d+1} f^{-1}(\tau,x)$.

Therefore, we have completed the proof of Theorem 2.2.

## A.2. 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.2.1** (i) $\{x_t\}$ is locally stationary with associated process $\{x_t(\tau)\}$, and all $x_t$ $(1 \le t \le n)$ have the same compact support $V = [a_{\min}, a_{\max}]$. Moreover, the density $f(\tau,x)$ of $x_t(\tau)$ is smooth in $\tau$. (ii) For each $\tau \in [0,1]$, $x_t(\tau)$ is a strictly stationary and $\alpha$-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.2.2** 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.2.3** For all $t$ and any $\tau \in [0,1]$, $x_t(\tau)$ is independent of $\{e_s, -\infty < s < \infty\}$.

**Assumption A.2.4** 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.2.5** (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.2.6** 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.2.1–A.2.4 allow us to approximate the locally stationary variable $x_t$ by stationary variable $x_t(\tau)$ when $\tau_t$ is in a small neighborhood of $\tau$. In this paper, we require the support of the locally stationary process to be compact. Assumption A.2.5 (i) imposes a smoothness condition on the unknown functions, which is to guarantee a certain rate of the convergence. Assumption A.2.5(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.2.6 imposes the rates of divergence on $k_{1j}$ and $k_{2j}$, which guarantee the convergence of the proposed estimators.

**Proof of Theorem 2.3.**

Let $D_{nj} = \text{diag}(\sqrt{n}I_{k_{1j}}, \sqrt{n}I_{k_{2j}})$ 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 (2018), 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 $x \in V$,

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

$$= \Phi_j(\tau,x)^\top U_{k_j}^{-1} D_{nj}^{-1} B_{nk_j}^\top (\gamma_{(j)} + e_{(j)}) + \begin{pmatrix} \sqrt{n}\gamma_{k_{1j}}(\tau) \\ \sqrt{n}\gamma_{k_{2j}}(x) \end{pmatrix}.$$

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 (2018).

**Proof of Theorem 2.4.**

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 $n \to \infty$

$$\sqrt{n}\Sigma_{nJ}^{-1/2}\left(\widehat{m}(\tau,x) - m(\tau,x)\right) \to_D N(0,1), \tag{47}$$

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

# References


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. (2018), 'Nonparametric localized bandwidth selection in kernel density estimation', *Econometric Reviews* **forthcoming**.

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. (2018), 'Additive nonparametric models with time variable and both stationary and nonstationary regressors', *Journal of Econometrics* **207**(1), 212–236.

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.

Fan, J. and Yao, Q. (2003), *Nonlinear Time Series: Nonparametric and Parametric Methods*, Springer, New York.

Hall, P. G. 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.

Hansen, L. P. and Hodrick, R. J. (1980), 'Forward exchange rates as optimal predictors of future spot rates: an econometric analysis', *The Journal of Political Economy* **88**(5), 829–853.

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.

Hodrick, R. J. (1992), 'Dividend yields and expected stock returns: alternative procedures for inference and measurement', *The Review of Financial Studies* **5**(3), 357–386.

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

Nielsen, J. P. and Sperlich, S. (2003), 'Prediction of stock returns: A new way to look at it', *Astin Bulletin* **33**(2), 399–417.

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* .

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.

Scholz, M., Sperlich, S. and Nielsen, J. P. (2016), 'Nonparametric long term prediction of stock returns with generated bond yields', *Insurance: Mathematics and Economics* **69**, 82–96.

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.