‘A FLEXIBLE SEMIPARAMETRIC FORECASTING MODEL FOR TIME SERIES’

AUTHOR(S):

Dr Degui Li

Professor Oliver Linton, Faculty of Economics, Austin Robinson Building, Sidgwick Avenue, Cambridge University, Cambridge, UK.

Professor Zudi Lu, School of Mathematical Sciences & Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, UK.
A Flexible Semiparametric Forecasting Model for Time Series

Degui Li*  Oliver Linton†  Zudi Lu‡
University of York  Cambridge University  University of Southampton

November 24, 2014

Abstract

In this paper, we propose a semiparametric procedure called the “Model Averaging MArginal Regression” (MAMAR) that is flexible for forecasting of time series. This procedure considers approximating a multivariate regression function by an affine combination of one-dimensional marginal regression functions. The weight parameters involved in the approximation are estimated by least squares on the basis of the first-stage nonparametric kernel estimates of the marginal regressions. Under some mild conditions, we have established asymptotic normality for the estimated weights and the regression function in two cases: Case I considers the number of the covariates is fixed while Case II allows the number of the covariates depending on the sample size and diverging. As the observations are assumed to be stationary and near epoch dependent, the approach developed is applicable to both the estimation and forecasting issues in time series analysis. Furthermore, the method and result are augmented by a simulation study and illustrated by an application in forecasting the high frequency volatility of the FTSE100 index.

JEL subject classifications: C14, C22.

Keywords: Forecasting, marginal regression, model averaging, kernel estimation, near epoch dependence, semiparametric estimation.

*Department of Mathematics, University of York, York, YO10 5DD, UK, email: degui.li@york.ac.uk.
†Faculty of Economics, Austin Robinson Building, Sidgwick Avenue, Cambridge University, Cambridge, CB3 9DD, UK, email: obl20@cam.ac.uk.
‡School of Mathematical Sciences & Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, SO17 1BJ, UK, email: Z.Lu@soton.ac.uk.
1 Introduction

In many situations of practical interest, we are faced with a large number of variables and uncertain functional forms. Linearity is widely adopted in macroeconometrics where data is limited, but for many relationships this linearity may imply absurd conclusions when covariates are pushed to extreme values. Moreover, in the regression settings, we may have to choose between a large number of covariates. In the case of time series, the problem can get even worse, since in both estimation and forecasting, all possible lags of all possible predictor variables may be the candidates and their influences are of unknown forms. One approach to deal with this problem is to use model selection tools that choose the best model according to some traditional criterion from a set of models. In some cases, such an approach can be very time consuming. Also, it may be neglecting features of the data that arrive through the models which are not selected but are almost as good as those which are selected. A popular method is to use model averaging whereby we fit a number of candidate models and then weight them according to some criterion (see, for example, Hansen 2007, Liang et al 2011). Another approach that is now popular in statistics is to use some penalization device to force many weights to be zero. For instance, the least absolute shrinkage and selection operator (LASSO) proposed by Tibshirani (1996, 1997), is the penalized least squares estimate with the $L_1$ penalty. The penalized regression with general $L_q$ penalty leads to a bridge regression (Frank and Friedman 1993, Fu 1998). Fan and Li (2001) used the smoothly clipped absolute deviation (SCAD) penalty in penalized likelihood estimation. For recent developments and surveys on model averaging and variable selection, the reader is referred to Claeskens and Hjort (2008), Fan and Lv (2008, 2010), Bühlmann and van de Geer (2011) and the references therein. However, most of the literature regarding model averaging and selection has been concerned with parametric models, which assume some parametrically linear or nonlinear relationships among the variables considered. In this paper, we will consider nonparametric and semiparametric models and will focus on with model averaging, leaving the penalty model selection issue in the setting of this paper to our following future work.

Specifically, let $(Y_t, X_t^\top)$ be a stationary time series process, where $X_t = (X_{t1}, \ldots, X_{td})^\top$ is a $d$-dimensional random vector and the superscript $\top$ stands for the transpose of a vector or matrix. In many applications, we need to consider estimating regression function $E(Y_t|X_t = x)$ with $x = (x_1, \ldots, x_d)^\top$, which can be well estimated by nonparametric method when the dimension $d$ is small, but very poorly if the dimension $d$ is high (say larger than 3) owing to the so-called “curse of dimensionality”. Various nonparametric and semiparametric models, such as additive models, varying coefficient models, partially linear models, have been studied to deal with the curse of dimensionality problem in the literature (see, for example, Fan and Yao 2003, Li and Racine 2007, Teräsvirta et al
2010). In the time series case, as mentioned above, the conditioning information may consist of an infinite number of lags, i.e., $d = \infty$. Linton and Sancetta (2009) established consistency of estimators of $E(Y_t|Y_{t-1}, Y_{t-2} \ldots)$ under weak conditions without any functional form restrictions beyond some limited smoothness, but rates of convergence are not available and practical performance is likely to be poor without further restrictions. Instead, it makes sense to use lower dimensional predictors, but which one? We next consider some explicit semiparametric models that have been tried to address the issue in nonlinear time series.

Linton and Mammen (2005) considered the semiparametric (volatility) regression model

$$E(Y_t^2|Y_{t-1}, Y_{t-2} \ldots) = \sum_{j=1}^{\infty} \psi_j(\theta)m(Y_{t-j}),$$

where $m(\cdot)$ is an unknown function and the parametric family $\{\psi_j(\theta), \theta \in \Theta, j = 1, \ldots, \infty\}$ satisfies some regularity conditions. This model includes the GARCH(1,1) as a special case and also includes an infinite set of lags. They assumed that $\{Y_t\}$ is stationary and geometrically mixing and thereby obtained a characterization of the function $m$ as the solution of a linear integral equation with intercept of the form $m^*_\theta(x) = \sum_{j=1}^{\infty} \psi_j(\theta)m_j(x)$, where $m_j(x) = E(Y_t^2|Y_{t-j} = x)$ for each $j$. They proposed an estimation strategy for the unknown quantities, which requires as input the estimation of $m_j(x)$ for $j = 1, 2, \ldots, J(T)$, where $J(T) = c\log T$ for some $c > 0$. They required to bound the estimation error of $m_j(x)$ uniformly over $x$ and over $j = 1, 2, \ldots, J(T)$. However, they provided only a sketch proof of this result in the case where the process is assumed to have compact support and to be strong mixing with geometric decay. A recent paper by Li et al (2012) provided a more rigorous and complete proof of this result. Linton and Mammen (2008) generalized this class of models to allow for exogenous regressors and more complicated dynamics. See Chen and Ghyssels (2010) for an application of these methods to volatility forecasting.

This general approach to modelling is promising but quite computationally demanding. In addition, the models considered thus far all have a finite number of unknown functions (for example, in Linton and Mammen (2005) only one unknown function was allowed), and so appear to be heavily over identified. In this paper, we aim at relaxing such restrictive assumptions and consider a semiparametric model that contains possibly infinitely many unknown functions all of which can enter into the prediction. This may be particularly useful in situations where there is a lot of nonlinearity and a rich dynamic structure. The most general version of our model is similar in some ways to the setting considered in Hansen (2007) except instead of observed covariates we have nonparametrically estimated ones. We call our method MAMAR (Model Averaging MArginal Regression). We obtain consistency and asymptotic normality of our procedure under general conditions. We further
apply our method to volatility forecasting (where the time series is long and (log) linear models are predominant) and obtain some satisfactory results.

The rest of the paper is organized as follows. The approximation of MAMAR is presented in Section 2 and the semiparametric estimation method is presented in Section 3. The asymptotic properties for the estimators of the optimal weights and nonparametric estimators for finite covariates case are provided in Section 4.1, and Section 4.2 gives the theoretical results when the dimension of the covariates is diverging. Discussions of some related topics are given in Section 5. Numerical evidence of our methodology is given in Section 6. Section 7 concludes this paper. All the technical lemmas and the proofs of the main results are collected in the appendix. A supplemental document gives the application of our method to Australian temperature data and the proofs of some auxiliary results.

## 2 Approximation with MAMAR

We model or approximate the conditional regression function $E(Y|X = x)$ by an affine combination of lower dimensional regression functions. Let $S_ℓ$ denote the set of all subsets of $S = \{1, 2, \ldots, d\}$ of $ℓ$ components, and this has cardinality $J_ℓ = \binom{d}{ℓ}$. For example, $S_2 = \{(1, 2), \ldots, (d - 1, d)\}$ has cardinality $d(d - 1)/2$. We model or approximate $m(x) = E(Y|X = x)$ by

$$m_w(x) = w_0 + \sum_{j=1}^{J} w_j E(Y|X_{(j)} = x_{(j)})$$

for some weights $w_j$, $j = 0, 1, \ldots, J$, where $X_{(j)} = (X_{i1}, \ldots, X_{ik_j})^T$ is a subset of $X$ and $x_{(j)} = (x_{i1}, \ldots, x_{ik_j})^T$. In general, $X_{(j)}$ and $X_{(k)}$ could have different dimensions and of course overlapping members. The union of $X_{(j)}$ may exhaust one or more of $S_ℓ$ or it may not. A simple special case that we focus on for much of the paper is where $J = d$ and $X_{(j)} = X_j$ is just the $j^{th}$ component and the covariates are non overlapping. This seems well suited to time series applications. In practice, one would not wish to take $k_j$ to be too large, so as to avoid the curse of dimensionality.

We could be thinking of this as a family of models within which there is a true member that corresponds to the true regression function $m(x)$ or we could be thinking of this as an approximating or model averaging device. Either way, we are seeking $w = (w_0, w_1, \ldots, w_J)^T$ that minimizes

$$E\left[ Y - w_0 - \sum_{j=1}^{J} w_j E(Y|X_{(j)}) \right]^2.$$  
(2.1)
In general, the minimizing weights may not be unique, but the minimization problem is a projection onto the space spanned by the functions \( \{ E(Y | X(j)), j = 1, \ldots, J \} \) and so there is a unique solution \( m_w(x) \). We shall focus on the special case where there is a unique vector \( w \) (which is generally true for the special case where \( J = d \) and \( X(j) = X_j \) is just the \( j \)th component and the covariates are non overlapping). In this case, the minimizer to (2.1), \( w_o = (w_{o,0}, w_{o,1}, \ldots, w_{o,J})^T \), satisfies

\[
w_{o,0} = (1 - \sum_{j=1}^J w_{o,j}) E(Y), \quad (w_{o,1}, \ldots, w_{o,J})^T = A^{-1}a,
\]

where \( A \) is a \( J \times J \) matrix whose \((i, j)\)th component is \( \text{Cov}(E(Y | X(i)), E(Y | X(j))) \), and \( a \) is a \( J \)-dimensional vector whose \( i \)th component is \( \text{Cov}(E(Y | X(i)), Y) \). If the model is true or the approximation is perfect, (2.1) is equal to zero at the optimal weights but it need not be so. Obviously the conditional component regressions \( E(Y | X(j) = x_j), j = 1, \ldots, J \), are unknown but low dimensional, so they can be well estimated by various nonparametric approaches. In Section 3, we will first estimate these conditional regression functions by the Nadaraya-Watson method and then use the least squares approach to obtain the estimator of \( w_o \). We can consider this approach as a form of model averaging where we are averaging the “models”: \( E(Y | X(j) = x(j)), j = 1, \ldots, J \), see Hansen (2007). We can also think of this as a pragmatic use of lower dimensional relationships to build a more complex predictor.\(^{1}\)

We now confine our attention to the simple special case where \( J = d \) and \( X(j) = X_j \) is just the \( j \)th component and the covariates are non overlapping. Such model setting is partially inspired by the recent work on the sure independence screening for variable selection in high-dimensional feature space, where the marginal regression is used to choose relevant covariates, see, for example, Fan and Lv (2008) and Fan et al (2011) where the observations are assumed to be independent. In this paper, we consider the stationary and nonlinear time series data, and discuss the case where the “model” is not necessarily true, and make a comparison with other “models”. Note that the above fit is equal to the full linear regression model fit in the parametric linear case.\(^{2}\) However, in the general

\(^{1}\)The idea of considering pairwise relationships has been considered in Hong (2000) for testing the serial independence of an observed scalar series \( Y_t \). In practice checking the independence of \( Y_t \) from \( Y_{t-1}, Y_{t-2}, \ldots \) is very difficult due to the curse of dimensionality. He thus proposed to check all pairwise joint relationships \( (Y_t, Y_{t-j}) \) for departures from the null.

\(^{2}\)This can also be true in infinite dimensional settings. Consider the AR(\( \infty \)) model of the form \( Y_t = \sum_{j=1}^{\infty} \rho_j Y_{t-j} + \varepsilon_t \) for some (declining) coefficients \( \rho_j \). Our general class of models would include processes of the form \( Y_t = \sum_{j=1}^{\infty} \theta_j E(Y_t | Y_{t-j}) + \varepsilon_t \). In the special linear Gaussian case, the two representations are equivalent (since all the \( E(Y_t | Y_{t-j}) \) are linear functions). However, in general they will be different, even in the linear but non-Gaussian case (Tong, 1990, p13).
nonparametric case, this is not necessarily so, i.e.,

$$E(Y|X) \neq w_0 + \sum_{j=1}^{d} w_j E(Y|X_j),$$

although there are clearly some nonlinear cases where this is so and where this would be a reasonable approximation. We discuss this further below.

The fit can be seen as an additive function of the individual components, i.e.,

$$w_0 + \sum_{j=1}^{d} w_j E(Y|X_j) =: w_0 + \sum_{j=1}^{d} g_j(X_j)$$

for a specific set of functions $g_j(X_j)$. Therefore, generally speaking,

$$\inf_w E\left[ Y - w_0 - \sum_{j=1}^{d} w_j E(Y|X_j) \right]^2 \geq \inf_h E\left[ Y - w_0 - \sum_{j=1}^{d} h_j(X_j) \right]^2,$$

which means that the best additive fit has lower mean squared error (if the model is true the MSE will be the same, but when the model is not true, the additive approximation is better). In fact, we can also interpret our procedure through the repeated projection argument that

$$\inf_w E\left[ E_{Add}(Y|X) - w_0 - \sum_{j=1}^{d} w_j E(Y|X_j) \right]^2,$$

where $E_{Add}(Y|X)$ is the best additive fit of $Y$ by the vector $X$. The space generated by $w_0 + \sum_{j=1}^{d} w_j E(Y|X_j)$ is a linear subspace of the space of additive functions and so we cannot do as well as unrestricted additive fitting (see, Nielsen and Linton 1998, Mammen et al 1999, Linton 2000, and Nielsen and Sperlich 2005 for more discussion on estimation of the additive models). In other words, our model is a restricted subclass of the class of additive models. In order to explore this further, we suppose that the regression function is additive, i.e.,

$$E(Y|X) = h_0 + \sum_{j=1}^{d} h_j(X_j)$$

for some functions $h_j$. We may ask under which conditions can we write (with probability one)

$$h_0 + \sum_{j=1}^{d} h_j(X_j) = \omega_0 + \sum_{j=1}^{d} \omega_j E(Y|X_j) \quad (2.3)$$
for some choice of weights $\omega_j$. Clearly, when the covariates are mutually independent (2.3) will hold, but we show next that the full independence is not necessary. A sufficient condition for (2.3) is that there exist constants $\{\alpha_{ji}, \beta_{ji}\}$ such that with probability one

$$E[h_j(X_j)|X_i] = \alpha_{ji} + \beta_{ji} h_i(X_i)$$

(2.4)

for all $1 \leq i, j \leq d$. If the condition (2.4) is satisfied, we have

$$E(Y|X_i) = h_i(X_i) + \sum_{j \neq i} E[h_j(X_j)|X_i] = h_i(X_i) \left(1 + \sum_{j \neq i} \beta_{ji}\right) + \sum_{j \neq i} \alpha_{ji}.$$ 

It follows that with the choice of $\omega_i = (1 + \sum_{j \neq i} \beta_{ji})^{-1}$ and $\omega_0 = h_0 - \sum_{i=1}^d \omega_i \sum_{j \neq i} \alpha_{ji}$, we can achieve the objective (2.3) with probability one. We next consider a non trivial special case where condition (2.4) is satisfied. Suppose that $E(Y|X) = \alpha + \sum_{j=1}^d \beta_j X_j^2$, where $X_j$ are pairwise standard normal with correlation coefficients $\rho_{ij}$, and $d \leq \infty$. Then we have $X_j = \rho_{ij} X_i + \sqrt{1 - \rho_{ij}^2} Z_{ij}$, where $Z_{ij}$ is standard normal and independent of $X_i$. Therefore, for all $i, j$ we have $E(X_j^2|X_i) = \rho_{ij}^2 X_i^2 + 1 - \rho_{ij}^2$, which satisfies our condition (2.4). This example is actually quite relevant to the case of volatility modelling. It was noted in Linton and Mammen (2005) (and elsewhere) that the marginal regression functions $E(Y^2_t|Y_{t-j})$, where $Y_t$ was stock returns, are quite similar in shape (and fairly close to quadratic functions). In fact, the similarity in shape of the marginal regression functions was a motivation for the class of semiparametric models considered there.

Let us now compare our model with the type of model considered in Linton and Mammen (2005), which is another kind of restricted additive model. Specifically, consider the regression model where it is hypothesized that

$$E(Y|X) = w_0 + \sum_{j=1}^d w_j m(X_j)$$

(2.5)

for some single unknown function $m(\cdot)$ and weights $w_j$. It is easy to see that in general (meaning when the true regression function does not satisfy (2.5)) the MSE of the two models is non-nested. That is, there are regression functions for which

$$\inf_w E \left[ Y - w_0 - \sum_{j=1}^d w_j E(Y|X_j) \right]^2 \leq \inf_{w,m} E \left[ Y - w_0 - \sum_{j=1}^d w_j m(X_j) \right]^2.$$ 

3In Linton and Mammen (2005), the function $m(\cdot)$ had the interpretation of a news impact curve. They took $d = \infty$ and restricted $w_j$ parametrically and thereby to decay rapidly to zero as $j \to \infty$. It seems that to allow for very large dimensions one must employ some restrictions on the additive regression.
In fact, we would generally expect this ordering of the MSE especially when \( d \) is large. This is actually the case we are most interested in.

We close with some further advantages of our approach. In the large \( d \) case, the smooth backfitting theory of Mammen \textit{et al} (1999) does not apply, whereas we allow for the case where \( d \to \infty \) in our theoretical development. One further advantage of our method is computational. As we can obtain the closed form for the parametric estimator of \( w_o \) and no iterative algorithm is involved (see Section 3 for details), the computational procedure of our method is not as time consuming as that for nonparametric additive models or restricted versions such as Linton and Mammen (2005).

The modelling approach is similar in some way to copulas. We allow general marginal regression relationships but glue them together in a parametric way through the weights to give the joint regression. A more general setting then would be

\[
E(Y|X) = C(E(Y|X_1), \ldots, E(Y|X_d); w),
\]

where \( w \) is a parameter vector and \( C \) is a “regression copula”, in our case linear in its arguments. We will discuss this further below in Section 5.2.

3 Semiparametric Estimation

In this section we define our estimation procedures using matrix formulae, which facilitate efficient coding. Without loss of generality we assume that the process is stationary and centered, i.e., \( E(Y) = 0 \), otherwise we replace \( Y \) by \( Y - E(Y) \) and \( Y_t \) by \( Y_t - \overline{Y} = Y_t - \frac{1}{n} \sum_{t=1}^{n} Y_t \). Suppose that we have stationary and weakly dependent observations \((Y_t, X_t^j), t = 1, \ldots, n\). Let \( m(x) = E(Y|X_t = x) \) and \( m_j(x_j) = E(Y_t|X_{t,j}) \), \( j = 1, \ldots, d \). We first estimate \( m_j(x_j) \) by using the Nadaraya-Watson kernel method

\[
\hat{m}_j(x_j) = \frac{\sum_{t=1}^{n} Y_t K\left(\frac{X_{t,j} - x_j}{h_j}\right)}{\sum_{t=1}^{n} K\left(\frac{X_{t,j} - x_j}{h_j}\right)},
\]

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

Since we are interested in estimating the marginal regression function at the sample points, we let \( \mathcal{M}_j = [m_j(X_{1j}), \ldots, m_j(X_{nj})]^T \) and \( \hat{\mathcal{M}}_j = [\hat{m}_j(X_{1j}), \ldots, \hat{m}_j(X_{nj})]^T \). Note that \( \hat{\mathcal{M}}_j \) is the Nadaraya-Watson estimator of \( \mathcal{M}_j \) and we have \( \hat{\mathcal{M}}_j = S_j \mathcal{Y} \), where \( S_j \) is the \( n \times n \) smoother matrix associated
with $X_j$, $\mathcal{Y}$ is the $n \times 1$ vector of observations on the response, $\mathcal{Y} = (Y_1, \ldots, Y_n)^\top$. Then, for given $w = (w_1, \ldots, w_d)^\top$,

$$\mathcal{M}_w = (w_1 S_1 + \cdots + w_d S_d) \mathcal{Y} =: S(w) \mathcal{Y}.$$  

As $\mathbb{E}(\mathcal{Y}) = 0$, it is easy to see that $w_{o,0} = 0$. Then, motivated by (2.1), to estimate $w^*_o = (w_{o,1}, \ldots, w_{o,d})^\top$, we define the least squares sample objective function by

$$Q(w) = (\mathcal{Y} - \mathcal{M}_w)^\top (\mathcal{Y} - \mathcal{M}_w) = \mathcal{Y}^\top (I_n - w_1 S_1 - \cdots - w_d S_d)^\top (I_n - w_1 S_1 - \cdots - w_d S_d) \mathcal{Y} = \text{Tr} [\mathcal{Y}^\top \mathcal{P}(w)],$$  

(3.2)

where $I_n$ is an $n \times n$ identity matrix and

$$\mathcal{P}(w) = (I_n - w_1 S_1 - \cdots - w_d S_d)^\top (I_n - w_1 S_1 - \cdots - w_d S_d)$$

$$= I_n - \sum_{j=1}^d w_j (S_j + S_j^\top) + \sum_{j=1}^d w_j^2 S_j^\top S_j + \sum_{i=1}^{d-1} \sum_{j=i+1}^d w_i w_j (S_i^\top S_j + S_j^\top S_i),$$

while $\text{Tr}(\cdot)$ is the trace of a square matrix. We then minimize $Q(w)$ with respect to the vector $w$ to obtain

$$\frac{\partial}{\partial w_i} Q(w) = \text{Tr} \left[ \mathcal{Y} \mathcal{Y}^\top \frac{\partial}{\partial w_i} \mathcal{P}(w) \right] = -\mathcal{Y}^\top (S_i + S_i^\top) \mathcal{Y} + 2 w_i \mathcal{Y}^\top S_i^\top S_i \mathcal{Y} + \sum_{j \neq i} w_j \mathcal{Y}^\top (S_i^\top S_j + S_j^\top S_i) \mathcal{Y} = 0.$$  

We can write this as

$$\hat{A} w = \hat{a}, \quad \hat{A} = \left( \hat{A}_{ij} \right)_{d \times d}, \quad \hat{a} = (\hat{a}_1, \ldots, \hat{a}_d)^\top,$$

where $\hat{A}_{ij} = \mathcal{Y}^\top (S_i^\top S_j + S_j^\top S_i) \mathcal{Y}$ and $\hat{a}_i = \mathcal{Y}^\top (S_i + S_i^\top) \mathcal{Y}$. Then we have

$$\hat{w} = (\hat{w}_{o,1}, \ldots, \hat{w}_{o,d})^\top = \hat{A}^{-1} \hat{a}.$$  

(3.3)

Defining

$$\hat{\mathcal{M}} = \begin{pmatrix} \hat{m}_1(X_{11}) & \cdots & \hat{m}_d(X_{1d}) \\ \vdots & \vdots & \vdots \\ \hat{m}_1(X_{n1}) & \cdots & \hat{m}_d(X_{nd}) \end{pmatrix},$$

by (3.2), $\hat{w}$ in (3.3) can be rewritten as

$$\hat{w} = (\hat{\mathcal{M}}^\top \hat{\mathcal{M}})^{-1} \hat{\mathcal{M}}^\top \mathcal{Y}.$$  

(3.4)
Finally, we can estimate the conditional regression function \( \sum_{i=1}^{d} w_{o,j} m_j(x_j) \) by

\[
\hat{m}(x) := \hat{m}(\tilde{x}) = \sum_{j=1}^{d} \hat{w}_{o,j} \hat{m}_j(x_j).
\]  

(3.5)

By the discussion in Section 2, \( \hat{m}(x) \) can be seen as an approximation to \( m(x) \) in the case where \( \sum_{i=1}^{d} w_{o,j} m_j(x_j) \) does not equal \( m(x) \).

When \( d \) is fixed, in Section 4.1 below, we will show that \( \hat{w} \) is asymptotically normal with root-\( n \) convergence rate and the nonparametric estimator \( \hat{m}(x) \) is also asymptotically normal with root-\( (nh) \) convergence rate. Thus, the curse of dimensionality is avoided. Furthermore, in Section 4.2, we will consider the more involved case that the dimension of \( X_t \) is diverging, i.e., \( d = d_n \to \infty \) as \( n \to \infty \), which is common in complex time series analysis.

4 Asymptotic Properties

In this paper, we assume that \( \{(Y_t, X_t^\prime), t \geq 1\} \) belongs to a class of stationary near epoch dependent (NED) or stable processes, which is more general than the commonly-assumed \( \alpha \)-mixing process. Based on a stationary process \( \{\varepsilon_t\}, \{Y_t\} \) and \( \{X_t\} \) are defined by

\[
Y_t = \Psi_Y(\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}, \ldots),
\]

\[
X_t = (X_{t1}, \ldots, X_{td})^\prime = \Psi_X(\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}, \ldots),
\]  

(4.1)

where \( \Psi_Y : \mathbb{R}^\infty \to \mathbb{R}^1 \) and \( \Psi_X : \mathbb{R}^\infty \to \mathbb{R}^d \) are two Borel measurable functions, and \( \{\varepsilon_t\} \) may be vector-valued. The definition of the NED process is provided as follows.

**Definition 4.1.** The stationary process \( \{(Y_t, X_t)\} \) is said to be near epoch dependent in \( L_\nu \) norm (NED in \( L_\nu \)) with respect to a stationary \( \alpha \)-mixing process \( \{\varepsilon_t\} \), if

\[
v_\nu(m) = E\left[|Y_t - Y_t^{(m)}|^{\nu} + \|X_t - X_t^{(m)}\|^{\nu}\right] \to 0, \quad \nu > 0,
\]  

(4.2)

as \( m \to \infty \), where \( |\cdot| \) and \( \|\cdot\| \) are the absolute value and the Euclidean norm, respectively, \( Y_t^{(m)} = \Psi_Y, m(\varepsilon_t, \ldots, \varepsilon_{t-m+1}), X_t^{(m)} = (X_{t1}^{(m)}, \ldots, X_{td}^{(m)})^\prime = \Psi_X, m(\varepsilon_t, \ldots, \varepsilon_{t-m+1}), \Psi_Y, m \) and \( \Psi_X, m \) are \( \mathbb{R}^1 \)- and \( \mathbb{R}^d \)-valued Borel measurable functions with \( m \) arguments, respectively, and \( v_\nu(m) \) is said to be the stability coefficients of order \( \nu \) of the process \( \{(Y_t, X_t)\} \).

From the above definition, we know that the NED process includes some popular time series models such as the AR(\( p \))-GARCH(1,1) model as a special case. The detailed discussion of the NED
process with applications in nonparametric estimation can be found in Lu and Linton (2007) and Li et al. (2012). In this section, we derive the asymptotic theory for \( \hat{w} \) and \( \hat{m}(x) \) for two cases: (i) the dimension of \( \{X_t\} \) is fixed; and (ii) the dimension of \( \{X_t\} \) increases with the sample size \( n \). We do not assume that the model is true, i.e., \( m_w(x) \) is not necessarily equal to \( m(x) \).

### 4.1 Case I: The dimension of \( X_t \) is fixed

We start with a simple case that \( d \) is fixed. In order to establish the asymptotic theory, we introduce some regularity conditions.

**Assumption 1** The kernel function \( K \) is continuous with a compact support. Furthermore, it satisfies \( \int K(u)du = 1 \), \( \int u^lK(u) = 0 \), for \( 1 \leq l \leq \gamma - 1 \), and \( 0 < \int u^\gamma K(u) < \infty \).

**Assumption 2**

(i) The joint density function of \( \{X_t\} \), \( f_X(\cdot) \), and the marginal density function of \( \{X_t\} \), \( f_j(\cdot) \), have continuous derivatives up the \((\gamma + 1)\)-order and \( \inf_{1 \leq j \leq d} \inf_{x \in \Omega} f_j(x) > 0 \), where \( \Omega_j \) is the compact support of \( X_tj \).

(ii) The joint density function of \( (X_{tj}, X_{t+k,j}) \), \( f_{j,k}(\cdot, \cdot) \), exists for \( 1 \leq j \leq d \) and \( k \geq 1 \), and satisfies that for some positive integer \( k^* \) and all \( k \geq k^* \), \( f_{j,k}(x_1, x_2) < C_f \) for \( 1 \leq j \leq d \), all \( (x_1, x_2) \in \mathbb{R}^2 \), \( 0 < C_f < \infty \).

(iii) The conditional density function of \( X_{tj} \) for given \( X_{tk} \), \( k \neq j \), \( f_{j|k}(\cdot|\cdot) \), exists and satisfies the Lipschitz continuous condition.

(iv) The conditional regression functions \( m(\cdot) \) and \( m_j(\cdot) \), \( 1 \leq j \leq d \), have continuous and bounded derivatives up to the \((\gamma + 1)\)-order.

**Assumption 3**

(i) \( \{(Y_t, X_t^t), t \geq 1\} \) is stationary NED in \( L^p_0 \)-norm with respect to a stationary \( \alpha \)-mixing process \( \{\varepsilon_t\} \) with \( \mathbb{E}[|Y_t|^{p_0}] < \infty \), where \( p_0 = 2 + \delta \), \( \delta > 0 \).

(ii) The mixing coefficient \( \alpha(\cdot) \) of the stationary \( \alpha \)-mixing process \( \{\varepsilon_t\} \) satisfies

\[
\alpha(t) \sim C_\alpha \theta_0^t, \quad \text{where} \quad 0 < C_\alpha < \infty \quad \text{and} \quad 0 < \theta_0 < 1.
\]

**Assumption 4**

(i) The bandwidths, \( h_j \), \( j = 1, \ldots, d \), satisfy \( h_j = c_j h \) for some positive constant \( c_j \), \( 1 \leq j \leq d \), and

\[
nh^{2\gamma} \rightarrow 0, \quad \frac{n^{\frac{\gamma}{2}} h}{\log n} \rightarrow \infty \quad \text{as} \quad n \rightarrow \infty.
\]
(ii) There exist two sequences of positive integers \( r_n \) and \( R_n \) such that
\[
 r_n \to \infty, \quad r_n = o\left(R_n \vee n^{1/2} h^{-\frac{p_0+2}{2p_0}}\right), \quad R_n \left(h + n^{2-p_0} h^{-1} \log n \right) \to 0.
\] (4.4)

Furthermore, let
\[
v_1(r_n) = O(h^2 \tau_n), \quad nh^{-\frac{p_0+2}{2p_0}} v_2(r_n) = o(1), \quad h^{-2}\left(v_2^{1/2}(r_n) + h^{-\frac{p_0-2}{2p_0}} v_1^{p_0}(r_n)\right) = o(1),
\] (4.5)

where \( \tau_n = \sqrt{\log \frac{n}{nh}} \) and \( v_\nu(\cdot) \) is defined in (4.2).

**Remark 4.1.** Assumption 1 is commonly used in the kernel estimation literature, see, for example, Wand and Jones (1995). Assumption 2 imposes some smoothness conditions on the density functions and regression functions. The compact support condition on \{X_t\} can be relaxed with the expense of more lengthy proofs. Assumption 3 provides the moment condition on \{Y_t\} as well as the mixing coefficient condition for \{e_t\}. Note that, in Assumption 3 (ii), we assume that \( \alpha \)-mixing coefficient decays at the geometric rate, which can be relaxed to the algebraic rate at the cost of more lengthy proofs and more complicated conditions on the bandwidth and stability coefficient. Assumption 4 gives some conditions on the bandwidths and stability coefficient of the NED process. In particular, the technical conditions in Assumption 4(ii) are similar to the corresponding conditions in Lu and Linton (2007) and Li et al (2012), and they can be satisfied by some interesting time series models under mild conditions. More discussion can be found in Section 4.1 of Lu and Linton (2007) and Remark 2.1 of Li et al (2012).

Before stating the main results, we need to introduce some notations. Define
\[
\eta_t = Y_t - \sum_{j=1}^{d} w_{o,j} m_j(X_{tj}), \quad \eta_{tj} = Y_t - E(Y_t | X_{tj}) = Y_t - m_j(X_{tj})
\]
and \( \beta_{jk}(X_{sk}) = E(m_j(X_{sk})|X_{sk}) \). Let \( \xi_t = (\xi_{t1}, \ldots, \xi_{td})^\top \) with
\[
\xi_{tj} = \eta_{tj} - \eta_{tj}^*, \quad \eta_{tj}^* = m_j(X_{tj}) \eta_t, \quad \eta_{tj}^* = \sum_{k=1}^{d} w_{o,k} \eta_k \beta_{jk}(X_{tk}).
\]

Define
\[
\Lambda = \begin{pmatrix}
E[m_1(X_{t1})m_1(X_{t1})] & \cdots & E[m_1(X_{t1})m_d(X_{td})] \\
\vdots & \ddots & \vdots \\
E[m_d(X_{td})m_1(X_{t1})] & \cdots & E[m_d(X_{td})m_d(X_{td})]
\end{pmatrix} \quad \text{and} \quad \Sigma = \sum_{t=-\infty}^{\infty} E[\xi_0 \xi_t^\top].
\]
We give the asymptotic distribution of $\hat{w}$ in the following theorem.

**Theorem 4.1.** Suppose that the assumptions 1–4 are satisfied and $\Lambda$ is positive definite. Then, we have

$$\sqrt{n}(\hat{w} - w^*_o) \xrightarrow{d} N(0, \Lambda^{-1}\Sigma \Lambda^{-1})$$

(4.6)

where $w^*_o = (w_{o,1}, \ldots, w_{o,d})^T$.

**Remark 4.2.** When the dimension $d$ is fixed, the above theorem shows that the parametric estimation of the optimal weights in the approximation of MAMAR can achieve the root-$n$ convergence rate although we replace $E(Y_t|X_{tj})$ by its nonparametric estimator.

Define

$$\Sigma_1(x) = ||K||_2^2 \times \text{diag}\left\{ \frac{c_1\sigma_1^2(x_1)}{f_1(x_1)}, \ldots, \frac{c_d\sigma_d^2(x_d)}{f_d(x_d)} \right\}$$

and $\sigma_j^2(x_j) = E[\eta_j^2|X_{tj} = x_j]$, where $||K||_2^2 = \int K^2(u)du$, $x = (x_1, \ldots, x_d)^T$ and $c_j$ is defined as in Assumption 4 (i). We next give the asymptotic distribution for $b_m(x)$.

**Theorem 4.2.** Suppose that the conditions in Theorem 4.1 are satisfied. Then, we have

$$\sqrt{n\hat{h}}(\hat{m}(x) - m_w(x)) \xrightarrow{d} N(0, \sigma_w^2(x))$$

(4.7)

where $m_w(x) = \sum_{j=1}^d w_{o,j}m_j(x_j)$ and $\sigma_w^2(x) = (w^*_o)^T\Sigma_1(x)w^*_o$.

**Remark 4.3.** Theorem 4.2 above shows that the proposed nonparametric estimator $\hat{m}(x)$ is asymptotically normal and enjoys the convergence rate of the nonparametric estimator in the standard univariate nonparametric regression centered around the approximating function $m_w(x)$. Furthermore, if $m_w(x) = m(x)$, i.e., the approximating model is correct, by (4.7), then

$$\sqrt{n\hat{h}}(\hat{m}(x) - m(x)) \xrightarrow{d} N(0, \sigma_w^2(x))$$

(4.8)

In that case, $\hat{m}(x)$ can be seen to be also to be an efficient estimator of $m(x)$ in the sense of Tibshirani (1984, Chapter 5). We could in principle test the specification $m_w(x)$ using for example the principles behind the Fan and Li (1996)'s test.

In the estimation methodology, we do not consider the detection of the relevant covariates and the inclusion of the irrelevant covariates may potentially affect the nonlinear time series forecasting issue discussed in Section 5.1. A possible way to address this issue is to first apply the data-driven cross-validation method introduced by Hall et al (2007) to choose the significant covariates and then use the proposed semiparametric method in both the nonlinear estimation and forecasting. However, in the nonlinear time series forecasting with very large lag terms studied in Sections 4.2 and 5.1 below, the implementation of the cross-validation method might be quite time-consuming, and thus needs to be modified or replaced by the penalized variable selection procedure.
4.2 Case II: The dimension of $X_t$ is diverging

In Theorem 4.1, we showed that our estimator $\hat{w}$ is a $\sqrt{n}$-consistent estimator of $w_0^*$ when the dimension $d$ is fixed.

We next consider the case that the dimension of $X_t$ increases with the sample size, which would potentially allow a large $d$ with applications in nonlinear forecasting involving very large lag terms. This section therefore considers the MAMAR approximation as stated in the previous sections but with $d$ replaced by $d_n$ which increases with the sample size $n$. In case of confusion, we let $\hat{w}(n)$ and $w_0^*(n)$ denote for $\hat{w}$ and $w_0^*$ with $d$ replaced by $d_n$. Further, define

$$E\left[ m_1(X_{t1})m_1(X_{t1}) \right] \cdots E\left[ m_1(X_{t1})m_{dn}(X_{tdn}) \right]$$

$$\cdots \cdots$$

$$E\left[ m_{dn}(X_{tdn})m_1(X_{t1}) \right] \cdots E\left[ m_{dn}(X_{tdn})m_{dn}(X_{tdn}) \right],$$

$\Sigma_n$ to be the $\Sigma$ with $d$ replaced by $d_n$, and $\Sigma_n(w) = \Lambda_n^{-1}\Sigma_n\Lambda_n^{-1}$. In order to establish asymptotic theory for this case, we also need the following regularity conditions in addition to Assumptions 1–4 in Section 4.1.

Assumption 5

(i) There exists a compact support $\Omega$ such that $\cup_j \Omega_j \subset \Omega$, where $\Omega_j$ is defined in Assumption 2 (i).

(ii) Let $h_j \equiv h$ for $j = 1, \ldots, d_n$, where $h$ satisfies the conditions in Assumption 4.

(iii) The largest and smallest eigenvalues of $\Sigma_n(w)$ are bounded away from zero and infinite.

(iv) The dimension of $X_t$, $d_n$, satisfies

$$d_n(\tau_n + h^\gamma) = o(1), \quad nd_nh^{2\gamma} = o(1), \quad nd_nh^{\frac{p_0+2}{p_0}}v_2(r_n) = o(1),$$

(4.9)

where $\tau_n$ is defined in Assumption 4 (ii).

Remark 4.4. The technical conditions in Assumption 5 (i) and (ii) are imposed to simplify the presentation of our theoretical results as well as the proofs. Assumption 5 (iii) is a commonly-used condition in high-dimensional statistical inference, see, for example, Fan and Peng (2004). Assumption 5 (iv) gives some restrictions on $d_n$, which however could increase with the sample size at some polynomial rate.

We next give the asymptotic normal distribution theory for $\hat{w}(n)$. 

14
As the dimension $d_n$ tends to infinity, we cannot state our asymptotic theory as done in Theorem 4.1. We will follow Fan and Peng (2004) and Lam and Fan (2008), and let $A_n$ be a $p \times d_n$ matrix with full row rank such that as $n \to \infty$

$$A_nA_n^T \to A^*,$$

where $A^*$ is a non-negative definite $p \times p$ matrix with $p$ being a fixed positive integer. In applications, if we are interested in the asymptotic behavior for the components of $\hat{w}(n)$, say the first $p$ weights, we may choose $A_n = [I_p, N_{p \times (d_n-p)}]$, where $I_p$ is a $p \times p$ identity matrix and $N_{k \times j}$ is a $k \times j$ null matrix. In particular, if $p = 1$, and $A_n$ is taken as a $1 \times d_n$ vector of components 0 except the $j$-th component 1, then the following theorem shows that the $j$-th component of $\hat{w}(n)$ is a consistent estimator of the $j$-th component of $w^*_o(n)$, for any $1 \leq j \leq d_n$.

**Theorem 4.3.** Suppose that the conditions of Theorem 4.1 and Assumption 5 are satisfied. Then, we have

$$\sqrt{n}A_n\Sigma_n^{-1/2}(w)[\hat{w}(n) - w^*_o(n)] \overset{d}{\to} N(0, A^*). \quad (4.10)$$

**Remark 4.5.** The above theorem indicates that when $d_n \to \infty$, the parametric estimator of the optimal weight $w^*_o(n)$ can achieve the root-$(n/d_n)$ convergence rate. That is $\|\hat{w}(n) - w^*_o(n)\| = O_P(\sqrt{d_n/n})$. Such a result is analogous to some existing results in the literature of statistics such as Theorems 1 and 2 in Lam and Fan (2008) with independent observations. By using Theorem 4.3, we can also prove the asymptotic distribution for nonparametric estimator as in Theorem 4.2, by assuming that $\sum_{j=1}^{d_n} w^*_o \sigma^2_j(x_j)/f_j(x_j)$ converges to a finite limit. An issue here is that the dimensions of the argument $x = (x_1, \ldots, x_{d_n})$ increase with sample size $n$. This result may be more naturally discussed in the context of forecasting, which we will consider in the next section.

5 Some Extensions

In this section, we provide the discussions on some relevant topics including the nonlinear forecasting issue in time series analysis, and applications to limited dependent variables and robust methods.
5.1 Nonlinear forecasting

We next study the problem of forecasting a future value \( Y_{T+k_0} \) by using given observations \( \{Y_t : 1 \leq t \leq T\} \), where \( k_0 \) is a fixed number. Consider the forecast of the form defined by

\[
Y_T + k_0 | T = w_0 + \sum_{j=1}^{d_T} w_j E(Y_{T+k_0} | Y_{T+1-j}). \tag{5.1}
\]

Note that for a stationary process, \( E(Y_{T+k_0} | Y_{T+1-j} = y) = E(Y_t | Y_{t+1-j-k_0} = y) \), which can be estimated using \( \{Y_t : 1 \leq t \leq T\} \), so long as \( d_T + k_0 \ll T \).

We next discuss different choices of \( d_T \) and \( w \) in time series forecasting. First, we consider an unstructured model where the weights are chosen by the predictive least squares unrestrictedly and the lag horizon \( d_T \) is fixed. For this case, we can use the semiparametric method introduced in Section 3 to estimate the optimal weights, and then get \( \hat{Y}_{T+k_0} | T \), the predicted value of \( Y_{T+k_0} | T \), by replacing \( E(Y_t | Y_{t+1-j-k_0} = y) \) by its corresponding nonparametric estimated value.

For the more general case that the lag horizon \( d_T \to \infty \) slowly (recommend taking \( d_T = c \log T \) for some constant \( c \)), we can still use the semiparametric method developed in this paper to predict \( Y_{T+k} \). For the case of \( d_T = c \log T \), by using Theorem 4.3, we can prove that

\[
\sqrt{Th/\sigma_w^2(T)}(\hat{Y}_{T+k_0} | T - \hat{Y}_{T+k_0} | T) \xrightarrow{d} N(0,1), \tag{5.2}
\]

where \( \sigma_w^2(T) \) is defined as \( \sigma_w^2 \) in Theorem 4.2 with \( d \) and \( (x_1, \ldots, x_d) \) replaced by \( d_T \) and \( (Y_T, \ldots, Y_{T-d_T+1}) \), respectively. We find that the convergence rate for the nonparametric predicted value can achieve root-(\( Th \)) if \( \sigma_w^2(T) \) tends to a positive constant as \( T \) tends to infinity, although \( c \log T \) lags are involved in forecasting.

Notice that the above discussion ignores the fact that the importance of more distant lags should be much less than those of more recent ones in time series analysis. That is the weights \( w_j \) should decay to zero as \( j \to \infty \). Hence, for a small enough \( \epsilon > 0 \), we can always find a positive integer \( d := d(\epsilon) \) such that \( |w_j| < \epsilon \) when \( j \geq d \), which implies that the weights \( w_j \) can be ignored when \( j \geq d \). There are a number of ways of imposing the decay through parameterizations. For example, we can impose a polynomial function on \( w_j \)

\[
w_j = \alpha_0 + \alpha_1 j^{-1} + \ldots + \alpha_k j^{-k}, \tag{5.3}
\]

where \( \alpha = (\alpha_0, \ldots, \alpha_k) \) are free parameters. Another popular approach is based on the ARMA class of time series models. Define the lag polynomials \( A(L) = \sum_{j=0}^{p} a_j L^j \) and \( B(L) = \sum_{j=0}^{q} b_j L^j \),
where $L$ is a lag operator, and let the $w_j$ solve the equation

$$
\sum_{j=0}^{\infty} w_j L^j = \frac{A(L)}{B(L)},
$$

which would happen uniquely under some conditions. Then the weights $w_j$, $j = 1, 2, \ldots$, just depend on the parameter vector $\theta = (a_0, \ldots, a_p, b_0, \ldots, b_q)^T$. The problem then reduces to choosing $\theta$ to minimize the sample least squares problem using some truncation (truncation may not be strictly necessary here since provided the parameters are inside the usual “stationary invertible” region, the weights should decay at some geometric rate).

For the case that the weights $w_j$ decay irregularly to zero, we may need to combine our semiparametric estimation method with some variable selection approach (such as the LASSO and SCAD penalty functions mentioned in Section 1 and the independence screening technique introduced by Fan and Lv (2008)) to choose the significant weights (or lags), and then study the nonlinear forecasting using the significant lags. An advantage of the variable selection methods lies in that they may force sufficiently small weights to be zeros and thus screen out the irrelevant lags (and irrelevant exogenous covariates), crucial for us to determine the dimension $d$ ($d_n$ or $d_T$) for the complex and high-dimensional time series analysis. This is a non-trivial variable selection issue beyond the scope of this paper, left for a future work.

### 5.2 Limited dependent variables and robust methods

As mentioned in Section 2 briefly, consider the more general setting

$$
E(Y|X) = C(E(Y|X_1), \ldots, E(Y|X_d); w),
$$

where $w$ is a parameter vector and $C$ is a regression copula. The connection with copulas can be made firmer in the case where the marginal regressions are monotonic in which case any regression function (with monotonic marginals) can be represented as (5.5) for some copula function $C$. The generalization encoded in (5.5) might be useful in the case of limited dependent variables. Suppose that $Y$ is binary but $X$ is continuously distributed. A common parametric model here would be probit or logit, where $P(Y = 1|X = x) = E(Y|X = x) = F(\beta_0 + \sum_{j=1}^{d} \beta_j x_j)$, where $F$ is the normal or logistic cumulative distribution function (c.d.f.). Therefore, consider

$$
P(Y = 1|X = x) = E(Y|X = x) = F\left(w_0 + \sum_{j=1}^{d} w_j F^{-1}(E(Y|X_j = x_j))\right),
$$
where $F$ is a known c.d.f. We may compare this with the generalized additive modelling approach, see Hastie and Tibshirani (1990) in which the marginal regression functions $E(Y|X_j = x_j)$ are replaced by free functions $m_j(x_j)$ and the weights $\{w_j\}$ are not needed. Estimation of this model can be carried out using quasi-likelihood, which entails a nonlinear optimization over $w$, albeit one that can be coded as iterative weighted least squares.

Finally, we could also allow the conditional expectations operator to be replaced by conditional quantiles. That is, we consider the model (or approximation)

$$Q_\alpha(Y|X) = w_0 + \sum_{j=1}^d w_j Q_\alpha(Y|X_j = x_j),$$

where $Q_\alpha$ denotes the level $\alpha$ conditional quantile function. This model can be estimated by linear quantile regression where the marginal quantile regressions are the covariates and so computationally this is also relatively simple.

## 6 Numerical Evidence

In this section, we are demonstrating certain advantages of the proposed nonlinear MAMAR method by a Monte Carlo simulation and a real data example, to uncover and understand the time series lag effects in applications of forecasting. Monte Carlo simulation example is provided in the first subsection, and the second one is the analysis of the high frequency (one minute) data from the FTSE100 index. The supplemental document gives another application of our method in analysis of the Australian annual mean temperature anomaly series.

### 6.1 A simulation study

In order to understand the performance of our proposed nonlinear MAMAR (simply denoted as NMA below) in comparison with other forecasting modelings such as linear modeling and additive modeling in different settings of real models, we consider the following model in our simulation study:

$$Y_t = \sum_{k=1}^9 g_{0k}(Y_{t-k}) + \varepsilon_t$$

$$g_{0k}(Y_{t-k}) = a_k Y_{t-k} + \delta \frac{\exp(-kY_{t-k})}{1 + \exp(-kY_{t-k})} + \gamma \cos(Y_{t-k}Y_{t-1}),$$

(6.1)
where $\varepsilon_t \sim$ i.i.d. $\mathcal{N}(0, \sigma^2)$, while the values of $\sigma^2$ and $a_k$’s, for $k = 1, 2, \ldots, 9$, are specified in Table 1, which are actually the estimated values of a linear AR(9) model using the whole time series data set AMTA.res in the supplemental document. Let $\delta$ and $\gamma$ be two constants for which we consider three cases of $\delta = 0$, $\delta = 0.1$ and $\delta = 0.5$, and three cases of $\gamma = 0$, $\gamma = 0.1$ and $\gamma = 0.5$. Although we can construct a more involved GARCH structure for $\varepsilon_t$ such as $\varepsilon_t = e_t \sigma_t$ with $\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \alpha_2 \varepsilon_{t-2}^2 + \beta_1 \sigma_{t-1}^2$ and $e_t$ i.i.d. $\mathcal{N}(0, 1)$, where $\alpha_0 > 0$, $\alpha_1$, $\alpha_2$ and $\beta_1$ are suitable non-negative constants, so that $Y_t$ is a NED process with respect to an $\alpha$-mixing process $\{\varepsilon_t\}$, we do not pursue this for simplicity. Note that under the conditions specified as above, it follows from Lu (1998) that the model has a strictly stationary solution which is geometrically ergodic and thus $\alpha$-mixing, a special NED process defined in Definition 4.1. Further, note that in model (6.1), as $(\delta, \gamma) = (0, 0)$, this is a linear autoregressive model of order 9; while as $\delta \neq 0$ but $\gamma = 0$, it is a nonlinear additive autoregressive model of order 9; and as $\gamma \neq 0$, this is a nonlinear autoregressive model of order 9 with interaction between $Y_{t-k}$ and $Y_{t-1}$.

Table 1: The noise variance $\sigma^2$ and the coefficients $a_k$’s in the model (6.1).

<table>
<thead>
<tr>
<th>$\sigma^2$</th>
<th>$a_1$</th>
<th>$a_2$</th>
<th>$a_3$</th>
<th>$a_4$</th>
<th>$a_5$</th>
<th>$a_6$</th>
<th>$a_7$</th>
<th>$a_8$</th>
<th>$a_9$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.08498</td>
<td>-0.1129</td>
<td>0.0245</td>
<td>-0.1892</td>
<td>-0.0820</td>
<td>-0.1962</td>
<td>-0.1232</td>
<td>0.1180</td>
<td>0.1282</td>
<td>-0.2407</td>
</tr>
</tbody>
</table>

We simulate the stationary time series data of size $n$ from model (6.1) by deleting the first 100 observations among the $(100 + n)$ observations generated through iteration of (6.1) with initial values of $Y_1 = \cdots = Y_9 = 0$ for each given pair of $(\delta, \gamma)$ with $\delta = 0$, $0.1$, $0.5$ and $\gamma = 0$, $0.1$, $0.5$. We partition the whole sample of size $n$ into two parts: the first part is an estimation sample of size $n_{\text{est}} = n - n_{\text{pred}}$ for model estimation, and the second part is a prediction sample of size $n_{\text{pred}} = 50$ for evaluation of different prediction methods. We are considering three cases of $n = 140$, $n = 200$ and $n = 250$ so that $n_{\text{est}} = 90$, $n_{\text{est}} = 150$, $n_{\text{est}} = 200$, respectively. Three prediction methods are compared: the linear AR model of order 9, the nonlinear MAMAR of lag 9 proposed in this paper, and the nonlinear additive AR model of order 9. The estimation procedures for the linear AR model and nonlinear additive AR model are based on the arima with “ML” method and the gam with smoothing splines, in the R packages STATS and GAM, respectively. As one referee suggested, for the additive modelling, we also considered the smooth backfitting method due to Mammen et al (1999) and Nielsen and Sperlich (2005) with the R package sBF. The estimation procedure for the
Figure 1: Simulation — Boxplot of 100 repetitions of the MSPE of one-step-ahead predictions for $\gamma = 0$: (a) $n_{\text{est}} = 90$, (b) $n_{\text{est}} = 150$, (c) $n_{\text{est}} = 200$. Here “AR” and “NMA” stand for the prediction based on the linear AR and nonlinear MAMAR, respectively, and “GAM” and “sBF” are based on additive AR modelling with the R packages of GAM (generalised additive model) and sBF (smooth backfitting), respectively.
Figure 2: Simulation — Boxplot of 100 repetitions of the MSPE of one-step-ahead predictions for $\gamma = 0.1$: (a) $n_{\text{est}} = 90$, (b) $n_{\text{est}} = 150$, (c) $n_{\text{est}} = 200$. Here “AR” and “NMA” stand for the prediction based on the linear AR and nonlinear MAMAR, respectively, and “GAM” and “sBF” are based on additive AR modelling with the R packages of GAM (generalised additive model) and sBF (smooth backfitting), respectively.
Figure 3: Simulation — Boxplot of 100 repetitions of the MSPE of one-step-ahead predictions for $\gamma = 0.5$: (a) $n_{est} = 90$, (b) $n_{est} = 150$, (c) $n_{est} = 200$. Here “AR” and “NMA” stand for the prediction based on the linear AR and nonlinear MAMAR, respectively, and “GAM” and “sBF” are based on additive AR modelling with the R packages of GAM (generalised additive model) and sBF (smooth backfitting), respectively.
nonlinear MAMAR approximation is based on that in Section 3 of this paper. We are examining the one-step-ahead prediction of $Y_{n_{est}+i}$, say $\hat{Y}_{n_{est}+i}$, for $i = 1, 2, \ldots, n_{pred}$, based on the estimated models, and consider the mean squared prediction error, defined by

$$\text{MSPE} = \frac{1}{n_{pred}} \sum_{i=1}^{n_{pred}} \left( Y_{n_{est}+i} - \hat{Y}_{n_{est}+i} \right)^2.$$ 

We repeat the simulation for 100 times, and the boxplots of the 100 MSPE values for three different methods with different pairs of $(\delta, \gamma)$ are depicted in Figures 1–3, corresponding to $\gamma = 0, 0.1, 0.5$, respectively. In each figure, there are three panels (a), (b) and (c), corresponding to $n_{est} = 90, 150, 200$, respectively, and in each panel, there are three sub-panels of boxplots, corresponding to $\delta = 0, 0.1, 0.5$, respectively, where “AR” and “NMA” stand for the methods of prediction based on the linear AR and nonlinear MAMAR, respectively, and “GAM” and “sBF” are for additive AR modelling with the R packages of GAM and sBF, respectively.

We first have a look at Figure 1 for $\gamma = 0$, where model (6.1) is a purely additive AR model of order 9. In the case of $\delta = 0$, model (6.1) further reduces to a linear AR model, where the linear “AR” method should perform the best in prediction, as confirmed in the first column of Figure 1. In this case, both “NMA” and “GAM” also appear quite acceptable with small values of MSPE and perform similarly in prediction, where for the small estimation sample size, “NMA” appears a bit better than “GAM”, while with larger estimation sample size, “GAM” is slightly better than “NMA”. In the case of $\delta \neq 0$, model (6.1) is a purely nonlinear additive model, where it follows from the second and third columns of Figure 1 that the linear “AR” method is much worse than both “NMA” and “GAM”, both of which perform again quite similarly in prediction although “GAM” is slightly better with the estimation sample size increasing. In this case, the performance of “sBF” is also similar to that of “GAM” in prediction but a bit inferior.

We next turn to Figures 2 and 3 for $\gamma \neq 0$, where model (6.1) is a nonlinear, but not purely additive, AR model of order 9 with interaction between $Y_{t-k}$ and $Y_{t-1}$. Both figures indicate that the linear “AR” is very poor. For the case of small value of $\gamma = 0.1$, where model (6.1) is close to a purely additive AR model, the performance of both “NMA” and “GAM” in Figure 2 looks similar to that in Figure 1, that is both “NMA” and “GAM” perform very similarly in prediction with “GAM” slightly better as the estimation sample size increases, while “sBF” appears poorer in prediction with larger $\delta$. However, as $\gamma = 0.5$, model (6.1) is far away from a purely additive AR model, and it clearly follows from Figure 3 that both “GAM” and “sBF” perform basically similarly but worse than our “NMA” method in prediction in particular with smaller sample size. Again in the case of the sample size as small as 90, “GAM” is better than “sBF”. Here it is worth noting that all four
methods look poor in prediction with much larger values of MSPE than those in Figures 1 and 2. This somehow indicates that the interaction between different lags should be taken into account in the prediction. In theory, this interaction can be much more easily incorporated into the prediction method proposed in our paper than that in additive models. However, practically, we need to deal with the selection of interactions among a large number of lag interactions. For example, there are 9 lags in model (6.1) and hence the number of lag interactions is 36 in total, which requires model selection techniques in prediction in particular when the estimation sample size is not that large, such as $n_{est} = 90$. We leave this for future research.

In summary, our proposed method performs quite well in prediction. When the actual model is a purely additive model, it performs quite close to the optimal additive prediction with either “GAM” or “sBF”. While a purely additive modelling structure is violated, it may even be better than the additive in prediction. The main advantage of our method is computational, in particular perhaps performance in the case where $d$ is large (see Section 6.2 below).

6.2 An application: Forecasting of the volatility with FTSE100 index

In this section, we are illustrating an application of the proposed nonlinear MAMAR to forecasting of the volatility with a high frequency (one minute) financial data set from the FTSE100 index. The data set consists of trading volume $v_t$, open price $o_t$, close price $c_t$, minimum price $\min_t$, and maximum price $\max_t$, of the index in each minute. The time period is the minutes within the trading days: 22–28 June 2012, of 2000 observations. We are concerned with the relationship between the volatility and the geometric return, defined, respectively, by

$$V_t = 100(\max_t - \min_t)/((\max_t + \min_t)/2),$$

and

$$R_t = 100\log(c_t/c_{t-1}),$$

as well as the volume series $v_t$. The three series are depicted in Figure 4.

In this example, we are interested in the $k$-step-ahead prediction of the volatility by using the information of the long lags (from lag 1 to lag 60 in view of the minute data within one hour of lags) of both volatility and return series, and also checking if the volume lags would be helpful in improving the prediction of the volatility. That is, we are using $X_t = (V_{t-1}, \cdots, V_{t-60}, R_{t-1}, \cdots, R_{t-60})^\top$ or $X_t = (V_{t-1}, \cdots, V_{t-60}, R_{t-1}, \cdots, R_{t-60}, v_{t-1}, \cdots, v_{t-60})^\top$ to predict $Y_t = V_{t+k}$, for $k = 1, 2, \cdots, 10$. In such high-dimensional case, how to select optimal lags is an interesting issue beyond the scope
Figure 4: The time series plots of the volume $v_t$, the geometric return $R_t$ and the volatility $V_t$. 
of this paper (c.f., the discussion in Subsection 5.1). We are only looking at the comparison of the forecasting based on the nonlinear MAMAR approximation with the linear AR forecasting as it is nearly impossible to do with additive modelling for such a high-dimensional case. We use the sample from the $M = (70 + k)$th observation to the $N = 1900$th observation as our estimation sample. Our evaluation sample of the prediction is the following $n_{pre} = 60$ observations right after the estimation sample. In order to avoid the serious impact of the extreme return 0.5132315 of the 1262th observation (see Figure 4) on the estimation of our model parameters, we tentatively delete it in our estimation step. We calculated MSPE of the $k$-step-ahead prediction of the volatility for the nonlinear MAMAR and the linear AR forecasting, respectively, for $k$ from 1 to 10, which are plotted in Figure 5, where (a) corresponds to the case of $X_t = (V_{t-1}, \ldots, V_{t-60}, R_{t-1}, \ldots, R_{t-60})^\top$, and (b) to $X_t = (V_{t-1}, \ldots, V_{t-60}, R_{t-1}, \ldots, R_{t-60}, v_{t-1}, \ldots, v_{t-60})^\top$. Clearly, the proposed model averaging method is overall preferred to the linear forecasting with smaller MSPEs except the $k = 7$ step ahead forecasting. In addition, comparing (a) with (b), it appears that the volume lags contribute little to the prediction of the volatility.

7 Conclusion

In this paper, we have proposed approximating a multivariate regression function by an affine combination of one-dimensional conditional component regression functions. A semiparametric method with the first-stage nonparametric kernel smoothing has been developed to estimate the weight parameters involved in the MAMAR approximation. Asymptotic properties for both the parametric and nonparametric estimators have been established under mild conditions. In particular, the parametric estimator is shown to be asymptotically normal with root-$n$ rate of convergence when the dimension of the covariates is fixed, and there is no curse of dimensionality for the nonparametric estimator. When the dimension increases with the sample size, the parametric estimator is shown to be asymptotically normal with root-$(n/d_n)$ rate of convergence. The observations in this paper are assumed to be stationary and near epoch dependent, which is very general and covers some popular time series models such as AR($p$)-GARCH(1,1) model. Hence, the developed approach is applicable to the nonlinear forecasting issues in time series analysis. Our methods and results are further augmented by a simulation study and an empirical application.
Figure 5: Comparison of the mean squares prediction error (MSPE) of the $k$-step-ahead prediction of the volatility against $k$ for the nonlinear MAMAR (NMA) and the linear AR forecasting methods: (a) Case of no lagged volumes used, (b) Case of the lagged volumes used.
8 Acknowledgements

The authors are grateful to the Co-Editor Professor Jianqing Fan, the Associate Editor and three anonymous referees for their valuable and constructive comments and suggestions. Thanks also go to the colleagues who commented on this paper and the participants of various conferences and seminars where this work was presented. The second author acknowledges the financial support from the ERC and the third author acknowledges the partial financial supports from the ARC Future Fellowship Scheme and Discovery Project (when he worked in Australia) and the European Commission’s Marie Curie Career Integration Grant (when in the UK).

Appendix: Proofs of the Main Results

In this appendix, we give the proofs of the asymptotic results stated in Sections 4. Some technical lemmas with proofs are provided in the supplemental document. In the sequel, $C$ denotes a positive constant, whose value may change from line to line.

Proof of Theorem 4.1. Recall that $\eta_t = Y_t - \sum_{j=1}^{d} w_{o,j} m_j(X_{tj})$ and $\eta_{tj} = Y_t - E(Y_t|X_{tj}) = Y_t - m_j(X_{tj})$. Furthermore, define

$$\mathcal{M} = \begin{bmatrix} m_1(X_{11}) & \cdots & m_d(X_{1d}) \\ \vdots & \ddots & \vdots \\ m_1(X_{n1}) & \cdots & m_d(X_{nd}) \end{bmatrix}, \quad \eta = (\eta_1, \ldots, \eta_n)^\top.$$ 

Observe that

$$\hat{w} = (\hat{\mathcal{M}}^\top \hat{\mathcal{M}})^{-1} \hat{\mathcal{M}}^\top \mathcal{Y}$$

$$= (\hat{\mathcal{M}}^\top \hat{\mathcal{M}})^{-1} \hat{\mathcal{M}}^\top (\mathcal{M}w_o^* + \eta)$$

$$= w_o^* + (\hat{\mathcal{M}}^\top \hat{\mathcal{M}})^{-1} \hat{\mathcal{M}}^\top (\mathcal{M} - \hat{\mathcal{M}})w_o^* + (\hat{\mathcal{M}}^\top \hat{\mathcal{M}})^{-1} \hat{\mathcal{M}}^\top \eta$$

$$=: w_o^* + \Pi_{n1} + \Pi_{n2}.$$

We first derive the leading term of $\Pi_{n1}$. Note that, for each $1 \leq j \leq d$,

$$\hat{m}_j(x_j) - m_j(x_j) = \frac{\sum_{t=1}^{n} \eta_{tj} K\left(\frac{X_{tj} - x_j}{h_j}\right)}{\sum_{t=1}^{n} K\left(\frac{X_{tj} - x_j}{h_j}\right)} + \frac{\sum_{t=1}^{n} m_j(X_{tj}) K\left(\frac{X_{tj} - x_j}{h_j}\right)}{\sum_{t=1}^{n} K\left(\frac{X_{tj} - x_j}{h_j}\right)} - m_j(x_j).$$
By Taylor’s expansion and Lemma B.1 in the supplemental document, uniformly for \( x_j \in \Omega_j \), we have

\[
\sum_{t=1}^{n} m_j(X_{tj})K\left(\frac{X_{tj} - x_j}{h_j}\right) - m_j(x_j) \overset{P}{\sim} \frac{(m * f_j^{(\gamma)}(x_j))}{f_j(x_j)} \mu_\gamma h_j^\gamma,
\]  

(A.1)

where \( a_n \overset{P}{\sim} b_n \) means that \( a_n/b_n = 1 + o_P(1) \), \((m * f_j^{(\gamma)}(x_j))\) is the \( \gamma \)-th derivative of \( m_j(z)f_j(z) \) which exits by Assumption 2 (i) and (iv), and \( \mu_\gamma = \int u^\gamma K(u)du \). By (A.1), we have, uniformly for \( x_j \in \Omega_j \),

\[
\hat{m}_j(x_j) - m_j(x_j) = \frac{\sum_{t=1}^{n} \eta_{tj}K\left(\frac{X_{tj} - x_j}{h_j}\right)}{\sum_{t=1}^{n} K\left(\frac{X_{tj} - x_j}{h_j}\right)} + \frac{m * f_j^{(\gamma)}(x_j)h_j^\gamma}{f_j(x_j)}(1 + o_P(1)).
\]  

(A.2)

On the other hand, by Lemma B.1 again, we can also prove

\[
\frac{1}{n} \mathcal{M}' \mathcal{M} \overset{P}{\sim} \frac{1}{n} \mathcal{M}' \mathcal{M} \overset{P}{\sim} \Lambda,
\]  

(A.3)

where \( \Lambda \) is assumed to be positive definite in Theorem 4.1. Then, by (A.1)–(A.3), we have

\[
\Pi_{n1} \overset{P}{\sim} (\mathcal{M}' \mathcal{M})^{-1} \Pi_{n3},
\]  

(A.4)

where

\[
\Pi_{n3} = \left\{ \sum_{t=1}^{n} m_j(X_{tj}) \sum_{k=1}^{d} w_{a,k} \left[ m_k(X_{tk}) - \hat{m}_k(X_{tk}) \right] \right\}^\top_{j=1,\ldots,d}
\]

\[
= \left\{ \sum_{t=1}^{n} m_j(X_{tj}) \sum_{k=1}^{d} w_{a,k} \left[ -\sum_{s=1}^{n} \eta_{sk}K\left(\frac{X_{sk} - X_{tj}}{h_k}\right) + O_P(h_k^\gamma) \right] \right\}^\top_{j=1,\ldots,d}
\]

\[
= \left\{ -\sum_{t=1}^{n} m_j(X_{tj}) \sum_{k=1}^{d} w_{a,k} \left[ \sum_{s=1}^{n} \eta_{sk}K\left(\frac{X_{sk} - X_{tj}}{h_k}\right) \right] \right\}^\top_{j=1,\ldots,d} + O_P(nh^\gamma)
\]

\[
= \left\{ -\sum_{s=1}^{d} \sum_{k=1}^{d} w_{a,k} \eta_{sk} \left[ \frac{1}{nh_k} \sum_{t=1}^{n} m_j(X_{tj})f_j^{-1}(X_{tj})K\left(\frac{X_{sk} - X_{tj}}{h_k}\right) \right] \right\}^\top_{j=1,\ldots,d} + O_P(nh^\gamma)
\]

where \( f_k(\cdot) \) is the marginal density function of \( X_{tk} \) and \( h \) is defined in Assumption 4 (i). If \( k = j \), by Lemma B.1, we have

\[
\frac{1}{nh_j} \sum_{t=1}^{n} m_j(X_{tj})f_j^{-1}(X_{tj})K\left(\frac{X_{sj} - X_{tj}}{h_j}\right) = m_j(X_{sj}) + o_P(1).
\]  

(A.5)
If \( k \neq j \), by Lemma B.1 again, we have

\[
\frac{1}{nh_k} \sum_{t=1}^{n} m_j(X_{tk}) f_k^{-1}(X_{tk}) K\left( \frac{X_{sk} - X_{tk}}{h_k} \right) = \beta_{jk}(X_{sk}) + o_P(1), \tag{A.6}
\]

where \( \beta_{jk}(X_{sk}) = \mathbb{E}(m_j(X_{sj})|X_{sk}) \). Then, by (A.5) and (A.6) and noting that \( \beta_{jj}(X_{sj}) = m_j(X_{sj}) \), we have

\[
\Pi_{n3} = -\left( \sum_{s=1}^{n} \sum_{k=1}^{d} w_{o,k} \eta_{sk} \beta_{jk}(X_{sk}) \right)^\top_{j=1,\ldots,d} + O_P(nh^\gamma)
\]

\[
= -\left( \sum_{t=1}^{n} \eta_{t1}^*, \ldots, \sum_{t=1}^{n} \eta_{td}^* \right)^\top + O_P(nh^\gamma), \tag{A.7}
\]

where \( \eta_{tj}^* = \sum_{k=1}^{d} w_{o,k} \eta_{tk} \beta_{jk}(X_{tk}) \). By (A.4) and (A.7), we have

\[
\Pi_{n1} = -(M^\top M)^{-1} \left( \sum_{t=1}^{n} \eta_{t1}^*, \ldots, \sum_{t=1}^{n} \eta_{td}^* \right)^\top + O_P(h^\gamma). \tag{A.8}
\]

We next consider \( \Pi_{n2} \). Observe that

\[
\Pi_{n2} = (M^\top \hat{\mathcal{M}})^{-1} \hat{\mathcal{M}}^\top \eta
\]

\[
= (M^\top M)^{-1} \hat{\mathcal{M}}^\top \eta (1 + o_P(1))
\]

\[
= (M^\top M)^{-1} [M^\top \eta + (\hat{\mathcal{M}} - M)^\top \eta] (1 + o_P(1)).
\]

Using (A.4) and Lemma B.3 in the supplemental document, we can show that the leading term of \( \Pi_{n2} \) is \((M^\top M)^{-1} M^\top \eta\) by noting that \( nh^{2\gamma} = o(1) \) and \( nh^{-\frac{p_0^2}{2k_0} v_2(r_n)} \to 0 \) and taking \( M_n = o(\sqrt{nh}) \). By letting \( \eta_{tj}^* = m_j(X_{tj}) \eta_t \), we have

\[
\Pi_{n2} = (M^\top M)^{-1} \left( \sum_{t=1}^{n} \eta_{t1}^*, \ldots, \sum_{t=1}^{n} \eta_{td}^* \right)^\top (1 + o_P(1)). \tag{A.9}
\]

Then, by (A.8), (A.9) and Lemma B.4 in the supplemental document, we can prove (4.6). Hence, the proof of Theorem 4.1 has been completed.
Proof of Theorem 4.2. Observe that
\[
\hat{m}_w(x) - m_w(x) = \sum_{j=1}^{d} \hat{w}_{o,j} \hat{m}_j(x_j) - \sum_{j=1}^{d} w_{o,j} m_j(x_j)
\]
\[
= \left[ \sum_{j=1}^{d} \hat{w}_{o,j} \hat{m}_j(x_j) - \sum_{j=1}^{d} w_{o,j} m_j(x_j) \right] + \left[ \sum_{j=1}^{d} w_{o,j} \hat{m}_j(x_j) - \sum_{j=1}^{d} w_{o,j} m_j(x_j) \right]
\]
\[=: \Pi_{n4} + \Pi_{n5}. \tag{A.10} \]

By Theorem 4.1, we can prove that
\[
\Pi_{n4} = O_P(\sqrt{n}) = o_P(\sqrt{nh}). \tag{A.11} \]

By the Cramér-Wold device (Billingsley 1968) and following the proof of Theorem 3.1 in Lu and Linton (2007), we can also prove that
\[
\sqrt{nh}\Pi_{n5} \xrightarrow{d} N\left(0, \sigma_w^2 \right). \tag{A.12} \]

Equations (A.10)–(A.12) imply that (4.7) holds. The proof of Theorem 4.2 is completed.

Proof of Theorem 4.3. The proof is similar to the proof of Theorem 4.1 with some modifications. Let \( \eta \) and \( \eta_j, 1 \leq j \leq d_n \), be defined as in the proof of Theorem 4.1 by replacing \( d \) by \( d_n \), and \( \eta = (\eta_1, \ldots, \eta_n)^t \).

Define the \( n \times d_n \) matrices by
\[
\mathcal{M}_n = \begin{bmatrix} m_1(X_{11}) & \cdots & m_{d_n}(X_{1d_n}) \\ \vdots & \vdots & \vdots \\ m_1(X_{n1}) & \cdots & m_{d_n}(X_{nd_n}) \end{bmatrix}, \quad \hat{\mathcal{M}}_n = \begin{bmatrix} \hat{m}_1(X_{11}) & \cdots & \hat{m}_{d_n}(X_{1d_n}) \\ \vdots & \vdots & \vdots \\ \hat{m}_1(X_{n1}) & \cdots & \hat{m}_{d_n}(X_{nd_n}) \end{bmatrix}.
\]

Letting \( \hat{\omega}(n) \) and \( \omega_o^*(n) \) be defined as \( \hat{\omega} \) and \( \omega_o^* \) with \( d \) replaced by \( d_n \), we can easily show that
\[
\mathcal{A}_n \Sigma_n^{-1/2}(w) [\hat{\omega}(n) - \omega_o^*(n)]
\]
\[
= \mathcal{A}_n \Sigma_n^{-1/2}(w) \left( \hat{\mathcal{M}}_n - \mathcal{M}_n \right)^{-1} \hat{\mathcal{M}}_n (\mathcal{M}_n - \hat{\mathcal{M}}_n) w_o^*(n) + \mathcal{A}_n \Sigma_n^{-1/2}(w) \left( \hat{\mathcal{M}}_n - \mathcal{M}_n \right)^{-1} \hat{\mathcal{M}}_n \eta
\]
\[=: \Pi_{n6} + \Pi_{n7}, \]

where, as in Theorem 4.3, \( \mathcal{A}_n \) is a \( p \times d_n \) matrix such that \( \mathcal{A}_n \mathcal{A}_n^t \) tends to a \( p \times p \) nonnegative matrix \( \mathcal{A}_* \), and \( \Sigma_n(w) = \Lambda_n^{-1} \Sigma_n \Lambda_n^{-1} \), \( \Lambda_n \) and \( \Sigma_n \) are defined as \( \Lambda \) and \( \Sigma \) with \( d \) replaced by \( d_n \).
We first derive the leading term of $\Pi_{n6}$. Using (A.1) in the proof of Theorem 4.1, we have, uniformly for $x \in \Omega$ and $1 \leq j \leq d_n$,

$$
\hat{m}_j(x) - m_j(x) = \frac{1}{n} \sum_{t=1}^{n} \eta_{tj} K\left(\frac{X_{tj} - x}{h}\right) + \frac{(m * f)^{(\gamma)}(x)}{f_j(x)} \mu_{\gamma} h^{\gamma}(1 + o_P(1)). \tag{A.13}
$$

Note that

$$
\frac{1}{n} \sum_{t=1}^{n} K\left(\frac{X_{tj} - x}{h}\right) = m_j(x),
$$

which implies that the smallest eigenvalue of $\frac{1}{n} M_n^\top M_n$ is larger than $\frac{1}{2} \lambda_{\min}$ in probability, where $\lambda_{\min} > 0$ is the smallest eigenvalue of $\Lambda_n$. As $d_n (\tau_n + h^\gamma) = o(1)$, we can show that the maximum eigenvalues (in absolute value) for the last three matrices (divided by $n$) on the right hand side of (A.14) tends to zero. Hence, $M_n^\top M_n$ is leading term of $\hat{M}_n^\top \hat{M}_n$, which leads to

$$
\frac{1}{n} \hat{M}_n^\top \hat{M}_n \sim \Lambda_n. \tag{A.15}
$$

The above result can be seen as an extension of (A.3). We thus have

$$
\Pi_{n6} \sim A_n \Sigma_n^{-1/2}(w) \left(M_n^\top M_n\right)^{-1} \Pi_{n8}, \tag{A.17}
$$

where $\Pi_{n8}$ is defined as $\Pi_{n3}$ with $d$ replaced by $d_n$. Then, by (A.5) and (A.6), similar to the proof of (A.7), we can prove that

$$
\Pi_{n8} = - \left[ \sum_{t=1}^{n} \eta_{t1}^*, \ldots, \sum_{t=1}^{n} \eta_{td_n}^* \right]^\top + O_P(nd_nh^\gamma), \tag{A.18}
$$

where $\eta_{tj}^* = \sum_{k=1}^{d_n} w_{n,k} \eta_{tk} \beta_{jk}(X_{tk})$ is defined as that in the proof of Theorem 4.1 to avoid the abuse of notations. By (A.17) and (A.18), we have

$$
\Pi_{n6} = - A_n \Sigma_n^{-1/2}(w) \left(M_n^\top M_n\right)^{-1} \left[ \sum_{t=1}^{n} \eta_{t1}^*, \ldots, \sum_{t=1}^{n} \eta_{td_n}^* \right]^\top + O_P(\sqrt{d_nh^\gamma}). \tag{A.19}
$$

We next consider $\Pi_{n7}$. Note that, by (A.16),

$$
\Pi_{n7} = A_n \Sigma_n^{-1/2}(w) \left(\hat{M}_n^\top \hat{M}_n\right)^{-1} \hat{M}_n \eta \\
= A_n \Sigma_n^{-1/2}(w) \left(\hat{M}_n^\top \hat{M}_n\right)^{-1} \left[ M_n^\top M_n \eta + (\hat{M}_n - M_n)^\top \eta \right] \left(1 + o_P(1)\right).
$$

32
As in the proof of Theorem 4.1, we can also show that the leading term of \( \Pi_n \) is \((M_n^T M_n)^{-1} M_n^T \eta\) by noting that \( nd_n h^{2\gamma} = o(1) \) and \( nd_n h^{-\frac{2\gamma+2}{p_0}} v_2(r_n) = o(1) \) and taking \( M_n = o(\sqrt{nh/d_n}) \). We thus have

\[
\Pi_n = (M_n^T M_n)^{-1} \left( \sum_{t=1}^{n} \eta_t, \ldots, \sum_{t=1}^{n} \eta_t' \right)^T (1 + o_P(1)).
\] (A.20)

Then, by (A.19), (A.20) and following the proof of Lemma B.4 in the supplemental document, we can prove (4.10). Then, the proof of Theorem 4.3 is completed.

References


PREPRINT SERIES 2014/2015

No. 1  ‘REMOVING THE FADDEEV-POPOV ZERO MODES FROM YANG-MILLS THEORY IN SPACETIMES WITH COMPACT SPATIAL SECTIONS’  Jos Gibbons & Atsushi Higuchi

No. 2  ‘INvariANTS OF SPECHT MODULES’  Stephen Donkin & Haralampos Geranios

No. 3  ‘A FLEXIBLE SEMIPARAMETRIC FORECASTING MODEL FOR TIME SERIES’  Degui Li, Oliver Linton & Zudi Lu