cemmap
centre for microdata methods and practice

A flexible semiparametric model for time series

Degui Li
Oliver Linton
Zudi Lu

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP28/12

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

A Flexible Semiparametric Model for Time Series

Degui Li* Monash University

Oliver Linton† University of Cambridge

Zudi Lu‡ University of Adelaide

August 6, 2012

Abstract

We consider approximating a multivariate regression function by an affine combination of one-dimensional conditional component regression functions. The weight parameters involved in the approximation are estimated by least squares on the first-stage nonparametric kernel estimates. We establish asymptotic normality for the estimated weights and the regression function in two cases: the number of the covariates is finite, and the number of the covariates is diverging. As the observations are assumed to be stationary and near epoch dependent, the approach in this paper is applicable to estimation and forecasting issues in time series analysis. Furthermore, the methods and results are augmented by a simulation study and illustrated by application in the analysis of the Australian annual mean temperature anomaly series. We also apply our methods to high frequency volatility forecasting, where we obtain superior results to parametric methods.

JEL subject classifications: C14, C22.

Keywords: Asymptotic normality, model averaging, Nadaraya-Watson kernel estimation, near epoch dependence, semiparametric method.

*Department of Econometrics and Business Statistics, Monash University, Caulfield East, VIC 3145, Australia. Email: degui.li08@gmail.com. Thanks to the ARC DECRA Fund for financial support.

†Faculty of Economics, Austin Robinson Building, Sidgwick Avenue, Cambridge, CB3 9DD, UK. Email: obl20@cam.ac.uk. Thanks to the ERC for financial support.

‡Section of Statistics, School of Mathematical Sciences, The University of Adelaide, SA 5005, Australia. Email: zudi.lu@adelaide.edu.au. Thanks to the ARC Future Fellowship for financial support.

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 implies absurd conclusions when covariates are pushed to extreme values. In regression settings, we may have to choose between a large number of covariates. In the time series case, the problem can get even worse, since in estimation and forecasting, all possible lags of all possible predictor variables may be candidates and their influence may be of unknown form. 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, this can be very time consuming. Also, such an approach may be neglecting features of the data that arrive through models that are not selected but which are almost as good as that which is 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 popular approach 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. However, most of the above literature on model averaging and selection has been concerned with parametric models, which assume some parametric linear or nonlinear relationships among the variables considered. In this paper, we will consider nonparametric and semiparametric models.

Let $(Y_t, X_t^\intercal)$ be a stationary time series process, where $X_t = (X_{t1}, \ldots, X_{td})^\intercal$ is a d-dimensional random vector and the superscript $^\intercal$ stands for the transpose of a vector or matrix. In many applications, we need to consider estimating regression function $\mathsf{E}(Y_t|X_t = x)$, where $x = (x_1, \ldots, x_d)^\intercal$, 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", Stone (1980). 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, 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 $\mathsf{E}(Y_t|Y_{t-1}, Y_{t-2} \ldots)$ 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

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

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 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 $m_j(x) = \mathsf{E}(Y_t^2|Y_{t-j} = x)$ 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 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. 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 obtain consistency and asymptotic normality of our procedures under general conditions. We further apply our methods to volatility forecasting (where the time series is long and (log) linear models are predominant) and to Australian temperature data (where the data is shorter but nonlinear parametric methods have already been considered) and obtain satisfactory results in both cases.

The rest of the paper is organized as follows. The model is presented in Section 2 and the estimation method is presented in Section 3. The asymptotic properties for the estimators of w_o 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 Appendix.

2 Model

We model or approximate the conditional regression function $\mathsf{E}(Y|X=x)$ 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 $J_\ell = \binom{d}{\ell}$. For example, $S_2 = \{(1,2),\ldots,(d-1,d)\}$ has cardinality $d(d-1)/2$. We model or approximate $m(x)=\mathsf{E}(Y|X=x)$ by

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

for some weights w_j, $j=0,1,\ldots,J$, where $X_{(j)} = (X_{i_1},\ldots,X_{i_{k_j}})^\intercal$ is a subset of X and $x_{(j)} = (x_{i_1},\ldots,x_{i_{k_j}})^\intercal$. 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 regressing function $m(x)$ or we could be thinking of this as an approximating or model averaging device. Either way, we are then seeking $w=(w_0,w_1,\ldots,w_J)^\intercal$ that minimizes

$$\mathsf{E}\Big[Y - w_0 - \sum_{j=1}^{J} w_j \mathsf{E}(Y|X_{(j)})\Big]^2. \tag{2.1}$$

In general, the minimizing weights may not be unique, but the minimization problem is a projection onto the space spanned by the functions $\{\mathsf{E}(Y|X_{(j)}), j=1,\ldots,J\}$ and so there 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), $w_o=(w_{o,0},w_{o,1},\ldots,w_{o,J})^\intercal$, satisfies

$$w_{o,0} = \big(1-\sum_{j=1}^{J} w_{o,j}\big)\mathsf{E}(Y), \quad (w_{o,1},\ldots,w_{o,J})^\intercal = A^{-1}a, \tag{2.2}$$

where A is a $J \times J$ matrix whose $(i,j)^{th}$ component is $\mathsf{Cov}(\mathsf{E}(Y|X_{(i)}), \mathsf{E}(Y|X_{(j)}))$, and a is a J-dimensional vector whose i^{th} component is $\mathsf{Cov}(\mathsf{E}(Y|X_{(i)}), Y)$. If the model is true, (2.1) is equal to zero at the optimal weights but it need not be so. Obviously the conditional component regressions $\mathsf{E}(Y|X_{(j)} = x_j)$, $j = 1, \ldots, J$, 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 we 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": $\mathsf{E}(Y|X_{(j)} = x_{(j)})$, $j = 1, \ldots, J$, 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. We 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 nonparametric case, this is not necessarily so, i.e.,

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

although there are clearly some nonlinear cases where this is so and where this would be a reasonable model. For example, when the regression model is additive and the covariates are mutually independent.

[1]The basic 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 is also true in infinite dimensional settings. Consider the AR(∞) model of the form

$$Y_t = \sum_{j=1}^{\infty} \rho_j Y_{t-j} + \varepsilon_t$$

for some (declining) coefficients ρ_j. Our general class of models would include processes of the form

$$Y_t = \sum_{j=1}^{\infty} \theta_j \mathsf{E}(Y_t|Y_{t-j}) + \varepsilon_t.$$

In the special linear Gaussian case, the two representations are equivalent (since all the $\mathsf{E}(Y_t|Y_{t-j})$ are linear functions). However, in general they will be different, even in the linear but non-Gaussian case (Tong, 1990, p13).

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

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

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

$$\inf_{w} \mathsf{E}\Big[Y - w_0 - \sum_{j=1}^{d} w_j \mathsf{E}(Y|X_j)\Big]^2 \geq \inf_{h} \mathsf{E}\Big[Y - w_0 - \sum_{j=1}^{d} h_j(X_j)\Big]^2,$$

which means that the 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

$$\inf_{w} \mathsf{E}\Big[\mathsf{E}_{Add}(Y|X) - w_0 - \sum_{j=1}^{d} w_j \mathsf{E}(Y|X_j)\Big]^2,$$

where $\mathsf{E}_{Add}(Y|X)$ is the best additive fit of Y by the vector X. The space generated by $w_0 + \sum_{j=1}^{d} w_j \mathsf{E}(Y|X_j)$ 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, and Linton 2000 for more discussion on estimation of the additive models). However, if we compare with the type of models considered in Linton and Mammen (2005), where there is only one unknown function m, then the MSE of the two models is non-nested. That is, there are regression functions for which

$$\inf_{w} \mathsf{E}\Big[Y - w_0 - \sum_{j=1}^{d} w_j \mathsf{E}(Y|X_j)\Big]^2 \leq \inf_{w;m} \mathsf{E}\Big[Y - w_0 - \sum_{j=1}^{d} w_j m(X_j)\Big]^2.$$

In fact, we would generally expect this ordering of the MSE especially when d is large.

The main advantage of our method is computational, and perhaps performance in the case where d is large. 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 the nonparametric additive models, or even variations such as Linton and Mammen (2005).

The modelling approach is also 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

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

where w is a parameter vector and $\mathcal{C}$ is a "regression copula", in our case known (given w). We will discuss this further below.

3 Estimation

In this section we define our estimation procedures using matrix formulae, which facilitate efficient coding. Without loss of generality we assume that $\mathsf{E}(Y) = 0$, otherwise we replace Y by $Y - \mathsf{E}(Y)$ 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 $(Y_t, X_t^\intercal)$, $t = 1, \ldots, n$. Let $m(x) = \mathsf{E}(Y_t | X_t = x)$ and $m_j(x_j) = \mathsf{E}(Y_t | X_{tj})$, $j = 1, \ldots, d$. We first estimate $m_j(x_j)$ by using the Nadaraya-Watson kernel method

$$\widehat{m}_j(x_j) = \frac{\sum_{t=1}^{n} Y_t K\left(\frac{X_{tj} - x_j}{h_j}\right)}{\sum_{t=1}^{n} K\left(\frac{X_{tj} - x_j}{h_j}\right)}, \tag{3.1}$$

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 $\mathcal{M}_j = \left[m_j(X_{1j}), \ldots, m_j(X_{nj})\right]^\intercal$ and $\widehat{\mathcal{M}}_j = \left[\widehat{m}_j(X_{1j}), \ldots, \widehat{m}_j(X_{nj})\right]^\intercal$. $\widehat{\mathcal{M}}_j$ is the Nadaraya-Watson estimator of $\mathcal{M}_j$ and we have

$$\widehat{\mathcal{M}}_j = \mathcal{S}_j \mathcal{Y},$$

where $\mathcal{S}_j$ is the $n \times n$ smoother matrix associated with X_j, $\mathcal{Y}$ is the $n \times 1$ vector of observations on the response, $\mathcal{Y} = (Y_1, \ldots, Y_n)^\intercal$. Then, for given $w = (w_1, \ldots, w_d)^\intercal$,

$$\widehat{\mathcal{M}}_w = (w_1 \mathcal{S}_1 + \cdots + w_d \mathcal{S}_d)\, \mathcal{Y} =: \mathcal{S}(w)\mathcal{Y}.$$

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

$$\begin{aligned} Q(w) &= (\mathcal{Y} - \widehat{\mathcal{M}}_w)^\intercal (\mathcal{Y} - \widehat{\mathcal{M}}_w) \\ &= \mathcal{Y}^\intercal (I - w_1 \mathcal{S}_1 - \cdots - w_d \mathcal{S}_d)^\intercal (I - w_1 \mathcal{S}_1 - \cdots - w_d \mathcal{S}_d) \mathcal{Y} \\ &= \mathrm{Tr}\left[\mathcal{Y}\mathcal{Y}^\intercal \mathcal{P}(w)\right], \end{aligned} \tag{3.2}$$

where I is the $n \times n$ identity matrix and

$$\begin{aligned} \mathcal{P}(w) &= (I - w_1 \mathcal{S}_1 - \cdots - w_d \mathcal{S}_d)^\intercal (I - w_1 \mathcal{S}_1 - \cdots - w_d \mathcal{S}_d) \\ &= I - \sum_{j=1}^{d} w_j (\mathcal{S}_j + \mathcal{S}_j^\intercal) + \sum_{j=1}^{d} w_j^2 \mathcal{S}_j^\intercal \mathcal{S}_j + \sum_{i=1}^{d-1} \sum_{j=i+1}^{d} w_i w_j (\mathcal{S}_i^\intercal \mathcal{S}_j + \mathcal{S}_j^\intercal \mathcal{S}_i), \end{aligned}$$

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

$$\frac{\partial}{\partial w_i}Q(w) = \mathrm{Tr}\left[\mathcal{Y}\mathcal{Y}^\intercal \frac{\partial}{\partial w_i}\mathcal{P}(w)\right] = -\mathcal{Y}^\intercal(\mathcal{S}_i + \mathcal{S}_i^\intercal)\mathcal{Y} + 2w_i\mathcal{Y}^\intercal\mathcal{S}_i^\intercal\mathcal{S}_i\mathcal{Y} + \sum_{j\neq i} w_j\mathcal{Y}^\intercal(\mathcal{S}_i^\intercal\mathcal{S}_j + \mathcal{S}_j^\intercal\mathcal{S}_i)\mathcal{Y} = 0.$$

We can write this as

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

where $\widehat{A}_{ij} = \mathcal{Y}^\intercal(\mathcal{S}_i^\intercal\mathcal{S}_j + \mathcal{S}_j^\intercal\mathcal{S}_i)\mathcal{Y}$ and $\widehat{a}_i = \mathcal{Y}^\intercal(\mathcal{S}_i + \mathcal{S}_i^\intercal)\mathcal{Y}$. Then we have

$$\widehat{w} = (\widehat{w}_{o,1}, \ldots, \widehat{w}_{o,d})^\intercal = \widehat{A}^{-1}\widehat{a}. \tag{3.3}$$

Defining

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

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

$$\widehat{w} = \left(\widehat{\mathcal{M}}^\intercal\widehat{\mathcal{M}}\right)^{-1}\widehat{\mathcal{M}}^\intercal\mathcal{Y}. \tag{3.4}$$

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

$$\widehat{m}(x) := \widehat{m}_{\widehat{w}}(x) = \sum_{j=1}^d \widehat{w}_{o,j}\widehat{m}_j(x_j). \tag{3.5}$$

By the discussion in Section 2, $\widehat{m}(x)$ can only be seen as the approximated value for $m(x)$ as $\sum_{i=1}^d w_{o,j}m_j(x_j)$ does not necessarily equal to $m(x)$ except the full linear regression case.

In Section 4.1, we will show that $\widehat{w}$ is asymptotically normal with root-n convergence rate. The nonparametric estimator $\widehat{m}(x)$ is also asymptotically normal with root-(nh) convergence rate, and thus the curse of dimensionality is avoided. Furthermore, in Section 4.2, we will consider the more general case that the dimension of X_t is diverging, i.e., $d_n \to \infty$ as $n \to \infty$, which is common in modern time series analysis.

4 Asymptotic properties

In this paper, we assume that $\{(Y_t, X_t^\intercal), t \geq 1\}$ belongs to a class of stationary near epoch dependent (NED) or stable processes, which is more general than the α-mixing process. Based on a stationary

process $\{\varepsilon_t\}$, $\{Y_t\}$ and $\{X_t\}$ are defined by

$$\begin{aligned} Y_t &= \Psi_Y(\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}, \ldots), \\ X_t &= (X_{t1}, \cdots, X_{td})^\intercal = \Psi_X(\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}, \ldots), \end{aligned} \tag{4.1}$$

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 NED process is provided as follows.

DEFINITION 4.1. *The stationary process $\{(Y_t,\ X_t^\intercal)\}$ is said to be near epoch dependent in L_ν norm (NED in L_ν) with respect to a stationary α-mixing process $\{\varepsilon_t\}$, if*

$$v_\nu(m) = \mathsf{E}\left[|Y_t - Y_t^{(m)}|^\nu + \|X_t - X_t^{(m)}\|^\nu\right] \to 0, \quad \nu > 0, \tag{4.2}$$

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

From the above definition, we know that the NED process includes the α-mixing process as a special case. The concept of NED process can date back to Ibragimov (1962), and it is further developed by Billingsley (1968), McLeish (1975a, 1975b, 1977) and Lin (2004), most of which assume that $\{\varepsilon_t\}$ is stationary martingale difference or φ-mixing. In this paper, we study the NED process with respect to a stationary α-mixing process $\{\varepsilon_t\}$ (α-mixing dependence is weaker than the φ-mixing). As pointed by Lu and Linton (2007), such NED process can easily cover some important compounded econometric processes, which are not covered by the α-mixing processes. Because of this, there has been extensive literature on estimation and testing issues under NED assumption, see, for example, Andrews (1995), Lu (2001), Ling (2007), Lu and Linton (2007) and Li *et al* (2012). In this paper, we estimate the weight w_o in the context of stationary NED process, which makes our methodology applicable for some popular time series models such as AR(p)-GARCH(1,1) model.

In this section, we derive the asymptotic theory for $\widehat{w}$ and $\widehat{m}(x)$ for two cases: (i) the dimension of $\{X_t\}$ is finite; 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 The dimension of $\{X_t\}$ is finite

We start with a simple case that d is fixed. 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 $\int K(u)du = 1$, $\int u^l K(u) = 0$, for $1 \le l \le \gamma - 1$, 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 $\{X_{tj}\}$, $f_j(\cdot)$, have continuous derivatives up the $(\gamma+1)$-order and $\inf_{1\le j\le d}\inf_{x\in\Omega_j} f_j(x) > 0$, where Ω_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 $1 \le j \le d$ and $k \ge 1$, and satisfies that for some positive integer k^* and all $k \ge k^*$, $f_{j,k}(x_1, x_2) < C_f$ for $1 \le j \le d$, 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}, $k \ne j$, $f_{j|k}(\cdot|\cdot)$, exists and satisfies the Lipschitz continuous condition.

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

ASSUMPTION 3 (i) $\{(Y_t, X_t^\intercal), t \ge 1\}$ is stationary NED in L_{p_0}-norm with respect to a stationary α-mixing process $\{\varepsilon_t\}$ with $\mathsf{E}\big[|Y_t|^{p_0}\big] < \infty$, where $p_0 = 2 + \delta$, $\delta > 0$.

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

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

ASSUMPTION 4 (i) The bandwidths, h_j, $j = 1, \ldots, d$, satisfy $h_j = c_j h$ for some positive constant c_j, $1 \le j \le d$, and

$$nh^{2\gamma} \to 0, \quad \frac{n^{\frac{p_0-2}{p_0}} h}{\log n} \to \infty \ \text{ as } n \to \infty. \tag{4.3}$$

(ii) There exists two sequences of positive integers r_n and R_n such that

$$r_n \to \infty, \quad r_n = o\big(R_n \vee n^{1/2} h^{-\frac{p_0+2}{2p_0}}\big), \quad R_n\Big(h + n^{\frac{2-p_0}{p_0}} h^{-1} \log n\Big) \to 0. \tag{4.4}$$

Furthermore, let

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

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

REMARK 4.1. Assumption 1 is a commonly-used condition on higher-order kernel function to reduce the influence of the asymptotic bias of the nonparametric estimators, 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 $\{\varepsilon_t\}$. Note that, in Assumption 3 (ii), we assume that α-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

$$\eta_t = Y_t - \sum_{j=1}^{d} w_{o,j} m_j(X_{tj}), \quad \eta_{tj} = Y_t - \mathsf{E}(Y_t|X_{tj}) = Y_t - m_j(X_{tj})$$

and $\beta_{jk}(X_{sk}) = \mathsf{E}\left(m_j(X_{sj})|X_{sk}\right)$. Let $\xi_t = (\xi_{t1}, \ldots, \xi_{td})^\intercal$ with

$$\xi_{tj} = \eta'_{tj} - \eta^*_{tj}, \quad \eta'_{tj} = m_j(X_{tj})\eta_t, \quad \eta^*_{tj} = \sum_{k=1}^{d} w_{o,k}\eta_{tk}\beta_{jk}(X_{tk}).$$

Define

$$\Lambda = \begin{pmatrix} \mathsf{E}\big[m_1(X_{t1})m_1(X_{t1})\big] & \cdots & \mathsf{E}\big[m_1(X_{t1})m_d(X_{td})\big] \\ \vdots & \vdots & \vdots \\ \mathsf{E}\big[m_d(X_{td})m_1(X_{t1})\big] & \cdots & \mathsf{E}\big[m_d(X_{td})m_d(X_{td})\big] \end{pmatrix} \quad \text{and} \quad \Sigma = \sum_{t=-\infty}^{\infty} \mathsf{E}\left[\xi_0\xi_t^\intercal\right].$$

We give the asymptotic distribution of $\widehat{w}$ in the following theorem.

THEOREM 4.1. *Suppose that the assumptions 1–4 are satisfied and* Λ *is positive definite. Then, we have*

$$\sqrt{n}\left(\widehat{w} - w_o^*\right) \xrightarrow{d} \mathsf{N}\left(\mathbf{0}, \Lambda^{-1}\Sigma\Lambda^{-1}\right), \tag{4.6}$$

where $w_o^* = (w_{o,1}, \ldots, w_{o,d})^\intercal$.

REMARK 4.2. When d is finite, the above theorem shows that the parametric estimator of the optimal weight can achieve the root-n convergence rate although we replace $\mathsf{E}(Y_t|X_{tj})$ by its nonparametric estimator.

Define

$$\Sigma_1(x) = \big(\int K^2(u)du \big) \mathsf{diag}\big\{ \frac{c_1\sigma_1^2(x_1)}{f_1(x_1)}, \ldots, \frac{c_d\sigma_d^2(x_d)}{f_d(x_d)} \big\} \quad \text{and} \quad \sigma_j^2(x_j) = \mathsf{E}\left[\eta_{tj}^2 | X_{tj} = x_j \right],$$

where $x = (x_1, \ldots, x_d)^\tau$ and c_j is defined as in Assumption 4 (i). We next give the asymptotic distribution for $\widehat{m}(x)$.

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

$$\sqrt{nh}\left(\widehat{m}(x) - m_w(x) \right) \xrightarrow{d} \mathsf{N}\left(0, \sigma_w^2 \right), \tag{4.7}$$

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

REMARK 4.3. Theorem 4.2 above shows that the proposed nonparametric estimator $\widehat{m}_w(x)$ is asymptotically normal and enjoys the convergence rate of the nonparametric estimator in the standard univariate nonparametric regression. Furthermore, if $m_w(x) = m(x)$, which indicates that there is no approximation bias, by (4.7), we can easily prove that

$$\sqrt{nh}\left(\widehat{m}(x) - m(x) \right) \xrightarrow{d} \mathsf{N}\left(0, \sigma_w^2 \right). \tag{4.8}$$

4.2 The dimension of $\{X_t\}$ is diverging

We next consider the case that the dimension of $\{X_t\}$ increases with the sample size, which would have potential applications in nonlinear forecasting with very large lag terms. It is well known that the additive model performs poorly for the high-dimensional case, which is our motivation to find an alternative method to deal with this case. In this section, we consider the semiparametric approximation as discussed in the previous sections but with d replaced by d_n which increases with the sample size n. To avoid confusion, we let $\widehat{w}(n)$ and $w_o^*(n)$ be defined as $\widehat{w}$ and w_0^* with d replaced by d_n. Define

$$\Lambda_n = \begin{pmatrix} \mathsf{E}\big[m_1(X_{t1}) m_1(X_{t1}) \big] & \cdots & \mathsf{E}\big[m_1(X_{t1}) m_{d_n}(X_{td_n}) \big] \\ \vdots & \vdots & \vdots \\ \mathsf{E}\big[m_{d_n}(X_{td}) m_1(X_{t1}) \big] & \cdots & \mathsf{E}\big[m_{d_n}(X_{td_n}) m_{d_n}(X_{td_n}) \big] \end{pmatrix},$$

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

ASSUMPTION 5 (i) There exists a compact support Ω such that $\cup_j \Omega_j \in \Omega$, where Ω_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

$$d_n(\tau_n + h^\gamma) = o(1), \quad n d_n h^{2\gamma} = o(1), \quad n d_n h^{-\frac{p_0+2}{p_0}} v_2(r_n) = o(1), \tag{4.9}$$

where τ_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 could increase with the sample size at some polynomial rate.

We next give the asymptotic normal distribution theory for $\widehat{w}(n)$. Let $\mathcal{A}_*$ be a given non-negative $p \times p$ matrix, for example the identity, and let $\mathcal{A}_n$ be a $p \times d_n$ matrix such that as $n \to \infty$

$$\mathcal{A}_n \mathcal{A}_n^\intercal \to \mathcal{A}_*.$$

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

$$\sqrt{n}\mathcal{A}_n \Sigma_n^{-1/2}(w)\big(\widehat{w}(n) - w_o^*(n)\big) \xrightarrow{d} \mathsf{N}\big(\mathbf{0}, \mathcal{A}_*\big). \tag{4.10}$$

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 $\|\widehat{w}(n) - w_o^*(n)\| = O_P(\sqrt{d_n/n})$. Such result is analogous to some existing results in statistics literature such as Theorems 1 and 2 in Lam and Fan (2008). By using Theorem 4.3, we can also prove the asymptotic distribution for nonparametric estimator as in Theorem 4.2. We will discuss this issue in details in the context of nonlinear forecasting in Section 5.2 below.

5 Some extensions

In this section, we give the discussions of some related topics including the nonlinear forecasting issue in time series analysis, and applications in 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 $\{Y_t : 1 \le t \le T\}$, where k_0 is a fixed number. Consider the forecast of the form defined by

$$\widetilde{Y}_{T+k_0|T} = w_0 + \sum_{j=1}^{d_T} w_j \mathsf{E}(Y_{T+k_0}|Y_{T+1-j}). \tag{5.1}$$

Note that for a stationary process, $\mathsf{E}(Y_{T+k_0}|Y_{T+1-j} = y) = \mathsf{E}(Y_t|Y_{t+1-j-k_0} = y)$, which can be estimated using $\{Y_t : 1 \le t \le T\}$, 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 $\widehat{Y}_{T+k_0|T}$, the predicted value of $\widetilde{Y}_{T+k_0|T}$, by replacing $\mathsf{E}(Y_t|Y_{t+1-j-k_0} = y)$ 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, under some conditions,

$$\sqrt{Th/\sigma_w^2(T)}\big(\widehat{Y}_{T+k_0|T} - \widetilde{Y}_{T+k_0|T}\big) \xrightarrow{d} \mathsf{N}\big(0, 1\big), \tag{5.2}$$

where $\sigma_w^2(T)$ is defined as σ_w^2 in Theorem 4.2 with d and $(x_1, \ldots, x_d)^\intercal$ replaced by d_T and $(Y_T, \ldots, Y_{T-d_T+1})^\intercal$, 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 $j \ge d$, 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 $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_k)^\intercal$ 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 lap operator, and let the w_j solve the equation

$$\sum_{j=0}^{\infty} w_j L^j = \frac{A(L)}{B(L)}, \tag{5.4}$$

which will happen uniquely under some conditions. Then the weights w_j, $j = 1, 2, \ldots$, just depend on the parameter vector $\boldsymbol{\theta} = (a_0, \ldots, a_p, b_0 \ldots, b_q)^\intercal$. The problem then reduces to choosing $\boldsymbol{\theta}$ 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). We will study this issue in our future research.

5.2 Limited dependent variables and robust methods

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

$$\mathsf{E}(Y|X) = \mathcal{C}(\mathsf{E}(Y|X_1), \ldots, \mathsf{E}(Y|X_d); w), \tag{5.5}$$

where w is a parameter vector and $\mathcal{C}$ 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 $\mathcal{C}$. 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 $\mathsf{P}(Y = 1|X = x) = \mathsf{E}(Y|X = x) = F(\beta_0 + \sum_{j=1}^{d} \beta_j x_j)$, where F is the normal or logistic cdf. Therefore, consider

$$\mathsf{P}(Y = 1|X = x) = \mathsf{E}(Y|X = x) = F\left(w_0 + \sum_{j=1}^{d} w_j F^{-1}(\mathsf{E}(Y|X_j = x_j))\right),$$

where F is a known c.d.f. Compare this with the generalized additive modelling approach, see Hastie and Tibshirani (1990) in which the marginal regression functions $\mathsf{E}(Y|X_j = x_j)$ 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 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_α denotes the level α 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 semiparametric method by Monte Carlo simulation and real data examples, to uncover and understand the time series lag effects in applications. Monte Carlo simulation example is provided in the first subsection, and the second one is the analysis of two real data sets, the Australian annual mean temperature anomaly series and the one minute data from the FTSE100 index.

6.1 Monte Carlo simulation

We consider the model in our simulation as follows:

$$Y_t = \sum_{k=1}^{9} g_{0k}(Y_{t-k}) + \varepsilon_t \tag{6.1}$$

with

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

and $\varepsilon_t \sim$ i.i.d. $\mathsf{N}(0, \sigma^2)$, where the values of σ^2 and a_k's, for $k = 1, 2, \cdots, 9$, 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 Section 6.2. Let δ and γ 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 ε_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. $\mathsf{N}(0,1)$, where $\alpha_0 > 0$, α_1, α_2 and β_1 are suitable non-negative constants, so that

Y_t is a NED process with respect to an α-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 strictly stationary solution that is geometrically ergodic and thus α-mixing, a special NED process defined in Definition 1.1, with mixing coefficient $\alpha(m) = O(\rho^m)$ as $m \to \infty$, $0 < \rho < 1$, for any real values of δ and γ, owing to the fact that $1 - \sum_{k=1}^{9} a_k z^k \neq 0$ for any $|z| \leq 1$ with a_k specified in Table 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 σ^2 and the coefficients a_k's in the simulating model (6.1).

σ^2	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9
0.08498	-0.1129	0.0245	-0.1892	-0.0820	-0.1962	-0.1232	0.1180	0.1282	-0.2407

It is worth mentioning that in the linear case, the fit in equations (2.1) and (2.2) is equal to the full linear regression model fit. However, this may not hold in the nonparametric case. For example, in model (6.1) with $\gamma \neq 0$, $\mathsf{E}(Y_t|X_t) \neq \sum_{j=1}^{d} w_j \mathsf{E}(Y_t|X_{tj})$, where $X_t = (X_{t1}, \ldots, X_{td})^\intercal$ with $X_{tj} = Y_{t-j}$ and $d = 9$. Furthermore, the fit in equations (2.1) and (2.2) can be seen as an additive function of the individual components, i.e., $\sum_{j=1}^{d} w_j \mathsf{E}(Y_t|X_{tj}) = \sum_{j=1}^{d} h_j(X_{tj})$ for a specific set of functions $h_j(X_{tj})$. Therefore, as introduced in Section 2,

$$\inf_{w} \mathsf{E}\Big[\mathsf{E}(Y_t|X_t) - w_0 - \sum_{j=1}^{d} w_j E(Y_t|X_{tj})\Big]^2 \geq \inf_{g} \mathsf{E}\Big[\mathsf{E}(Y_t|X_t) - w_0 - \sum_{j=1}^{d} g_j(X_{tj})\Big]^2, \tag{6.2}$$

which indicates that the additive fit could reduce the mean squared error. In this simulation, we are interested in understanding how large the difference between the LHS and the RHS of (6.2) from the prediction perspective by Monte Carlo simulation.

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 (δ, γ) 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 $n_{est} = n - n_{pred}$ for model estimation, and the second part is a prediction sample of size $n_{pred} = 50$

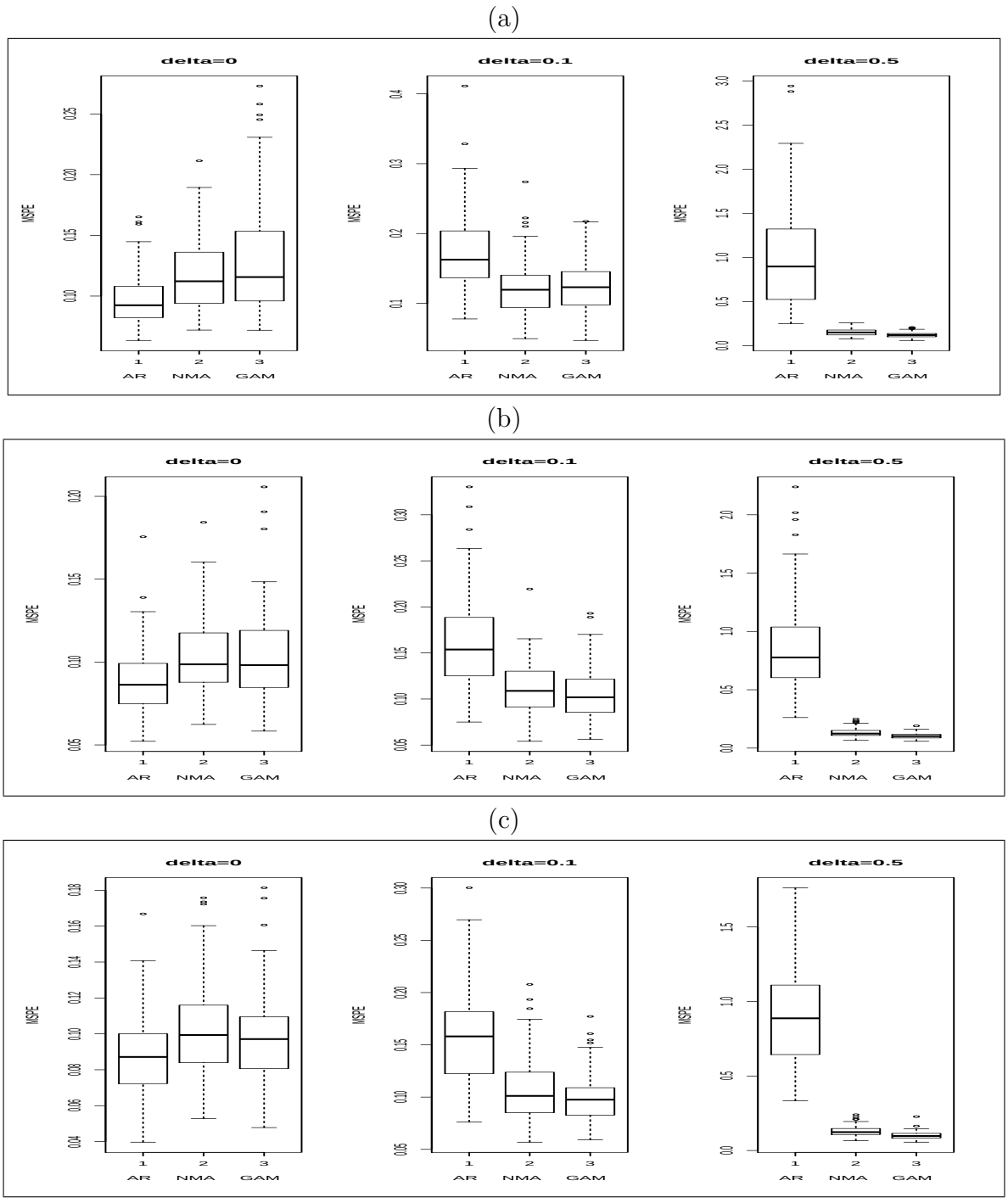


Figure 1: *Simulation — Boxplot of 100 repetitions of the mean squared error of* $n_{pred} = 50$ *one-step-ahead predictions for* $\gamma = 0$*: (a)* $n_{est} = 90$*, (b)* $n_{est} = 150$*, (c)* $n_{est} = 200$*. Here "AR", "NMA" and "GAM" stand for the prediction based on linear AR, semiparametric nonlinear model average and additive AR modelling, respectively.*

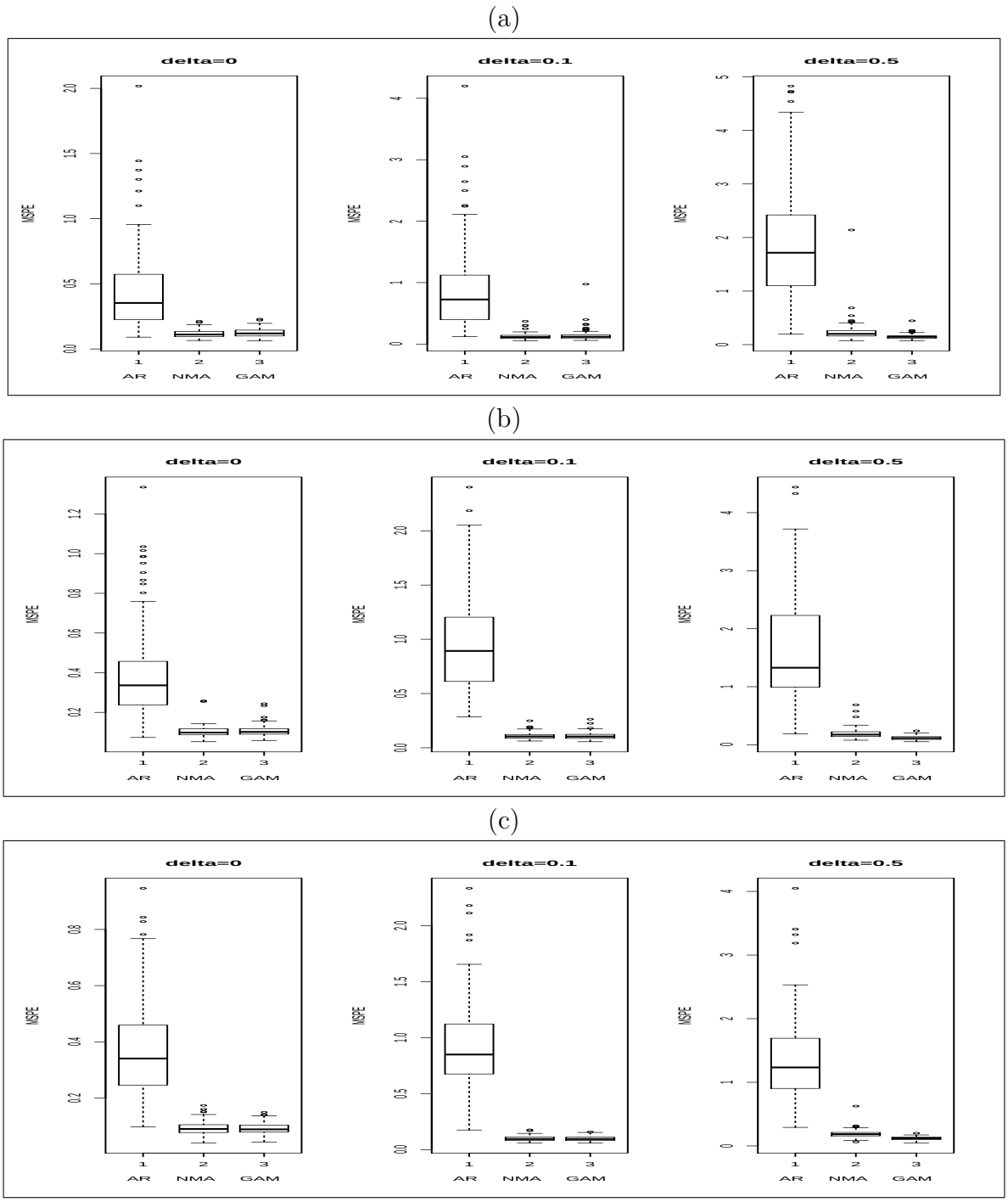


Figure 2: *Simulation — Boxplot of 100 repetitions of the mean squared error of $n_{pred} = 50$ one-step-ahead predictions for $\gamma = 0.1$: (a) $n_{est} = 90$, (b) $n_{est} = 150$, (c) $n_{est} = 200$. Here "AR", "NMA" and "GAM" stand for the prediction based on linear AR, semiparametric nonlinear model average and additive AR modelling, respectively.*

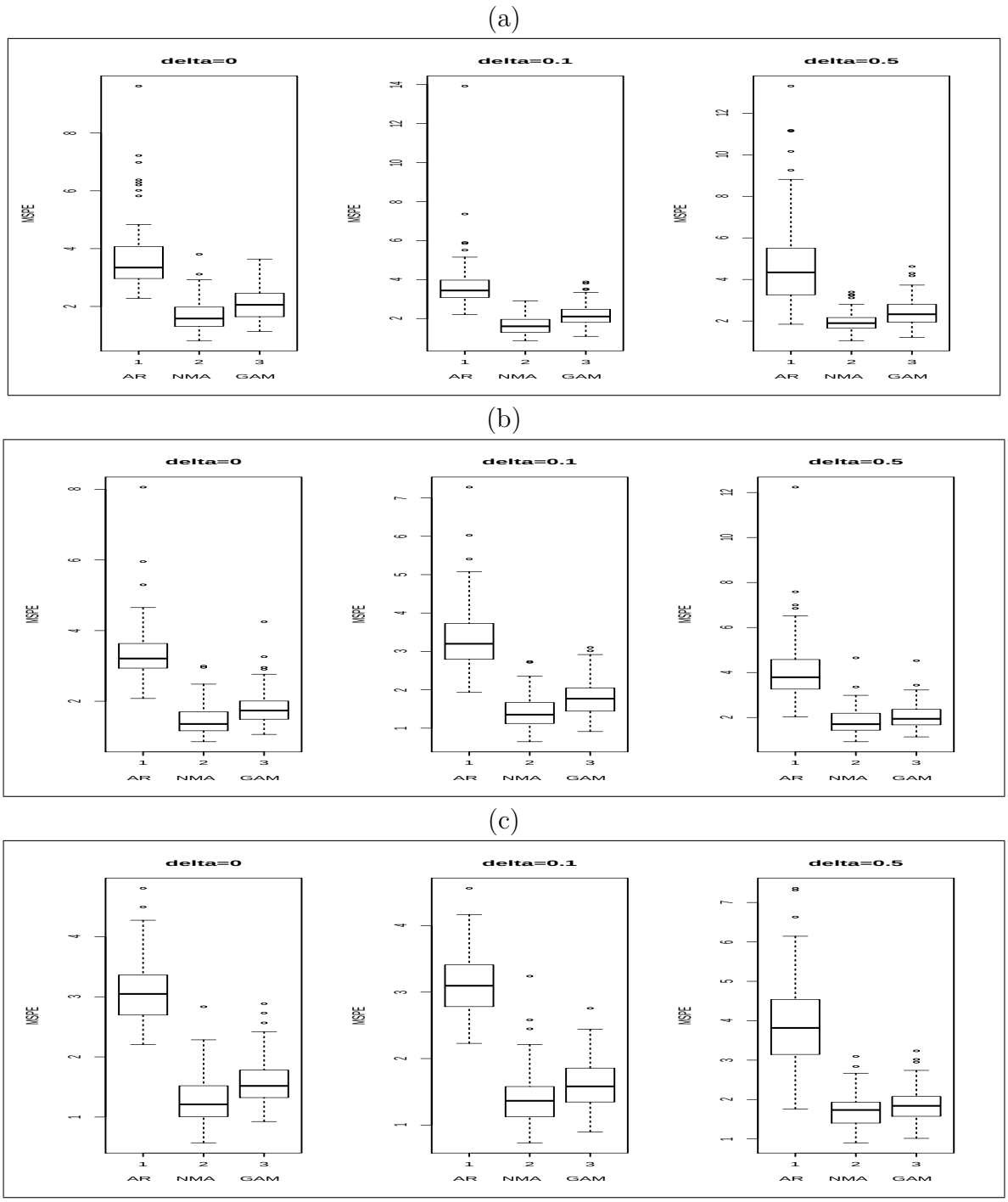


Figure 3: *Simulation — Boxplot of 100 repetitions of the mean squared error of* $n_{pred} = 50$ *one-step-ahead predictions for* $\gamma = 0.5$*: (a)* $n_{est} = 90$*, (b)* $n_{est} = 150$*, (c)* $n_{est} = 200$*. Here "AR", "NMA" and "GAM" stand for the prediction based on linear AR, semiparametric nonlinear model average and additive AR modelling, respectively.*

for evaluation of different prediction methods. We are considering three cases of $n = 140$, $n = 200$ and $n = 250$ so that $n_{est} = 90$, $n_{est} = 150$, $n_{est} = 200$, respectively. Three prediction methods are compared: linear AR model of order 9, semiparametric nonlinear model averaging of lag 9 (proposed in this paper), and the nonlinear additive AR model of order 9. The estimation procedures for linear AR model and nonlinear additive AR model are based on arima with "ML" method and gam with gaussian family and smoothing splines in the R packages STATS and GAM, respectively. The estimation procedure for semiparametric model averaging 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 $i = 1, 2, \cdots, n_{pred}$, based on the estimated models, and consider the mean squared prediction error, defined by

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

We repeat the simulation for 100 times, and the boxplots of the 100 MSPE values for three different methods with different pairs of (δ, γ) 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", "NMA" and "GAM" stand for the methods of prediction based on linear AR, semiparametric nonlinear model average and additive AR modelling, 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.

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. 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 'GAM" performs much worse than our "NMA" method in prediction although all three 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 our "NMA" prediction method than that in "GAM" model. 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 "NMA" method performs quite well in prediction. When the actual model is a purely additive model, it performs quite close to the optimal additive prediction "GAM". While a purely additive model is violated, it may even be better than the "GAM" in prediction. The main advantage of our method is computational, and perhaps performance in the case where d is large.

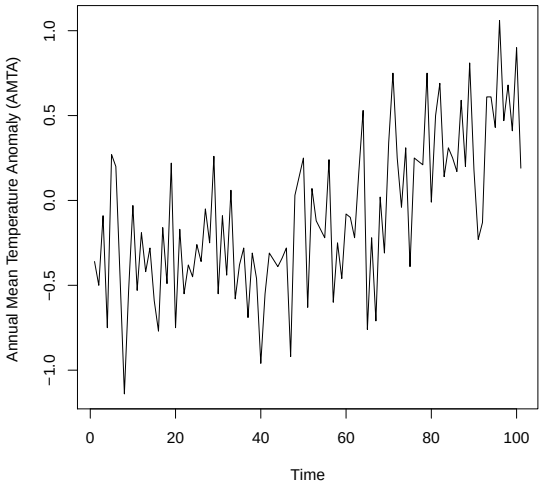
6.2 Two real data examples

6.2.1 Australian annual mean temperature anomaly (AMTA) series

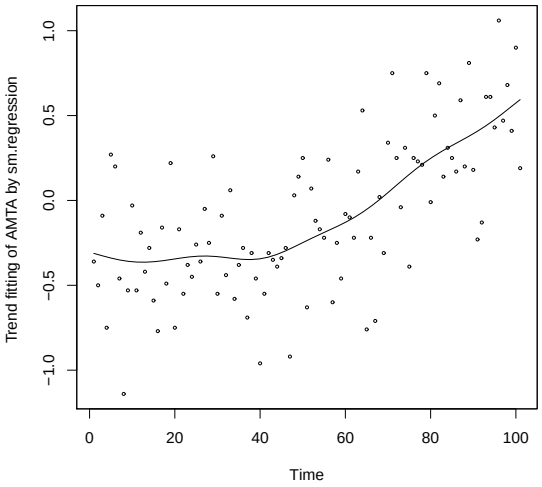
Our first real data set is the Australian annual mean temperature anomaly (AMTA) series starting from 1910 to year 2010, downloaded at http://www.bom.gov.au/cgi-bin/climate/change/timeseries.cgi. The time series plot of the data is illustrated in Figure 4(a). Obviously there is an increasing nonlinear trend in the AMTA series, indicating a global warming effect in Australia. We therefore remove this nonlinear trend by using the sm.regression function in R SM package, which is plotted in Figure 4(b); the resulting residual series, AMTA.res, plotted in Figure 4(c), appears stationary.

We are analysing the stationary series, AMTA.res, the sample size of which is $n = 101$. We partition it into two sub-samples for model estimation and prediction evaluation by taking $n_{pred} = 10$ so that the estimation sample is of size $n_{est} = n - n_{pred} = 91$. We first apply linear AR(p) analysis to the estimation sub-sample of AMTA.res by using R with p from 1 to 12. According to Akaike Information Criterion (AIC), AR(9) model is selected (see Table 2). We also tried ARMA(9,1) model for the series, the AIC value of which is 56.97. This demonstrates that the AR(9) model is reasonable for the series AMTA.res, the estimated coefficients of which are reported in Table 3, indicating the

(a)



(b)



(c)

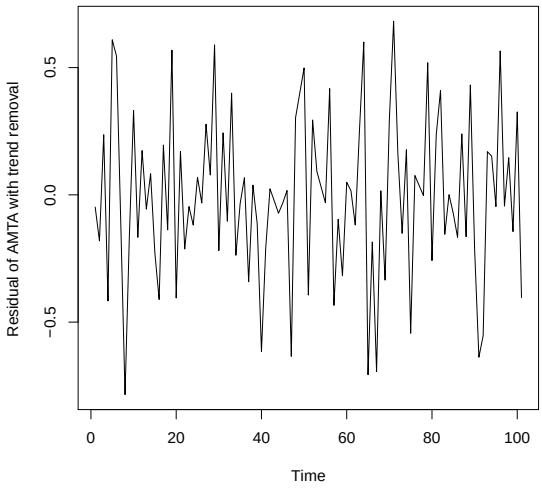


Figure 4: *Australian annual mean temperature anomaly (AMTA): (a) The series starting from 1910 to year 2010; (b) The trend fitting by sm.regression. (c) The resulting residual series, AMTA.res, after trend removal.*

long term lag effects up to 9 years in the annual mean temperature anomaly series. We can clearly see that the coefficients of ar1 – ar7 are insignificant from zero at the 5% significance level in this linear analysis.

Table 2: AIC values of AR(p) models with order p from 1 to 12 for the estimation sample of AMTA.res.

AR order	$p=1$	$p=2$	$p=3$	$p=4$	$p=5$	$p=6$
AIC	56.4	58.39	58.11	59.49	59.49	60.87
AR order	$p=7$	$p=8$	$p=9$	$p=10$	$p=11$	$p=12$
AIC	62.52	59.13	**54.97**	56.97	58.97	58.37

Table 3: The estimated coefficients (coef.) and their standard errors (s.e.) in the selected linear AR(9) model for the series, AMTA.res:

	intercept	ar1	ar2	ar3	ar4
coef.	-0.0074	-0.1086	0.0161	-0.1637	-0.0962
s.e.	0.0197	0.1029	0.1012	0.1007	0.1004
	ar5	ar6	ar7	ar8	ar9
coef.	-0.1429	-0.1154	0.1251	0.2075	-0.2775
s.e.	0.1017	0.1046	0.1041	0.1077	0.1088

However, checking the kernel density estimate of the AMTA.res series in Figure 5, we find that it is not a Gaussian series, showing that some nonlinear effects may exist in this series. We apply the semiparametric method proposed in this paper to explore the individual nonlinear lag effects. We examine the lags from 1 to 9. Denote the AMTA.res series by y_t. We estimate the nonlinear individual lag effects, $\mathsf{E}(y_t|y_{t-k})$, for $k = 1, 2, 3, \cdots, 9$, the local constant estimators of which, by applying sm.autoregression in R sm package, are plotted in Figure 6(a)–(i). The nonlinear individual lag effects appear clear in Figure 6(a)–(c) and (f)–(g). We then apply model averaging by considering

minimisation of $\mathsf{E}\left(y_t - w_0 - \sum_{k=1}^{9} w_k \mathsf{E}(y_t|y_{t-k})\right)^2$ with respect to $w_0, w_1, \cdots, w_9$, and a least squares estimate is made by minimising $\sum_{t=10}^{n_{est}} \left[y_t - w_0 - \sum_{k=1}^{9} w_k \hat{\mathsf{E}}(y_t|y_{t-k})\right]^2$, where $\hat{\mathsf{E}}(y_t|y_{t-k})$ is the local constant estimator of $\mathsf{E}(y_t|y_{t-k})$. The estimated values of w_k, $k = 0, 1, 2, \cdots, 9$, are given in Table 4.

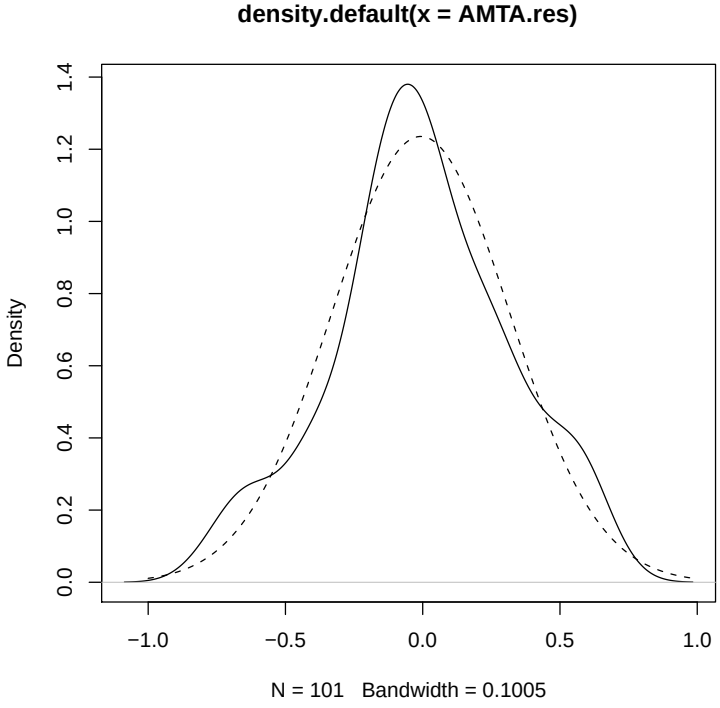


Figure 5: *Density estimate of AMTA.res: The solid line for kernel estimate and the dashed line for the Gaussian density with the same mean and variance.*

Table 4 shows that the coefficients of the lag effects in nonparametric model averaging are mostly significant at 5% significance level. We particularly note that the effects of the lags 1–6 that are insignificant in the linear AR model (Table 3) become significantly different from zero in nonparametric model averaging (Table 4), indicating that these lag effects may essentially be nonlinear in annual mean temperature anomaly. In addition, it follows from Table 5 that the mean of residual squares for our proposed "NMA" is 0.04652649, much smaller than that of linear "AR", 0.08261, in the estimation sample. This is further evidenced by the mean squared error of the one-step-ahead prediction, MSPE, over the evaluation sample of size $n_{pred} = 10$, which is 0.1136318 and 0.1001797 for the linear "AR" and our proposed "NMA", respectively.

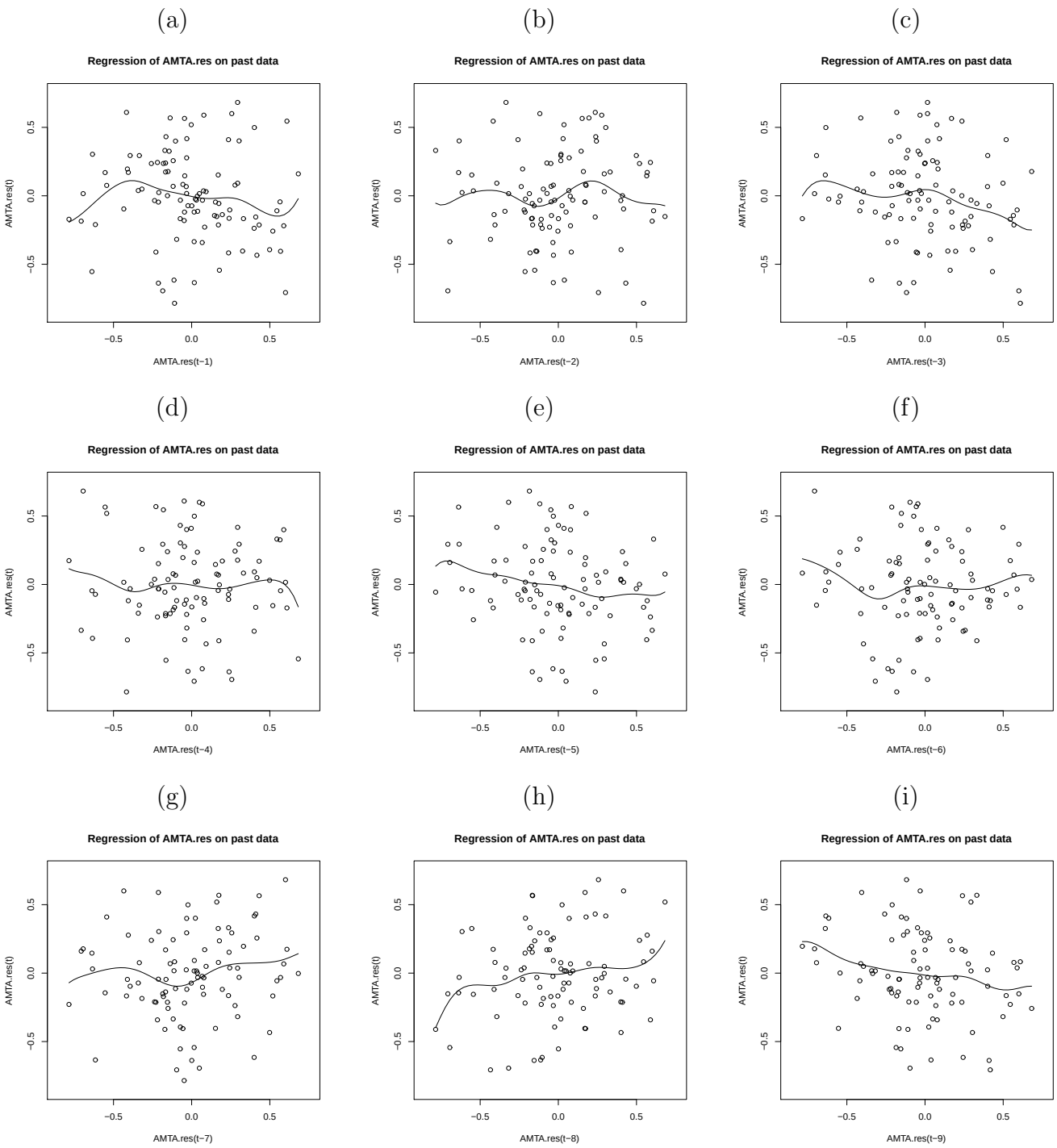


Figure 6: *Individual lag effects for AMTA.res, estimated by sm.autoregression in R: (a) lag=1; (b) lag=2; (c) lag=3; (d) lag=4; (e) lag=5; (f) lag=6; (g) lag=7; (h) lag=8; (i) lag=9.*

Table 4: The estimated coefficients and their standard error (s.e.) in the nonparametric model averaging with lags from 1 to 9 for the estimation sub-sample of AMTA.res:

	w_0	w_1	w_2	w_3	w_4
value	0.06507	0.87392	1.80548	1.13594	2.13717
s.e.	0.02551	0.34763	0.46953	0.37083	0.94460
	w_5	w_6	w_7	w_8	w_9
value	1.15743	1.05129	0.87480	0.72317	0.80995
s.e.	0.51662	0.40191	0.46996	0.35264	0.30814

We also considered the comparison with the additive AR(9) model fitting. The 9 component functions in the fitted additive AR(9) model are plotted in Figure 7, where X_i stands for the lag i variable of y_t. From Table 5, the mean of residual squares of the additive fitting "GAM" is 0.04372112, slightly smaller than that of our "NMA", in the estimation sample. However, the MSPE for the additive model "GAM" is 0.1333408, larger than that of our proposed "NMA", 0.1001797. This again shows that our approach is a useful tool to uncover nonlinear lag effects with simple calculations and performs basically well.

Table 5: The comparison of mean of residual squares (MRS) and mean squared prediction error (MSPE) for the linear "AR", "NMA" and "GAM".

	linear "AR"	"NMA"	"GAM"
MRS (in sample)	0.08261	0.04652649	0.04372112
MSPE (out of sample)	0.1136318	0.1001797	0.1333408

6.2.2 FTSE100 real data

Our second real data set is a one minute financial data set of 2000 observations from the FTSE100 index, consisting 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

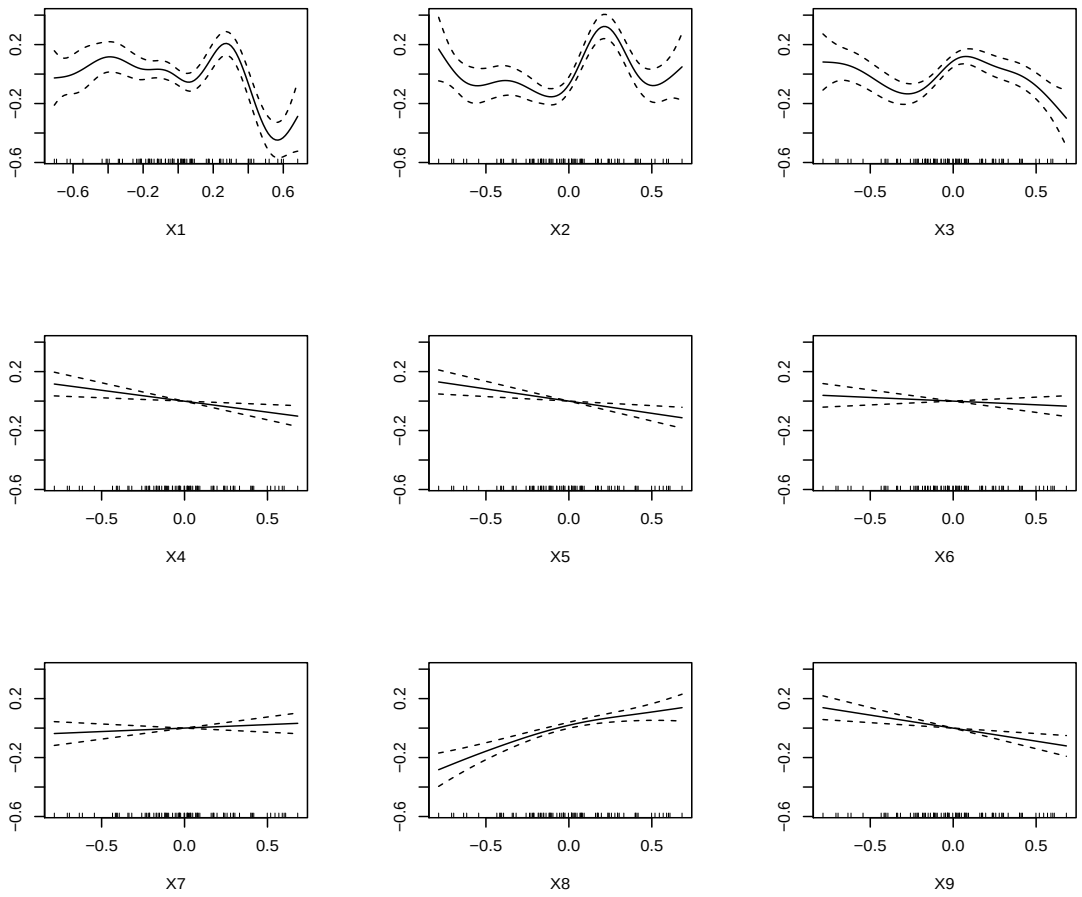


Figure 7: *Additive modelling for the estimation sub-sample of AMTA.res: The solid line for loess based estimate of each additive component and the dashed lines for the 95% pointwise confidence interval.*

from 22–28 June 2012. We are concerned with the relationship of 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 8.

Differently from the short lags in the above AMTA 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) 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 $X_t = (V_{t-1}, \cdots, V_{t-60}, R_{t-1}, \cdots, R_{t-60})^\intercal$ or $X_t = (V_{t-1}, \cdots, V_{t-60}, R_{t-1}, \cdots, R_{t-60}, v_{t-1}, \cdots, v_{t-60})^\intercal$ to predict $Y_t = V_{t+k}$, for $k = 1, 2, \cdots, 10$. In the last real example, we have shown some of the advantages of our proposed "NMA" method in comparison with the additive model. Therefore here we only look at the comparison of the "NMA" forecasting method with the linear forecasting. 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 8) 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 "NMA" and the Linear forecasting, respectively, for k from 1 to 10, which are plotted in Figure 9, where (a) corresponds to the case of $X_t = (V_{t-1}, \cdots, V_{t-60}, R_{t-1}, \cdots, R_{t-60})^\intercal$, and (b) to $X_t = (V_{t-1}, \cdots, V_{t-60}, R_{t-1}, \cdots, R_{t-60}, v_{t-1}, \cdots, v_{t-60})^\intercal$. Clearly, overall our proposed "NMA" is 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 first-stage nonparametric kernel smoothing has been developed to estimate the weight parameters involved in the approximation. Asymptotic properties for both the parametric and nonparametric estimators have been established under mild conditions. In particular, the parametric estimator

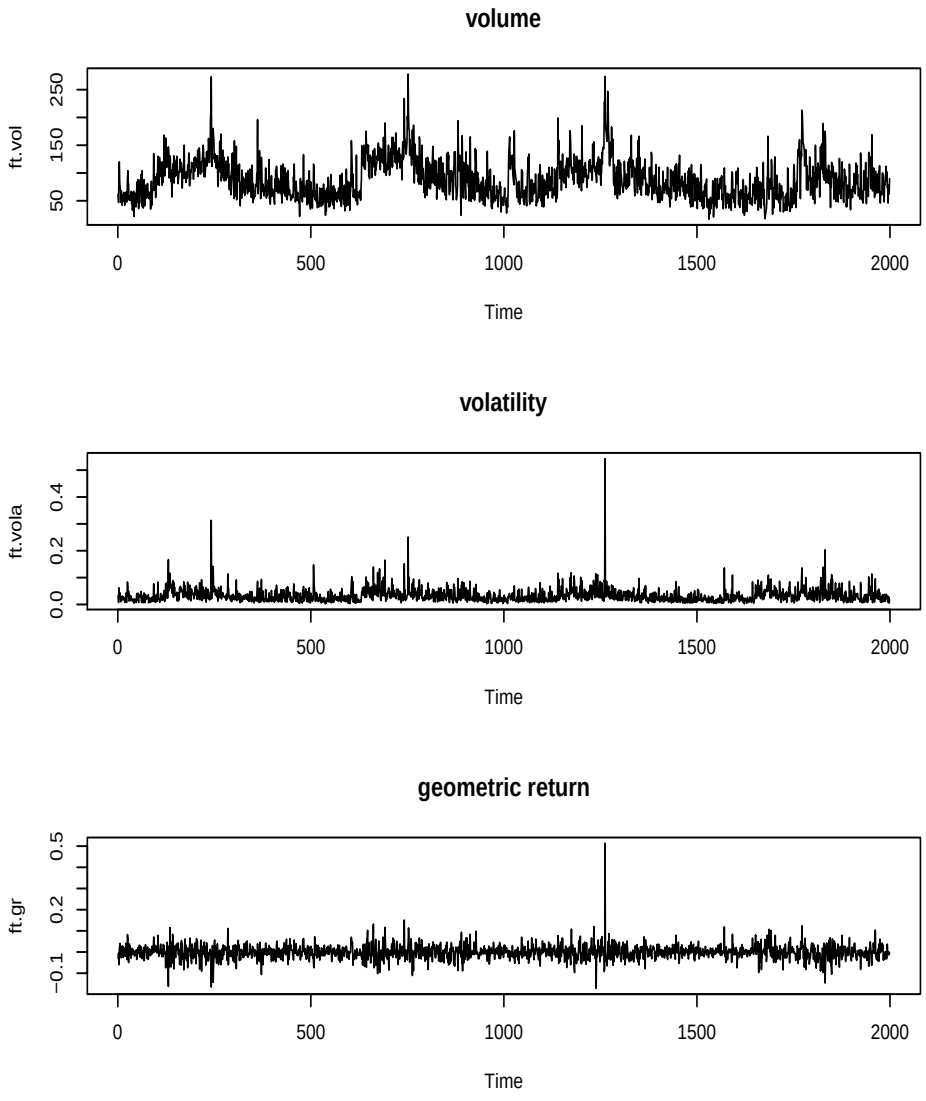
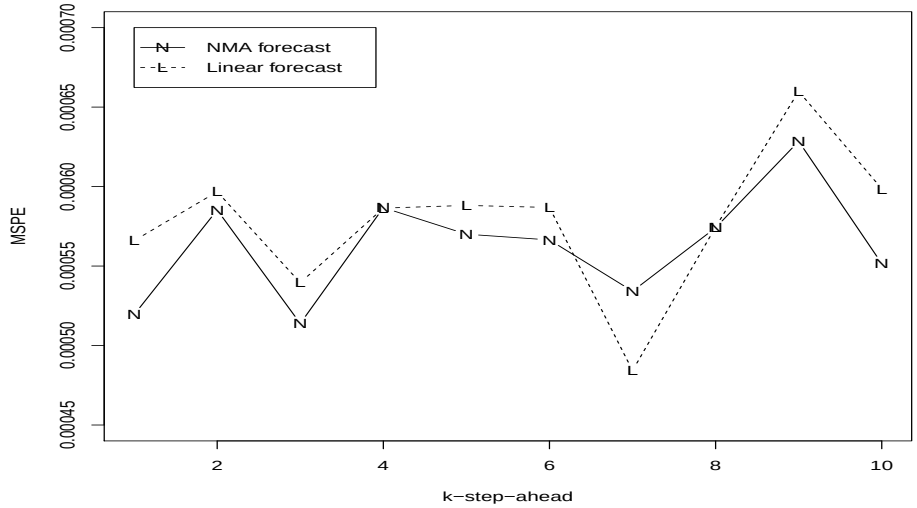


Figure 8: *The time series plots of the volume v_t, the geometric return R_t and the volatility V_t.*

(a)



(b)

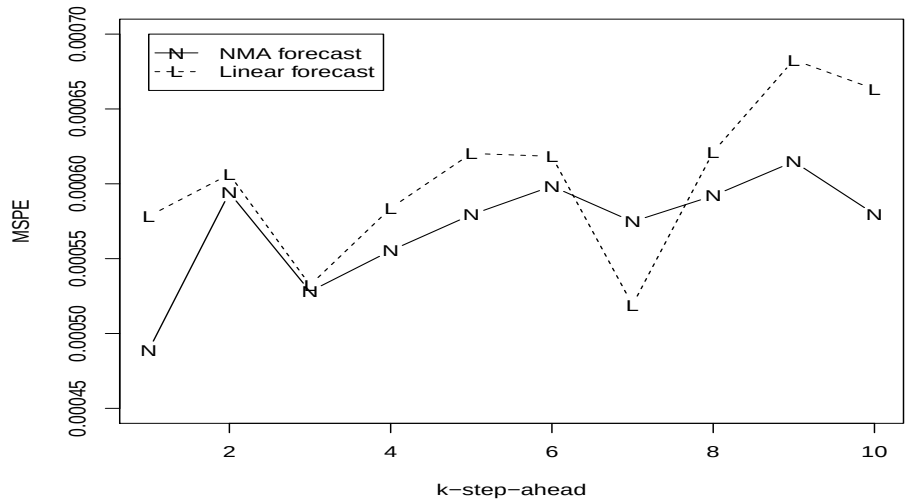


Figure 9: *Comparison of the mean squares prediction error (MSPE) of the k-step-ahead prediction of the volatility against k for the NMA and the Linear forecasting methods: (a) Case of no lagged volumes used, (b) Case of the lagged volumes used.*

is shown to be asymptotically normal with root-n rate of convergence when the dimension of the covariates is finite, 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 estimation and forecasting issues in time series analysis. Our methods and results are further augmented by a simulation study and two applications.

Appendix

In this appendix, we first give the detailed proofs of the asymptotic results stated in Sections 4, and then provide some technical lemmas. In the sequel, C denotes a positive constant, whose value may change from line to line.

A Proofs of the main results

Proof of Theorem 4.1. Recall that

$$\eta_t = Y_t - \sum_{j=1}^{d} w_{o,j} m_j(X_{tj})$$

and

$$\eta_{tj} = Y_t - \mathsf{E}(Y_t|X_{tj}) = Y_t - m_j(X_{tj}).$$

Furthermore, define

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

Observe that

$$\begin{aligned} \widehat{w} &= \left(\widehat{\mathcal{M}}^{\intercal}\widehat{\mathcal{M}}\right)^{-1} \widehat{\mathcal{M}}^{\intercal}\mathcal{Y} \\ &= \left(\widehat{\mathcal{M}}^{\intercal}\widehat{\mathcal{M}}\right)^{-1} \widehat{\mathcal{M}}^{\intercal}(\mathcal{M}w_o^* + \eta) \\ &= w_o^* + \left(\widehat{\mathcal{M}}^{\intercal}\widehat{\mathcal{M}}\right)^{-1} \widehat{\mathcal{M}}^{\intercal}(\mathcal{M} - \widehat{\mathcal{M}})w_o^* + \left(\widehat{\mathcal{M}}^{\intercal}\widehat{\mathcal{M}}\right)^{-1} \widehat{\mathcal{M}}^{\intercal}\eta \\ &=: w_o^* + \Pi_{n1} + \Pi_{n2}. \end{aligned}$$

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

$$\widehat{m}_j(x_j) - m_j(x_j) = \frac{\sum_{t=1}^n \eta_{tj} K\big(\frac{X_{tj}-x_j}{h_j}\big)}{\sum_{t=1}^n K\big(\frac{X_{tj}-x_j}{h_j}\big)} + \frac{\sum_{t=1}^n m_j(X_{tj}) K\big(\frac{X_{tj}-x_j}{h_j}\big)}{\sum_{t=1}^n K\big(\frac{X_{tj}-x_j}{h_j}\big)} - m_j(x_j).$$

By Taylor's expansion and Lemma B.1 in Appendix B, uniformly for $x_j \in \Omega_j$, we have

$$\frac{\sum_{t=1}^n m_j(X_{tj}) K\big(\frac{X_{tj}-x_j}{h_j}\big)}{\sum_{t=1}^n K\big(\frac{X_{tj}-x_j}{h_j}\big)} - m_j(x_j) \overset{P}{\sim} \frac{(m*f)_j^{(\gamma)}(x_j)}{f_j(x_j)} \mu_\gamma h_j^\gamma, \tag{A.1}$$

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

$$\widehat{m}_j(x_j) - m_j(x_j) = \frac{\sum_{t=1}^n \eta_{tj} K\big(\frac{X_{tj}-x_j}{h_j}\big)}{\sum_{t=1}^n K\big(\frac{X_{tj}-x_j}{h_j}\big)} + \frac{m*f_j^{(\gamma)}(x_j) h_j^\gamma \mu_\gamma}{f_j(x_j)}(1+o_P(1)). \tag{A.2}$$

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

$$\widehat{\mathcal{M}}^\intercal \widehat{\mathcal{M}} \overset{P}{\sim} \mathcal{M}^\intercal \mathcal{M} \overset{P}{\sim} \Lambda, \tag{A.3}$$

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

$$\Pi_{n1} \overset{P}{\sim} \big(\mathcal{M}^\intercal \mathcal{M}\big)^{-1} \Pi_{n3}, \tag{A.4}$$

where

$$\begin{aligned}
\Pi_{n3} &= \Big\{ \sum_{t=1}^n m_j(X_{tj}) \sum_{k=1}^d w_{o,k} \big[m_k(X_{tk}) - \widehat{m}_k(X_{tk}) \big] \Big\}_{j=1,\ldots,d}^\intercal \\
&= \Big\{ \sum_{t=1}^n m_j(X_{tj}) \sum_{k=1}^d w_{o,k} \Big[\frac{-\sum_{s=1}^n \eta_{sk} K\big(\frac{X_{sk}-X_{tk}}{h_k}\big)}{\sum_{s=1}^n K\big(\frac{X_{sk}-X_{tk}}{h_k}\big)} + O_P(h_k^\gamma) \Big] \Big\}_{j=1,\ldots,d}^\intercal \\
&= \Big\{ -\sum_{t=1}^n m_j(X_{tj}) \sum_{k=1}^d w_{o,k} \cdot \frac{\sum_{s=1}^n \eta_{sk} K\big(\frac{X_{sk}-X_{tk}}{h_k}\big)}{\sum_{s=1}^n K\big(\frac{X_{sk}-X_{tk}}{h_k}\big)} \Big\}_{j=1,\ldots,d}^\intercal + O_P(nh^\gamma) \\
&= \Big\{ -\sum_{s=1}^n \sum_{k=1}^d w_{o,k} \eta_{sk} \Big[\frac{1}{nh_k} \sum_{t=1}^n m_j(X_{tj}) f_k^{-1}(X_{tk}) K\big(\frac{X_{sk}-X_{tk}}{h_k}\big) \Big] \Big\}_{j=1,\ldots,d}^\intercal + O_P(nh^\gamma)
\end{aligned}$$

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

$$\frac{1}{nh_j}\sum_{t=1}^{n} m_j(X_{tj}) f_j^{-1}(X_{tj}) K\Big(\frac{X_{sj} - X_{tj}}{h_j}\Big) = m_j(X_{sj}) + o_P(1). \tag{A.5}$$

If $k \neq j$, by Lemma B.1 again, we have

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

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

$$\begin{aligned}
\Pi_{n3} &= -\Big(\sum_{s=1}^{n}\sum_{k=1}^{d} w_{o,k}\eta_{sk}\beta_{jk}(X_{sk})\Big)_{j=1,\ldots,d}^{\intercal} + O_P(nh^{\gamma}) \\
&= -\Big(\sum_{t=1}^{n}\eta_{t1}^{*}, \ldots, \sum_{t=1}^{n}\eta_{td}^{*}\Big)^{\intercal} + O_P(nh^{\gamma}),
\end{aligned} \tag{A.7}$$

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

$$\Pi_{n1} = -\big(\mathcal{M}^{\intercal}\mathcal{M}\big)^{-1}\Big(\sum_{t=1}^{n}\eta_{t1}^{*}, \ldots, \sum_{t=1}^{n}\eta_{td}^{*}\Big)^{\intercal} + O_P(h^{\gamma}). \tag{A.8}$$

We next consider Π_{n2}. Observe that

$$\begin{aligned}
\Pi_{n2} &= \big(\widehat{\mathcal{M}}^{\intercal}\widehat{\mathcal{M}}\big)^{-1}\widehat{\mathcal{M}}^{\intercal}\eta \\
&= \big(\mathcal{M}^{\intercal}\mathcal{M}\big)^{-1}\widehat{\mathcal{M}}^{\intercal}\eta(1 + o_P(1)) \\
&= \big(\mathcal{M}^{\intercal}\mathcal{M}\big)^{-1}\big[\mathcal{M}^{\intercal}\eta + \big(\widehat{\mathcal{M}} - \mathcal{M}\big)^{\intercal}\eta\big](1 + o_P(1)).
\end{aligned}$$

Using (A.4) and Lemma B.3 in Appendix B, we can show that the leading term of Π_{n2} is $\big(\mathcal{M}^{\intercal}\mathcal{M}\big)^{-1}\mathcal{M}^{\intercal}\eta$ by noting that $nh^{2\gamma} = o(1)$ and $nh^{-\frac{p_0+2}{p_0}} v_2(r_n) \to 0$ and taking $M_n = o(\sqrt{nh})$. By letting $\eta'_{tj} = m_j(X_{tj})\eta_t$, we have

$$\Pi_{n2} = \big(\mathcal{M}^{\intercal}\mathcal{M}\big)^{-1}\Big(\sum_{t=1}^{n}\eta'_{t1}, \ldots, \sum_{t=1}^{n}\eta'_{td}\Big)^{\intercal}(1 + o_P(1)). \tag{A.9}$$

Then, by (A.8), (A.9) and Lemma B.4 in Appendix B, we can prove (4.6). Hence, the proof of Theorem 4.1 has been completed. ■

PROOF OF THEOREM 4.2. Observe that

$$\begin{aligned}\widehat{m}_w(x) - m_w(x) &= \sum_{j=1}^{d} \widehat{w}_{o,j}\widehat{m}_j(x_j) - \sum_{j=1}^{d} w_{o,j}m_j(x_j) \\ &= \Big[\sum_{j=1}^{d} \widehat{w}_{o,j}\widehat{m}_j(x_j) - \sum_{j=1}^{d} w_{o,j}\widehat{m}_j(x_j)\Big] \\ &\quad + \Big[\sum_{j=1}^{d} w_{o,j}\widehat{m}_j(x_j) - \sum_{j=1}^{d} w_{o,j}m_j(x_j)\Big] \\ &=: \Pi_{n4} + \Pi_{n5}. \end{aligned} \tag{A.10}$$

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 and following the proof of Theorem 3.1 in Lu and Linton (2007), we can also prove that

$$\sqrt{nh}\Pi_{n5} \xrightarrow{d} \mathsf{N}\left(0, \sigma_w^2\right). \tag{A.12}$$

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 η_t and η_{tj}, $1 \le j \le d_n$, be defined as in the proof of Theorem 4.1 by replacing d by d_n, and $\eta = \big(\eta_1, \ldots, \eta_n\big)^\intercal$. Define the $n \times d_n$ matrices by

$$\mathcal{M}_n = \begin{bmatrix} m_1(X_{11}) & \cdots & m_{d_n}(X_{1d_n}) \\ \vdots & \vdots & \vdots \\ m_1(X_{n1}) & \cdots & m_{d_n}(X_{nd_n}) \end{bmatrix}, \quad \widehat{\mathcal{M}}_n = \begin{bmatrix} \widehat{m}_1(X_{11}) & \cdots & \widehat{m}_{d_n}(X_{1d_n}) \\ \vdots & \vdots & \vdots \\ \widehat{m}_1(X_{n1}) & \cdots & \widehat{m}_{d_n}(X_{nd_n}) \end{bmatrix}.$$

Letting $\widehat{w}(n)$ and $w_o^*(n)$ be defined as $\widehat{w}$ and w_0^* with d replaced by d_n, we can easily show that

$$\begin{aligned} &\mathcal{A}_n\Sigma_n^{-1/2}(w)\big[\widehat{w}(n) - w_0^*(n)\big] \\ &= \mathcal{A}_n\Sigma_n^{-1/2}(w)\Big(\widehat{\mathcal{M}}_n^\intercal\widehat{\mathcal{M}}_n\Big)^{-1}\widehat{\mathcal{M}}_n^\intercal(\mathcal{M}_n - \widehat{\mathcal{M}}_n)w_o^*(n) \\ &\quad + \mathcal{A}_n\Sigma_n^{-1/2}(w)\Big(\widehat{\mathcal{M}}_n^\intercal\widehat{\mathcal{M}}_n\Big)^{-1}\widehat{\mathcal{M}}_n^\intercal\eta \\ &=: \Pi_{n6} + \Pi_{n7}, \end{aligned}$$

where, as in Theorem 4.3, $\mathcal{A}_n$ is a $p \times d_n$ matrix such that $\mathcal{A}_n\mathcal{A}_n^\intercal$ tends to a $p \times p$ nonnegative matrix $\mathcal{A}_*$, and $\Sigma_n(w) = \Lambda_n^{-1}\Sigma_n\Lambda_n^{-1}$, Λ_n and Σ_n are defined as Λ and Σ with d replaced by d_n.

We first derive the leading term of Π_{n6}. Using (A.1) in the proof of Theorem 4.1, we have, uniformly for $x \in \Omega$ and $1 \le j \le d_n$,

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

Note that

$$\widehat{\mathcal{M}}_n^\intercal \widehat{\mathcal{M}}_n = \mathcal{M}_n^\intercal \mathcal{M}_n + \big(\widehat{\mathcal{M}}_n - \mathcal{M}_n\big)^\intercal \mathcal{M}_n + \mathcal{M}_n^\intercal \big(\widehat{\mathcal{M}}_n - \mathcal{M}_n\big) + \big(\widehat{\mathcal{M}}_n - \mathcal{M}_n\big)^\intercal \big(\widehat{\mathcal{M}}_n - \mathcal{M}_n\big). \tag{A.14}$$

It is easy to see that

$$\frac{1}{n}\mathcal{M}_n^\intercal \mathcal{M}_n \overset{P}{\sim} \Lambda_n, \tag{A.15}$$

which implies that the smallest eigenvalue of $\frac{1}{n}\mathcal{M}_n^\intercal\mathcal{M}_n$ is larger than $\frac{1}{2}\lambda_{\min}$ in probability, where $\lambda_{\min} > 0$ is the smallest eigenvalue of Λ_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, $\mathcal{M}_n^\intercal\mathcal{M}_n$ is leading term of $\widehat{\mathcal{M}}_n^\intercal\widehat{\mathcal{M}}_n$, which leads to

$$\frac{1}{n}\widehat{\mathcal{M}}_n^\intercal \widehat{\mathcal{M}}_n \overset{P}{\sim} \Lambda_n. \tag{A.16}$$

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

$$\Pi_{n6} \overset{P}{\sim} \mathcal{A}_n \Sigma_n^{-1/2}(w) \Big(\mathcal{M}_n^\intercal \mathcal{M}_n\Big)^{-1} \Pi_{n8}, \tag{A.17}$$

where Π_{n8} is defined as Π_{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

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

where $\eta_{tj}^* = \sum_{k=1}^{d_n} w_{o,k}\eta_{tk}\beta_{jk}(X_{tk})$ 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

$$\Pi_{n6} = -\mathcal{A}_n \Sigma_n^{-1/2}(w)\big(\mathcal{M}_n^\intercal \mathcal{M}_n\big)^{-1}\Big[\sum_{t=1}^n \eta_{t1}^*, \ldots, \sum_{t=1}^n \eta_{td_n}^*\Big]^\intercal + O_P(\sqrt{d_n}h^\gamma). \tag{A.19}$$

We next consider Π_{n7}. Note that, by (A.16),

$$\begin{aligned}\Pi_{n7} &= \mathcal{A}_n\Sigma_n^{-1/2}(w)\big(\widehat{\mathcal{M}}_n^\intercal\widehat{\mathcal{M}}_n\big)^{-1}\widehat{\mathcal{M}}_n^\intercal\eta \\ &= \mathcal{A}_n\Sigma_n^{-1/2}(w)\big(\mathcal{M}_n^\intercal\mathcal{M}_n\big)^{-1}\Big[\mathcal{M}_n^\intercal\eta + \big(\widehat{\mathcal{M}}_n - \mathcal{M}_n\big)^\intercal\eta\Big](1+o_P(1)).\end{aligned}$$

As in the proof of Theorem 4.1, we can also show that the leading term of Π_{n7} is $\big(\mathcal{M}_n^\intercal\mathcal{M}_n\big)^{-1}\mathcal{M}_n^\intercal\eta$ by noting that $nd_nh^{2\gamma} = o(1)$ and $nd_nh^{-\frac{p_0+2}{p_0}}v_2(r_n) = o(1)$ and taking $M_n = o(\sqrt{nh/d_n})$. We thus have

$$\Pi_{n7} = \big(\mathcal{M}_n^\intercal\mathcal{M}_n\big)^{-1}\big(\sum_{t=1}^n\eta'_{t1},\ldots,\sum_{t=1}^n\eta'_{td_n}\big)^\intercal(1+o_P(1)). \tag{A.20}$$

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

B Technical lemmas

We next give some technical lemmas, which have been used to prove the main results in Appendix A. Define

$$W_{nj}(x_j) = \frac{1}{nh_j}\sum_{t=1}^n Y_tK\Big(\frac{X_{tj}-x_j}{h_j}\Big).$$

LEMMA B.1. *Suppose that the assumptions 1–4 are satisfied. Then, we have, uniformly for* $1\le j\le d$,

$$\sup_{x_j\in\Omega_j}\big|W_{nj}(x_j) - \mathsf{E}\left[W_{nj}(x_j)\right]\big| = O_P\left(\tau_n\right), \tag{B.1}$$

where $\tau_n = \sqrt{\frac{\log n}{nh}}$ *is defined as in Section 4.1.*

PROOF. The uniform consistency result (B.1) follows directly from Theorem 3.1 in Li *et al* (2012). ■

The next lemma, which can be found in Bradley (1983), will help us use coupling technique to get the order of the nonparametric type U-statistic, which is crucial for the proofs of Theorems 4.1 and 4.3.

LEMMA B.2. *Let* X *and* Y *be real-valued random variable and* $\mathbb{R}^p$*-valued random vector,* $p\ge 1$. *Assume that* $\mathsf{E}|X|^\gamma<\infty$ *and let* $0<\epsilon\le\mathsf{E}^{\frac{1}{\gamma}}|X|^\gamma$. *Then, there exists (after replacing the underlying*

probability space by a bigger one if necessary) a random variable X_* such that (i) X_* has the same distribution as X, and is independent of Y; and (ii)

$$\mathsf{P}\left(|X - X_*| \geq \epsilon\right) \leq 18\Big(\frac{\mathsf{E}^{\frac{1}{\gamma}}|X|^{\gamma}}{\epsilon}\Big)^{\frac{\gamma}{2+\gamma}} \alpha^{\frac{2\gamma}{2+\gamma}}(X, Y). \tag{B.2}$$

For $j = 1, \ldots, d$, define

$$\overline{U}_{nj} = \sum_{t=1}^{n} \sum_{s \neq t} \eta_t \eta_{sj} K\Big(\frac{X_{sj} - X_{tj}}{h_j}\Big), \tag{B.3}$$

$$\overline{U}_{nj}^{(r_n)} = \sum_{t=1}^{n} \sum_{s \neq t} \eta_t^{(r_n)} \eta_{sj}^{(r_n)} K\Big(\frac{X_{sj}^{(r_n)} - X_{tj}^{(r_n)}}{h_j}\Big), \tag{B.4}$$

where $\eta_t^{(r_n)}$ and $\eta_{tj}^{(r_n)}$ are defined as $Y_t^{(r_n)}$ and $X_t^{(r_n)}$ in Definition 1.1. By the definitions of η_t and η_{tj}, it is easy to prove that

$$\mathsf{E}\Big(|\eta_t - \eta_t^{(r_n)}|^{\nu} + \sum_{j=1}^{d} |\eta_{tj} - \eta_{tj}^{(r_n)}|^{\nu}\Big) \leq C v_{\nu}(r_n), \quad 0 \leq \nu \leq p_0, \tag{B.5}$$

where C is a positive constant. We next calculate the orders for both $\overline{U}_{nj}$ and $\overline{U}_{nj}^{(r_n)}$.

LEMMA B.3. *Suppose that the conditions of Theorem 4.1 are satisfied. Then, we have*

$$\max_{1 \leq j \leq d} \mathsf{E}\Big[\big(\overline{U}_{nj}^{(r_n)}\big)^2\Big] \leq C_* n^2 M_n^2 h, \tag{B.6}$$

where C_* is a positive constant and $M_n \geq n^{\zeta}$ for some $\zeta > 0$. Furthermore, we have

$$\max_{1 \leq j \leq d} \big|\overline{U}_{nj}\big| = O_P\left(n M_n h^{1/2} + n^2 h^{\frac{p_0 - 2}{2p_0}} v_2^{1/2}(r_n)\right). \tag{B.7}$$

PROOF. Observe that

$$\begin{aligned}
&\sum_{t=1}^{n} \sum_{s \neq t} \eta_t \eta_{sj} K\Big(\frac{X_{sj} - X_{tj}}{h_j}\Big) - \sum_{t=1}^{n} \sum_{s \neq t} \eta_t^{(r_n)} \eta_{sj}^{(r_n)} K\Big(\frac{X_{sj}^{(r_n)} - X_{tj}^{(r_n)}}{h_j}\Big) \\
&= \sum_{t=1}^{n} \sum_{s \neq t} \big(\eta_t - \eta_t^{(r_n)}\big) \eta_{sj} K\Big(\frac{X_{sj} - X_{tj}}{h_j}\Big) + \sum_{t=1}^{n} \sum_{s \neq t} \eta_t^{(r_n)} \big(\eta_{sj} - \eta_{sj}^{(r_n)}\big) K\Big(\frac{X_{sj} - X_{tj}}{h_j}\Big) \\
&=: \Xi_{n1}(j) + \Xi_{n2}(j).
\end{aligned} \tag{B.8}$$

Letting $B_n = h_j^{-\frac{1}{p_0}}$, by (B.5) and some standard calculations, we have

$$\begin{aligned}
\mathsf{E}\big[|\Xi_{n1}(j)|\big] &= \sum_{t=1}^{n}\sum_{s\neq t}\mathsf{E}\Big[|\eta_t-\eta_t^{(r_n)}||\eta_{sj}|K\Big(\frac{X_{sj}-X_{tj}}{h_j}\Big)\Big] \\
&= \sum_{t=1}^{n}\sum_{s\neq t}\mathsf{E}\Big[|\eta_t-\eta_t^{(r_n)}||\eta_{sj}|I\big(|\eta_{sj}|\le B_n\big)K\Big(\frac{X_{sj}-X_{tj}}{h_j}\Big)\Big] \\
&+ \sum_{t=1}^{n}\sum_{s\neq t}\mathsf{E}\Big[|\eta_t-\eta_t^{(r_n)}||\eta_{sj}|I\big(|\eta_{sj}|> B_n\big)K\Big(\frac{X_{sj}-X_{tj}}{h_j}\Big)\Big] \\
&= O\Big(n^2B_nv_2^{1/2}(r_n)h^{1/2}+n^2B_n^{-\frac{p_0-2}{2}}v_2^{1/2}(r_n)\Big) \\
&= O\Big(n^2h^{\frac{p_0-2}{2p_0}}v_2^{1/2}(r_n)\Big) \qquad \text{(B.9)}
\end{aligned}$$

uniformly for $1\le j\le d$, where $I(\cdot)$ is the indicator function. Analogously, we can also show that

$$\mathsf{E}\left(|\Xi_{n2}(j)|\right) = O\Big(n^2h^{\frac{p_0-2}{2p_0}}v_2^{1/2}(r_n)\Big) \qquad \text{(B.10)}$$

uniformly for $1\le j\le d$.

By (B.6), (B.8)–(B.10), we can prove (B.7).

We next give the detailed proof of (B.6). By standard calculation, we have

$$\begin{aligned}
\mathsf{E}\Big[\big(\overline{U}_{nj}^{(r_n)}\big)^2\Big] = \sum_{t_1=1}^{n}\sum_{t_2=1}^{n}\sum_{s_1\neq t_1}\sum_{s_2\neq t_2}\mathsf{E}\Big[&\eta_{t_1}^{(r_n)}\eta_{t_2}^{(r_n)}\eta_{s_1j}^{(r_n)}\eta_{s_2j}^{(r_n)} \\
&\times K\Big(\frac{X_{s_1j}^{(r_n)}-X_{t_1j}^{(r_n)}}{h_j}\Big)K\Big(\frac{X_{s_2j}^{(r_n)}-X_{t_2j}^{(r_n)}}{h_j}\Big)\Big]. \qquad \text{(B.11)}
\end{aligned}$$

Without loss of generality, we only consider the case of $t_1<t_2<s_1<s_2$. Let $\eta_{t_1*}^{(r_n)}$ be the random variable, which has the same distribution of $\eta_{t_1}^{(r_n)}$ but independent of $\Big(\eta_{t_2}^{(r_n)},\eta_{s_1j}^{(r_n)},\eta_{s_2j}^{(r_n)}\Big)^{\intercal}$. Letting $X=\eta_{t_1}^{(r_n)}$, $Y=\Big(\eta_{t_2}^{(r_n)},\eta_{s_1j}^{(r_n)},\eta_{s_2j}^{(r_n)}\Big)^{\intercal}$ and $X_*=\eta_{t_1*}^{(r_n)}$ in Lemma B.2, it is easy to show that

$$\mathsf{E}\Big[\eta_{t_1*}^{(r_n)}\eta_{t_2}^{(r_n)}\eta_{s_1j}^{(r_n)}\eta_{s_2j}^{(r_n)}K\Big(\frac{X_{s_1j}^{(r_n)}-X_{t_1j}^{(r_n)}}{h_j}\Big)K\Big(\frac{X_{s_2j}^{(r_n)}-X_{t_2j}^{(r_n)}}{h_j}\Big)\Big]=0. \qquad \text{(B.12)}$$

Letting

$$\Delta_{t_1,t_2,s_1,s_2}=\eta_{t_1}^{(r_n)}\eta_{t_2}^{(r_n)}\eta_{s_1j}^{(r_n)}\eta_{s_2j}^{(r_n)}-\eta_{t_1*}^{(r_n)}\eta_{t_2}^{(r_n)}\eta_{s_1j}^{(r_n)}\eta_{s_2j}^{(r_n)},$$

by (B.12), we have

$$\begin{aligned}\mathsf{E}\Big[\big(\overline{U}_{nj}^{(r_n)}\big)^2\Big] &= C\Big\{\sum_{t_1=1}^{n}\sum_{t_2=t_1+1}^{t_1+M_n}\sum_{s_1=t_2+1}^{n}\sum_{s_2=s_1+1}^{n}\mathsf{E}\Big[\eta_{t_1}^{(r_n)}\eta_{t_2}^{(r_n)}\eta_{s_1j}^{(r_n)}\eta_{s_2j}^{(r_n)}\\ &\quad\times K\Big(\frac{X_{s_1j}^{(r_n)}-X_{t_1j}^{(r_n)}}{h_j}\Big)K\Big(\frac{X_{s_2j}^{(r_n)}-X_{t_2j}^{(r_n)}}{h_j}\Big)\Big]\\ &\quad+\sum_{t_1=1}^{n}\sum_{t_2=t_1+M_n+1}^{n}\sum_{s_1=t_2+1}^{n}\sum_{s_2=s_1+1}^{n}\mathsf{E}\Big[\Delta_{t_1,t_2,s_1,s_2}\\ &\quad\times K\Big(\frac{X_{s_1j}^{(r_n)}-X_{t_1j}^{(r_n)}}{h_j}\Big)K\Big(\frac{X_{s_2j}^{(r_n)}-X_{t_2j}^{(r_n)}}{h_j}\Big)\Big]\Big\}\\ &=: \Xi_{n3}(j)+\Xi_{n4}(j).\end{aligned}\tag{B.13}$$

For $\epsilon>0$, let

$$\overline{\Delta}_{t_1,t_2,s_1,s_2}=\Delta_{t_1,t_2,s_1,s_2}I\big(|\eta_{t_1}^{(r_n)}-\eta_{t_1*}^{(r_n)}|<\epsilon\big)$$

and

$$\widetilde{\Delta}_{t_1,t_2,s_1,s_2}=\Delta_{t_1,t_2,s_1,s_2}I\big(|\eta_{t_1}^{(r_n)}-\eta_{t_1*}^{(r_n)}|\ge\epsilon\big).$$

It is easy to see that

$$\begin{aligned}\Xi_{n4}(j) &= C\Big\{\sum_{t_1=1}^{n}\sum_{t_2=t_1+M_n+1}^{n}\sum_{s_1=t_2+1}^{n}\sum_{s_2=s_1+1}^{n}\mathsf{E}\Big[\overline{\Delta}_{t_1,t_2,s_1,s_2}K\Big(\frac{X_{s_1j}^{(r_n)}-X_{t_1j}^{(r_n)}}{h_j}\Big)K\Big(\frac{X_{s_2j}^{(r_n)}-X_{t_2j}^{(r_n)}}{h_j}\Big)\Big]\\ &\quad+\sum_{t_1=1}^{n}\sum_{t_2=t_1+M_n+1}^{n}\sum_{s_1=t_2+1}^{n}\sum_{s_2=s_1+1}^{n}\mathsf{E}\Big[\widetilde{\Delta}_{t_1,t_2,s_1,s_2}K\Big(\frac{X_{s_1j}^{(r_n)}-X_{t_1j}^{(r_n)}}{h_j}\Big)K\Big(\frac{X_{s_2j}^{(r_n)}-X_{t_2j}^{(r_n)}}{h_j}\Big)\Big]\Big\}\\ &=: \Xi_{n5}(j)+\Xi_{n6}(j).\end{aligned}\tag{B.14}$$

Note that uniformly for $1\le j\le d$,

$$\Xi_{n5}(j)=O\big(\epsilon n^4h^2\big).\tag{B.15}$$

On the other hand, by Lemma B.2, we have, uniformly for $1\le j\le d$,

$$\Xi_{n6}(j)=O\Big(n^3\epsilon^{-\frac{p_0-2}{p_0}}\sum_{t=M_n}^{\infty}\alpha^{\frac{p_0-2}{p_0+1}}(t)\Big).\tag{B.16}$$

By (B.15), (B.16) and taking $\epsilon=(nh^2)^{-\frac{p_0}{2(p_0-1)}}\big[\sum_{t=M_n}^{\infty}\alpha^{\frac{p_0-2}{p_0+1}}(t)\big]^{\frac{p_0}{2(p_0-1)}}$, we have

$$\Xi_{n4}(j)=O\Big(n^4h^2\big(nh^2\big)^{-\frac{p_0}{2(p_0-1)}}\Big[\sum_{t=M_n}^{\infty}\alpha^{\frac{p_0-2}{p_0+1}}(t)\Big]^{\frac{p_0}{2(p_0-1)}}\Big)=o(n^2M_n^2h)\tag{B.17}$$

uniformly for $1 \le j \le d$, as the α-mixing coefficient decays at the geometric rate and $M_n \ge n^{\zeta}$, where $\zeta > 0$.

On the other hand, note that

$$\begin{aligned}\Xi_{n3}(j) = C\Big\{ &\sum_{t_1=1}^{n} \sum_{t_2=t_1+1}^{t_1+M_n} \sum_{s_1=t_2+1}^{n} \sum_{s_2=s_1+1}^{s_1+M_n} \mathsf{E}\Big[\eta_{t_1}^{(r_n)} \eta_{t_2}^{(r_n)} \eta_{s_1 j}^{(r_n)} \eta_{s_2 j}^{(r_n)} \\ &\times K\Big(\frac{X_{s_1 j}^{(r_n)} - X_{t_1 j}^{(r_n)}}{h_j}\Big) K\Big(\frac{X_{s_2 j}^{(r_n)} - X_{t_2 j}^{(r_n)}}{h_j}\Big)\Big] \\ &+ \sum_{t_1=1}^{n} \sum_{t_2=t_1+M_n}^{n} \sum_{s_1=t_2+1}^{n} \sum_{s_2=s_1+M_n+1}^{n} \mathsf{E}\Big[\eta_{t_1}^{(r_n)} \eta_{t_2}^{(r_n)} \eta_{s_1 j}^{(r_n)} \eta_{s_2 j}^{(r_n)} \\ &\times K\Big(\frac{X_{s_1 j}^{(r_n)} - X_{t_1 j}^{(r_n)}}{h_j}\Big) K\Big(\frac{X_{s_2 j}^{(r_n)} - X_{t_2 j}^{(r_n)}}{h_j}\Big)\Big]\Big\}.\end{aligned}$$

Similarly to the calculation of $\Xi_{n4}(j)$, we can also show that

$$\Xi_{n3}(j) = O\left(n^2 M_n^2 h\right) + o\left(n^2 M_n^2 h\right) \tag{B.18}$$

uniformly for $1 \le j \le d$.

Then, (B.6) follows from (B.13), (B.17) and (B.18). Then, we have completed the proof of Lemma B.3. ■

LEMMA B.4. *Suppose that the conditions of Theorem 4.1 are satisfied. Then, we have*

$$\frac{1}{\sqrt{n}} V_n \xrightarrow{d} \mathsf{N}\big(\mathbf{0}, \Sigma\big), \tag{B.19}$$

where Σ is defined in Section 4.1 and

$$V_n = \Big[\sum_{t=1}^{n} (\eta'_{t1} - \eta^*_{t1}), \ldots, \sum_{t=1}^{n} (\eta'_{td} - \eta^*_{td})\Big]^{\intercal}.$$

PROOF. For $1 \le j \le d$, let $\xi_{tj} = \eta'_{tj} - \eta^*_{tj}$ and

$$\xi_{tj}^{(r_n)} = m_j(X_{tj}^{(r_n)}) \eta_t^{(r_n)} - \sum_{k=1}^{d} w_{o,k} \eta_{tk}^{(r_n)} \beta_{jk}(X_{tk}^{(r_n)}).$$

Note that

$$\sum_{t=1}^{n} \xi_{tj} = \sum_{t=1}^{n} \big[\xi_{tj}^{(r_n)} - \mathsf{E}(\xi_{tj}^{(r_n)})\big] + \sum_{t=1}^{n} \big(\xi_{tj} - \xi_{tj}^{(r_n)}\big) + \sum_{t=1}^{n} \big[\mathsf{E}(\xi_{tj}) - \mathsf{E}(\xi_{tj}^{(r_n)})\big] =: \sum_{k=7}^{9} \Xi_{nk}(j). \tag{B.20}$$

By Definition 4.1 and the fact that $nv_2(r_n) = o(1)$ implied by (4.5), we can prove that

$$\Xi_{n8}(j) + \Xi_{n9}(j) = nv_2^{1/2}(r_n) = o_P(\sqrt{n}) \tag{B.21}$$

uniformly for $1 \leq j \leq d$.

By (B.21), we have

$$V_n = \Big\{ \sum_{t=1}^{n} \big[\xi_{t1}^{(r_n)} - \mathsf{E}(\xi_{t1}^{(r_n)})\big], \ldots, \sum_{t=1}^{n} \big[\xi_{td}^{(r_n)} - \mathsf{E}(\xi_{td}^{(r_n)})\big] \Big\}^{\intercal} + o_P(\sqrt{n}) = V_n^{(r_n)} + o_P(\sqrt{n}). \tag{B.22}$$

Then, by Theorem 1.7 in Bosq (1998), we have

$$\frac{1}{\sqrt{n}} V_n^{(r_n)} \xrightarrow{d} \mathsf{N}\big(\mathbf{0}, \Sigma(r_n)\big), \tag{B.23}$$

where

$$\Sigma(r_n) = \sum_{k=-\infty}^{\infty} \mathsf{Cov}(\xi_0^{(r_n)}, \xi_k^{(r_n)}), \quad \xi_t^{(r_n)} = \Big(\xi_{t1}^{(r_n)}, \ldots, \xi_{td}^{(r_n)}\Big)^{\intercal}.$$

Observe that

$$\sum_{k=1}^{\infty} \mathsf{Cov}(\xi_0^{(r_n)}, \xi_k^{(r_n)}) = \sum_{k=1}^{M_n^*} \mathsf{Cov}(\xi_0^{(r_n)}, \xi_k^{(r_n)}) + \sum_{k=M_n^*+1}^{\infty} \mathsf{Cov}(\xi_0^{(r_n)}, \xi_k^{(r_n)}), \tag{B.24}$$

where $M_n^* = n^{\frac{1}{2}} h^{-\frac{p_0+2}{2p_0}}$.

As the α-mixing coefficient decays at the geometric rate, we can prove that

$$\sum_{k=M_n^*+1}^{\infty} \mathsf{Cov}(\xi_0^{(r_n)}, \xi_k^{(r_n)}) = o(1), \tag{B.25}$$

as $r_n = o(M_n^*)$ by the second term in Assumption 4 (ii). Meanwhile, by Definition 1.1 and (4.5) in Assumption 4, we can also prove that

$$\begin{aligned} \sum_{k=1}^{M_n^*} \mathsf{Cov}(\xi_0^{(r_n)}, \xi_k^{(r_n)}) &= \sum_{k=1}^{M_n^*} \mathsf{Cov}(\xi_0, \xi_k) + O\left(n^{1/2} h^{-\frac{p_0+2}{2p_0}} v_2^{1/2}(r_n)\right) \\ &= \sum_{k=1}^{\infty} \mathsf{Cov}(\xi_0, \xi_k) + o\left(1\right). \end{aligned} \tag{B.26}$$

Similarly, we also have

$$\mathsf{Var}(\xi_0^{(r_n)}) = \mathsf{Var}(\xi_0) + o(1), \quad \sum_{k=-\infty}^{-1} \mathsf{Cov}(\xi_0^{(r_n)}, \xi_k^{(r_n)}) = \sum_{k=-\infty}^{-1} \mathsf{Cov}(\xi_0, \xi_k) + o\left(1\right),$$

which together with (B.24)–(B.26), lead to

$$\Sigma(r_n) \to \Sigma, \quad \text{as } n \to \infty.$$

Then, the proof of Lemma B.4 is completed. ■

References

[1] Andrews, D. W. K., 1995. Nonparametric kernel estimation for semiparametric models. *Econometric Theory* 11, 560–596.

[2] Billingsley, P., 1968. *Convergence of Probability Measures*. Wiley, New York.

[3] Bradley, R., 1983. Approximation theorems for strongly mixing random variables. *Michigan Mathematical Journal* 60, 69–81.

[4] Bosq, D., 1998. *Nonparametric Statistics for Stochastic Processes: Estimation and Prediction*. Springer.

[5] Chen, X., Ghysels, E., 2010. News-Good or Bad- and its impact on volatility predictions over multiple horizons. *Review of Financial Studies* 24, 46–81

[6] Dehling, H., Wendler, M., 2010. Central limit theorem and the bootstrap for U–statistics of strongly mixing data. *Journal of Multivariate Analysis* 101, 126–137.

[7] Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96, 1348–1360.

[8] Fan, J., Peng, H., 2004. Nonconcave penalized likelihood with a diverging number of parameters. *Annals of Statistics* 32, 928–961.

[9] Fan, J., Yao, Q., 2003. *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer, New York.

[10] Frank, I. E., Friedman, J. H., 1993. A statistical view of some chemometrics regression tools. *Technometrics* 35, 109–148.

[11] Fu, W., 1998. Penalized regression: the bridge versus LASSO. *Journal of Computational and Graphical Statistics* 7, 397–416.

[12] Hastie, T. J., Tibshirani, R. J., 1990. *Generalized Additive Models*. Chapman & Hall/CRC .

[13] Hansen, B. E., 2007. Least Squares Model Averaging. *Econometrica* 75, 1175–1189.

[14] Hong, Y., 2000. Generalized spectral tests for serial dependence. *Journal of the Royal Statistical Society Ser. B* 62, 557–574

[15] Lam, C., Fan, J., 2008. Profile-kernel likelihood inference with diverging number of parameters. *Annals of Statistics* 36, 2232–2260.

[16] Li, D., Lu, Z., Linton, O., 2012. Local linear fitting under near epoch dependence: uniform consistency with convergence rates. Forthcoming in *Econometric Theory.*

[17] Liang, H., Zou, G. H., Wan, A. T. K., Zhang, X. Y., 2011. Optimal weight choice for frequentist model average estimators. *Journal of the American Statistical Association* 106, 1053–1066.

[18] Lin, Z., 2004. Strong near epoch dependence. *Science in China* (Series A) 47, 497–507.

[19] Ling, S., 2007. Testing for change points in time series models and limiting theorems for NED sequences. *Annals of Statistics* 35, 1213–1237.

[20] Linton, O., 2000. Efficient estimation of generalized additive nonparametric regression models. *Econometric Theory* 16, 502–523.

[21] Linton, O. B., Mammen, E., 2005. Estimating semiparametric ARCH(∞) models by kernel smoothing methods. *Econometrica* 73, 771–836.

[22] Linton, O. B., Mammen, E., 2008. Nonparametric transformation to white noise. *Journal of Econometrics* 141, 241–264.

[23] Linton, O. B., Sancetta, A., 2009. Consistent estimation of a general nonparametric regression function in time series. *Journal of Econometrics* **152**, 70–78.

[24] Lu, Z., 2001. Asymptotic normality of kernel density estimators under dependence. *Annals of the Institute of Statistical Mathematics* 53, 447–468.

[25] Lu, Z., Linton, O., 2007. Local linear fitting under near epoch dependence. *Econometric Theory* 23, 37–70.

[26] Mammen, E., Linton, O., Nielsen, J. 1999. The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Annals of Statistics* 27, 1443–1490.

[27] McLeish, D. L., 1975a. A maximal inequality and dependent strong laws. *Annals of Probability* 3, 826–836.

[28] McLeish, D. L., 1975b. Invariance principles for dependent variables. *Zeitschrift fur Wahrscheinlichskeittheorie und verwandte Gebiete* 32, 165–178.

[29] McLeish, D. L., 1977. On the invariance principle for nonstationary mixingales. *Annals of Probability* 5, 616–621.

[30] Nielsen, J., Linton, O., 1998. An optimization interpretation of integration and back–fitting estimators for separable nonparametric models. *Journal of the Royal Statistical Society, Series B* 60, 217–222.

[31] Stone, C. J., 1980. Optimal rates of convergence for nonparametric estimators. *Annals of Statistics* 8, 1348–1360.

[32] Teräsvirta, T., Tjøstheim, D., Granger, C., 2010. *Modelling Nonlinear Economic Time Series*. Oxford University Press.

[33] Tibshirani, R. J., 1996. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Ser. B* 58, 267–288.

[34] Tibshirani, R. J., 1997. The LASSO method for variable selection in the Cox model. *Statistics in Medicine* 16, 385–395.

[35] Tong, H., 1990. *Non-linear Time Series: A Dynamical System approach*. Oxford University Press, Oxford.

[36] Wand, M. P., Jones, M. C., 1995. *Kernel Smoothing*. Chapman and Hall.