Semiparametric Averaging of Nonlinear Marginal Logistic Regressions and Forecasting for Time Series Classification

Rong Peng, Zudi Lu

School of Mathematical Sciences, and Southampton Statistical Science Research Institute, University of Southampton, Southampton, SO17 1BJ, United Kingdom

ARTICLE INFO

Keywords:
Binary time series classification
Forecasting
Model average
MAMaLoR
Logistic marginal regression
Semi-parametric likelihood estimation

ABSTRACT

Binary classification is an important issue in many applications but mostly studied for independent data in the literature. A binary time series classification is investigated by proposing a semiparametric procedure named "Model Averaging nonlinear MArginal LOgistic Regressions" (MAMaLoR) for binary time series data based on the time series information of predictor variables. The procedure involves approximating the logistic multivariate conditional regression function by combining low-dimensional non-parametric nonlinear marginal logistic regressions, in the sense of Kullback-Leibler distance. A time series conditional likelihood method is suggested for estimating the optimal averaging weights together with local maximum likelihood estimations of the nonparametric marginal time series logistic (auto)regressions. The asymptotic properties of the procedure are established under mild conditions on the time series observations that are of β-mixing property. The procedure is less computationally demanding and can avoid the "curse of dimensionality" for, and be easily applied to, high dimensional lagged information based nonlinear time series classification forecasting. The performances of the procedure are further confirmed both by Monte-Carlo simulation and an empirical study for market moving direction forecasting of the financial FTSE 100 index data.

1. Introduction

Time series data lagged information has been useful for forecasting of future. Traditionally, for continuous-valued time series data, ARIMA based analysis is well developed and applied (c.f., Box, Jenkins, Reinsel & Ljung (2015)). Further development of nonlinear and nonparametric analysis of that kind of time series data can be found in Tong (1990), Fan & Yao (2003), Gao (2007) and Terasvirta, Tjostheim, Granger et al. (2010) for comprehensive reviews. Particularly, curse of dimensionality is a common challenging issue when faced a large number of time series lagged observations. Various semiparametric models are hence developed, which however usually involve expensive computations (c.f., the above-mentioned references). For more recent applications to multivariate time series under spatial and machine learning settings, the readers are referred to Al-Sulami, Jiang, Lu & Zhu (2017) and Hofert, Prasad & Zhu (2021) on the related issues. Alternatively, Li, Linton & Lu (2015) have recently introduced a novel procedure for forecasting the unknown future by conditional time series regression with high-dimensional time series lagged predictor vector, namely the Model Averaging MArginal Regressions (MAMAR). This is a very flexible procedure for time series forecasting based on the idea of model averaging the low-dimensional marginal forecasts. See also Chen, Li, Linton & Lu (2016, 2018) for more recent developments on the approach under continuous valued time series response.

However, in many situations of practical time series forecasting, such continuous response based procedure is not always adequate. In this paper we are concerned with binary valued time series classification forecasting. Observations like the market price moving (up/down) direction forecasting and the default/non-default credit scoring classification are actually discretely binary-valued. Binary data is a kind of important data with logistic regression analysis developed popularly for many applications though mostly under independent data in the literature (c.f., Cox & Snell (1989)). Our aim in this paper is to suggest a novel semiparametric procedure, named "Model Averaging nonlinear MArginal LOgistic Regressions" (MAMaLoR) for binary time series classification based on the information of a large number of lagged predictors, by extending the MAMAR idea of Li et al. (2015) to binary-valued time series nonlinear classification. This is motivated by the needs of wide practical applications, such as the financial examples mentioned. We are aware that such binary-valued time series data exist in wide applications beyond finance, though the financial application is particularly examined in this paper. Indeed, binary classification has been thought of as one of the most important problems in machine learning and statistics (c.f., Ryabko & Mary (2013)).

Email addresses and Correspondence to: rp2e13@soton.ac.uk (R. Peng), Z.Lu@soton.ac.uk (Z. Lu)

Within the discrete-valued time series models, linear autoregression technique is very popular. The history of analysing and modelling discrete-valued time series by a linear structure goes back to Jacobs & Lewis (1978), who proposed the DARMA (discrete mixed autoregressive-moving average) process. However its long term foresting performance is not as good as expected. McKenzie (1985) has alternatively proposed the INARMA (Integer-valued autoregressive-moving average) model, which is still well applied even today. Further developments include Waller, Carlin, Xia & Gelfand (1997) on hierarchica dynamic generalized linear mixed model for spatial time series problems, and Shephard (1995) on generalised linear autoregressive moving average model (GLARMA) applied in many different fields such as Rydberg & Shephard (2003) and Liesenfeld, Nolte & Pohlmeier (2006) in financial modelling and Turner, Hayen, Dunsmuir & Finch (2011) and Buckley & Bulger (2012) in epidemiological assessments and clinical management. Similarly, an Integer-valued GARCH model (INGARCH) has been proposed by Ferland, Latour & Oraichi (2006) in the spirit of the generalised autoregressive conditional heteroskedastic model (GARCH). In addition, the general latent-based time series models including the binary case are proposed by Davis & Wu (2009) (Experiment 2 on Page 743) and de Oliveira Maia, Barreto-Souza, de Souza Bastos & Ombao (2021) (Subsection 2.3). For a comprehensive review on the related developments, the reader is referred to Davis, Dunsmuir & Wang (1999) , Davis, Holan, Lund & Ravishanker (2016) and the references therein.

Though linearity is widely adopted in the literature, it may often be too strong to be appreciated when dealing with unknown data. The assumptions made on linear parametric relationship may be incorrect if we don't have prior knowledge about the true relationship between the predictors and the response. Indeed, in the case of time series classification and forecasting, the influences of the predictor variables and their lags on the response are usually of unknown forms. Differently from the parametric discrete-valued time series models above, in this paper, we will therefore suggest utilising nonparametric method, where the estimation of conditional regression functions is data driven. Here, in our proposed MAMaLoR procedure for binary time series classification, it involves approximating the logistic multivariate conditional regression function by combining low-dimensional nonlinear marginal logistic regressions which will be estimated non-parametrically in the first step of our procedure. A popular nonparametric approach in the literature is local fitting or kernel smoother of unknown functions (c.f., Fan, Farmen & Gijbels (1998a)), which can be estimated via technique of either maximum likelihood or least square method. Differently from Li et al. (2015), for our binary time series data, maximum likelihood method is preferred for nonparametric local linear fitting of the low-dimensional conditional marginal logistic regressions. The idea of maximum likelihood local fitting can be traced back to Tibshirani & Hastie (1987) and Fan & Gijbels (1995) for independent and identically distributed (i.i.d.) data, and Fan & Yao (1998) extending to stochastic regression. We will apply the maximum likelihood local fitting of the conditional marginal logistic regressions with the uniform consistency in the time series setting, which is required in the second step of combining those marginal logistic regressions for classification forecasting in our MAMaLoR procedure. Hence, in this paper, we will consider the maximum likelihood local fitting method under the data dependence of a so-called β-mixing conditions. For a more detailed discussion on β-mixing conditions, the reader is referred to Doukhan, Massart & Rio (1995) [Section2.4]. Theoretically, we will establish the asymptotic properties for our MAMaLoR procedure under β-mixing conditions.

Another advantage with our MAMaLoR procedure to be noted is that it overcomes the so-called "curse of dimensionality" (c.f., Seifert & Gasser (1996)), when a large number of time series lagged predictors are taken account of, leading to high dimensional conditional logistic regression functions. For multivariate nonparametric models with the increase of dimension d, it is well known that the performance may become worse or even useless (when d is beyond 2) as the sample size is required to increase exponentially to get the same quality of estimation for one dimensional function. In our MAMaLoR procedure, we consider combining low-dimensional marginal non-parametric nonlinear logistic regressions, and hence "curse of dimensionality" is flexibly avoided for time series binary classification similarly to that for the regression in Li et al. (2015). This is different than perhaps it initially looks when compared to the popular semiparametric generalised additive model (GAM) (Hastie & Tibshirani, 1987). When we only consider the one-dimensional marginal non-parametric logistic regressions for combination, our MAMaLoR shares a similar model form as a special case of GAM, but the GAM still suffers from heavier computational costs and other deficiencies in forecasting due to possible overfitting in particular in the case of small samples but with a relatively large number of time series lagged predictors, while these difficulties are more easily avoided in MAMaLoR. In addition, the low-dimensional marginal non-parametric logistic regressions could also be two-dimensional for combination in our MAMaLoR, where it is not of a GAM form (see more discussion on this in Section 2 below). We will also show in the data examples that our MAMaLoR procedure is not only easy to implement, but also works better in classification forecasting than GAM.

The structure of the rest of the paper is as follows: In Section 2, we provide the basic ideas on the proposed MAMaLoR procedure. Estimations for the MAMaLoR procedure with asymptotic properties established under β-mixing properties are given in Section 3. In Section 4 the numerical examples including a simulation and an application to forecasting the market price moving direction of FTSE 100 data will be demonstrated. Section 5 gives the conclusion. All the proofs will be relegated to an Appendix.

2. Model averaging marginal nonlinear logistic regressions

We are concerned with the binary classification forecasting. Let (Y_t, X_t^T) be a stationary time series process with Y_t the response of binary values of 0 and 1 at time t and $X_t = (x_{1t}, ..., x_{dt})^T$ a d-dimensional random vector representing the available information up to time $t-1$, where the components of X_t may involve the concerned time series predictor variables including lagged ones so that the dimension d may be rather large as in Li et al. (2015) in practice.

In general, we denote by I_{t-1} for all the information up to time $t-1$ about time series Y_t. So the regression problem is to estimate the conditional probability for classification forecasting:

$$p_t = P(Y_t = 1|I_{t-1}). \tag{1}$$

Because of the curse of dimensionality, it is well known that a direct nonparametric estimation of p_t performs very poor. We suggest the semiparametric procedure, Model Averaging nonlinear MArginal LOgistic Regressions (MAMaLoR), for binary time series classification by extending the MAMAR idea of Li et al. (2015), consisting of two steps as follows.

First, we would like to look at the marginal foresting effects based on part of the available information, say each component, of X_t. Then define the marginal forecasting probability based on the jth component (x_{jt}) as follows:

$$p_{jt} = P(Y_t = 1|x_{jt}),\ j = 1, ..., d. \tag{2}$$

A popular idea to model the conditional probability p_{jt} is by logistic regression. If we let F be the logistic cumulative distribution function(c.d.f), i.e., $F(u) = \frac{e^u}{1+e^u}$, then the marginal nonparametric logistic regression is

$$logit(p_{jt}) \equiv \log \frac{p_{jt}}{1-p_{jt}} = f_j(x_{jt}), \tag{3}$$

where $f_j(x_{jt})$ can be a nonlinear function of x_{jt}, and we hence have:

$$p_{jt} = F(f_j(x_{jt})). \tag{4}$$

Our second step is to combine the marginal logistic regressions together with a constant to approximate our concerned p_t in (12) by using the idea of model average as follows:

$$\begin{aligned} logit(p_t) &\approx \alpha_0 + \alpha_1 logit(p_{1t}) + ... + \alpha_d logit(p_{dt}) \\ &= \alpha_0 + \alpha_1 f_1(x_{1t}) + ... + \alpha_d f_d(x_{dt}) \equiv f_t^{MA}, \end{aligned} \tag{5}$$

where $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, ..., \alpha_d)$ is the vector of unknown coefficients. Indeed, the direct motivation for equation (5) comes from the model averaging by combining the easily estimable marginal logit forecasts to approximate the high-dimensional logit forecast that is hard to be well estimated due to curse of dimensionality for a relatively large d, so equation (5) represents an approximation, rather than an exact equality, similar to that in Li et al. (2015).

This can be seen as a model average as the $\boldsymbol{\alpha}$ can be seen as the weights assigned to different marginal estimations (c.f., Li et al. (2015)). Here we use the affine combination in equation (5) because it is flexible and easy to apply for classification forecasting and also much less overfitting than the GAM for forecasting in application. These advantages are similar to those in Li et al. (2015) with regression forecasting.

Let $F^{-1}(\cdot)$ be the inverse function of $F(\cdot)$. Then (5) can alternatively be expressed as

$$F^{-1}(p_t) = \log(\frac{p_t}{1-p_t}) \approx \alpha_0 + \sum_{j=1}^{d} \alpha_j F^{-1}(E(Y_t|x_{jt})), \tag{6}$$

where $E(Y_t|x_{jt}) = P(Y_t = 1|x_{jt}) = p_{jt}$. Therefore our (6) can be seen as a logit transformed extension of the MAMAR procedure of Li et al. (2015), in which $E(Y_t|I_{t-1})$ is approximated by $\alpha_0 + \sum_{j=1}^{d} \alpha_j E(Y_t|x_{jt})$ in terms of $\mathcal{L}_2$ distance, that $E\{E(Y_t|I_{t-1}) - \alpha_0 - \sum_{j=1}^{d} \alpha_j E(Y_t|x_{jt})\}^2$ is minimised with respect to $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, ..., \alpha_d)$. Differently from this $\mathcal{L}_2$ distance in Li et al. (2015), our approximation in (5) and (6) is based on the Kullback-Leibler distance (KL-distance), a natural distance function from a "true" probability distribution, $p_{yt} = P(Y_t = y|I_{t-1}) = p_t^y(1-p_t)^{1-y}$, to a "target" probability distribution, $q_{yt} = q_t^y(1-q_t)^{1-y}$, for $y = 0, 1$, with $q_t = q_t(\boldsymbol{\alpha}) = F(f_t^{MA})$ and f_t^{MA} defined in (5),

$$KL(p_{yt}, q_{yt}) = E_{p_{yt}}\{\log(p_{yt}/q_{yt})\}, \tag{7}$$

which is minimised with respect to $\boldsymbol{\alpha}$; we denote this minimiser by $\boldsymbol{\alpha}_0$. Note that (5) or (6) is a kind of approximation to the binary-valued distribution in $p_t = P(Y_t = 1|x_{1t}, \cdots, x_{dt})$. So this KL distance is appropriate to measure the closeness of the approximation of distribution, which is widely applied (c.f., Zhang, Yu, Zou & Liang (2016)). We hence need to estimate the minimiser by maximum likelihood estimation below.

We make some comments before ending this section. Firstly, in this paper we focus on the MAMaLoR procedure as given in (5) or (6) for easy implementation, but the basic idea underlying our proposed method can apply more than this. In general, estimation of the conditional probability p_t of $Y_t = 1$ given $X_t = (x_{1t}, \cdots, x_{dt})$ by nonparametric logistic regression for classification suffers from curse of dimensionality if $d > 3$, but we can well estimate the low-dimensional marginal conditional probabilities. We therefore try to approximate this high-dimensional conditional probability p_t by the affine combination, in logit transformation, of low-dimensional marginal conditional probabilities, say one-dimensional p_{jt}, $j = 1, \cdots, d$, as done above for simplicity in this paper. Here our MAMaLoR approximation given in (5) or (6) shares a similar model form as a special case of GAM (Hastie & Tibshirani (1987)), but it more easily avoids the shortcomings that the GAM suffers from, such as heavier computational costs and other deficiencies in forecasting due to possible overfitting with GAM in particular in the case of relatively small samples but with a larger number of time series lagged predictors. In addition, the low-dimensional marginal non-parametric logistic regressions could also be two-dimensional for combination in our MAMaLoR. Note that p_{jt}'s used in the combination approximation (5) could be replaced or added by other low-, say two-, dimensional marginal conditional probabilities $p_{jkt} = P(Y_t = 1|x_{jt}, x_{kt})$, for $j, k = 1, \cdots, d$, in the approximation, where it is not of a GAM form. However, this approximation would lead to additional issues including more careful variable selection needed for a good classification forecasting when d is large (c.f., Chen et al. (2018)), so we leave this problem for study in other work. Secondly, our combination idea for binary forecasting above is different from that of Lahiri & Yang (2016). In Lahiri & Yang (2016), it is based on discriminant analysis idea with copula applied to combine the conditional marginal distributions of two components of X_t, say x_{1t} and x_{2t}, given the binary response $Y_t = 1$ (in the notation of our paper) to model the conditional joint distribution of (x_{1t}, x_{2t}) given $Y_t = 1$. They suppose both conditional marginal distributions of x_{1t} and x_{2t} given the binary response $Y_t = 1$ as well as the copula function are known with parametric distributions respectively up to some unknown parameters. They mainly focus on the case $d = 2$, rather than $d > 3$ as addressed in this paper. When $d = 2$, we can also estimate the conditional joint probability density function of (x_{1t}, x_{2t}) given $Y_t = 1$ non-parametrically via the equality $f(x_1, x_2|Y = 1) = P(Y = 1|x_1, x_2) f_{X_1,X_2}(x_1, x_2)/P(Y = 1)$, where $f_{X_1,X_2}(x_1, x_2)$ stands for the joint probability density function of (x_{1t}, x_{2t}) while $P(Y = 1|x_1, x_2)$ is just what we are concerned with above.

3. Estimation and Properties

3.1. Estimation

We articulate the estimation for the MAMaLoR procedure in two steps.

Step one, to estimate the weight coefficients in (5), as $f_j(x_{jt})$'s are unknown, we need to estimate these marginal conditional regression first. Here nonparametric smoother is used to estimate the marginal probability $p_{jt} = E(Y_t = 1|x_{jt})$ through that given in (3). We suggest applying maximum likelihood local linear fitting (c.f., Fan et al. (1998a)) for estimation of $f_j(\cdot)$ in (3) as it is one-dimension and Y_t given x_{jt} follows $Bernoulli(p_{jt})$ distribution. Note that by taking the Taylor expansion of $f_j(x_{jt})$ at an arbitrary point x_{j0} given it is differentiable, then as x_{jt} is close to x_{j0}, it gives an approximation

$$\begin{aligned} f_j(x_{jt}) &\approx f_j(x_{j0}) + f_j'(x_{j0})(x_{jt} - x_{j0}) \\ &\equiv \beta_1 + \beta_2(x_{jt} - x_{j0}), \text{ if } |x_{jt} - x_{j0}| \le h, \end{aligned} \tag{8}$$

where h is a bandwidth to be appropriately selected. Then under the conditional independence of Y_t given the relevant information up to time $(t-1)$ along t, define the conditional local log likelihood function for (3) and (8) by:

$$\begin{aligned}\ell(\boldsymbol{\beta}, x_{j0}, h) = \sum_{t=1}^{n} [Y_t(\beta_1 + \beta_2(x_{jt} - x_{j0})) \\ - \log(1 + \exp(\beta_1 + \beta_2(x_{jt} - x_{j0})))] K_h(x_{jt} - x_{j0}),\end{aligned} \tag{9}$$

where $K_h(\cdot) = h^{-1}K(\cdot/h)$ with $K(\cdot)$ a kernel function on R^1 (c.f. Jones, Davies & Park (1994)). The aim is to estimate $\boldsymbol{\beta} = (\beta_1, \beta_2^T) = (f_j(x_{j0}), f_j'(x_{j0}))^T$, that is,

$$\begin{bmatrix} \hat{f}_j(x_{j0}) \\ \hat{f}_j'(x_{j0}) \end{bmatrix} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \arg\max_{\beta_1, \beta_2} \ell(\boldsymbol{\beta}, x_{j0}, h). \tag{10}$$

By solving the optimisation, which is easy as it could be seen as a locally weighted linear regression, we then get the estimation at x_{j0} as the intercept $\hat{f}_j(x_{j0})$ in the equation (8). Since x_{j0} is chosen arbitrarily, by letting x_{j0} go through each point in x_{jt}, we then get the estimated marginal probability $\hat{p}_{jt} = F(\hat{f}_j(x_{jt}))$, where we recall $F(y) = e^y/(1+e^y)$.

Step two, now we can try to estimate the coefficients in (5) together with replacing the $f_j(x_{jt})$'s by $\hat{f}_j(x_{jt})$'s. That is, we would like to estimate the minimiser that minimises (7) by using maximum likelihood estimation.

Under the conditional independence of Y_t given the relevant information up to time $(t-1)$ along t, following from (5), we can define the (approximate) conditional likelihood function as follows:

$$\begin{aligned} L(\boldsymbol{\alpha}) &= \prod_{t=1}^{n} P(Y_t = y_t | I_{t-1}; \boldsymbol{\alpha}) \\ &= \prod_{t=1}^{n} (p_t(\boldsymbol{\alpha}))^{y_t} (1 - p_t(\boldsymbol{\alpha}))^{1-y_t}, \end{aligned} \tag{11}$$

where

$$p_t(\boldsymbol{\alpha}) = \frac{e^{\alpha_0 + \sum_{j=1}^{d} \alpha_j f_j(x_{jt})}}{1 + e^{\alpha_0 + \sum_{j=1}^{d} \alpha_j f_j(x_{jt})}}. \tag{12}$$

Note that (11) can be also viewed as a kind of composite likelihood; see Varin, Reid & Firth (2011). Then taking nature log of the equation (11) together with (12), with $f_j(x_{jt})$'s replaced by $\hat{f}_j(x_{jt})$'s, we define the log conditional likelihood function (scaled by $1/n$) as follows

$$\hat{l}(\boldsymbol{\alpha}) = \frac{1}{n} \sum_{t=1}^{n} \left[y_t(\alpha_0 + \sum_{j=1}^{d} \alpha_j \hat{f}_j(x_{jt})) - \log(1 + e^{\alpha_0 + \sum_{j=1}^{d} \alpha_j \hat{f}_j(x_{jt})}) \right]. \tag{13}$$

In order to control the impacts of the poor estimate of $f_j(\cdot)$'s at the extreme x_{jt}'s, we slightly modify the estimation procedure with the log-likelihood given in (13), and define the following modified log-likelihood function:

$$l_n(\boldsymbol{\alpha}) = l_n(\hat{\mathbf{f}}(\cdot), \boldsymbol{\alpha}) = \frac{1}{n} \sum_{t=1}^{n} \left[\left\{ Y_t \left(\alpha_0 + \sum_{j=1}^{d} \alpha_j \hat{f}_j(x_{jt}) \right) \right\} - \log \left\{ 1 + \exp \left(\alpha_0 + \sum_{j=1}^{d} \alpha_j \hat{f}_j(x_{jt}) \right) \right\} \right] w(X_t), \tag{14}$$

which asymptotically corresponds to the population log-likelihood function:

$$l(\mathbf{f}(\cdot), \boldsymbol{\alpha}) = E \left[\left\{ Y_t \left(\alpha_0 + \sum_{j=1}^{d} \alpha_j f_j(x_{jt}) \right) \right\} - \log \left\{ 1 + \exp \left(\alpha_0 + \sum_{j=1}^{d} \alpha_j f_j(x_{jt}) \right) \right\} \right] w(X_t), \tag{15}$$

where $\mathbf{f}(\cdot) = (f_1(\cdot), \cdots, f_d(\cdot))^T$, $\hat{\mathbf{f}}(\cdot)$ is defined similarly with estimated elements, $X_t = (x_{1t}, ..., x_{dt})$ and $w(X_t) = \prod_{j=1}^{d} \mathbf{I}_{(c_{0j} \le x_{jt} \le c_{1j})}$ is a weight function controlling the edge effects in the estimation with $\mathbf{I}_{(\cdot)}$ being an indicator function

and $c_{0j} < c_{1j}$ appropriately chosen. For example, in practice, c_{0j} and c_{1j} may be chosen to include all observations, or as 0.1 and 0.9 quantiles of the sample $x_{jt}, t = 1, 2, \cdots, n$, if there are extreme outliers, which are hence removed from estimation by using this control weight function Lu, Tjøstheim & Yao (2007)[Section 3.2]. Note that $\hat{\boldsymbol{\alpha}} = \arg\max_{\alpha} l_n(\hat{\mathbf{f}}(\cdot), \boldsymbol{\alpha})$ gives the estimator $\hat{\boldsymbol{\alpha}}$ from sample data and $\boldsymbol{\alpha}^{(0)} = \arg\max_{\alpha} l(\mathbf{f}(\cdot), \boldsymbol{\alpha})$ gives the true parameter vector $\boldsymbol{\alpha}_0 = (\alpha_{00}, \alpha_{01}, ..., \alpha_{0d})^T$.

We now take the first order derivative of the modified log-likelihood function (14) with respect to α_j:

$$\frac{\partial l_n(\boldsymbol{\alpha})}{\partial \alpha_j} = \frac{1}{n}\sum_{t=1}^{n}[y_t \hat{f}_j(x_{jt}) - \hat{p}_t \hat{f}_j(x_{jt})]w(X_t), \tag{16}$$

where

$$\hat{p}_t = \hat{p}_t(\boldsymbol{\alpha}) = \frac{e^{\alpha_0 + \sum_{j=1}^{d} \alpha_j \hat{f}_j(x_{jt})}}{1 + e^{\alpha_0 + \sum_{j=1}^{d} \alpha_j \hat{f}_j(x_{jt})}}. \tag{17}$$

The second order derivative, which is also known as the Hessian matrix, is negative definite:

$$\frac{\partial^2 l_n(\boldsymbol{\alpha})}{\partial \alpha_j \partial \alpha_k} = -\frac{1}{n}\sum_{t=1}^{n} \hat{f}_j(x_{jt})\hat{p}_t(1-\hat{p}_t)\hat{f}_k(x_{kt})w(X_t). \tag{18}$$

This is to say, the likelihood function is concave and hence has a unique maximiser.

From the computational perspective, note that equation (5) looks like a logistic linear regression with $\hat{f}_j(x_{jt})$ given, which means we can apply relevant technique and algorithm developed in GLM with logistic regression. Therefore our MAMaLoR procedure is easy to implement in computation. In addition, both marginal nonparametric logistic regression estimation by local linear fitting and parametric affine combination estimation are applied in our method, so the MAMaLoR procedure is of "semiparametric" nature.

3.2. Asymptotic properties

In this section, we present the large sample property of asymptotic normality for our proposed MAMaLoR procedure. We would like to first show $\hat{\boldsymbol{\alpha}} \to \boldsymbol{\alpha}^{(0)}$ in probability as $n \to \infty$.

For notational ease below, we define

$$p_t(\mathbf{f}(\cdot), \boldsymbol{\alpha}) = \frac{e^{\alpha_0 + \sum_{j=1}^{d} \alpha_j f_j(x_{jt})}}{1 + e^{\alpha_0 + \sum_{j=1}^{d} \alpha_j f_j(x_{jt})}}. \tag{19}$$

Note that $p_t(\boldsymbol{\alpha}) = p_t(\mathbf{f}(\cdot), \boldsymbol{\alpha})$ and $\hat{p}_t(\boldsymbol{\alpha}) = p_t(\hat{\mathbf{f}}(\cdot), \boldsymbol{\alpha})$.

In addition, we suppose (Y_t, X_t^T) are β-mixing, for which we give the following definition:

Definition. Let $Z_t = (Y_t, X_t)$ be a strictly stationary time series. The process Z_t is said to be β-mixing if

$$\beta(n) = E\left\{\sup_{B \in \mathcal{F}_{t+n}^{\infty}} |P(B) - P(B|Z_t, Z_{t-1}, ...)|\right\} \to 0,$$

as $n \to \infty$, where $\mathcal{F}_{t+n}^{\infty}$ is the information field (a so-called σ-algebra) of $\{Z_s, s \geq t+n\}$.

For the completion of our results, we now introduce the following assumptions.

A1 (i) We assume (Y_t, X_t) (with Y_t being binary) is strictly stationary process under β-mixing condition. There exists $b > \max(2(\rho r + 1)/(\rho r - 2), (r + a)/(1 - 2/\rho))$ and $a \geq (r\rho - 2)r/(2 + r\rho - 4r)$, such that $\beta(t) = O(t^{-b})$; (ii) for any $t_1 < \cdots < t_s$ and $1 \leq s \leq 2r$, the joint probability density function of $(X_{t_1}, \cdots, X_{t_s}) := g_{X_{t_1}, \cdots, X_{t_s}}(x_1, \cdots, x_s)$ is bounded above uniformly; (iii) there exists $\rho > 4 - 2/r$ in R and $r \geq 1$ in Z, such that $E|X_t|^{\rho r} < \infty$.

A2 The weight function $w(X_t) = \prod_{j=1}^{d} I_{(c_{0j} \leq x_{jt} \leq c_{1j})}$ with $c_{0j} < c_{1j}$ appropriately chosen, where $I_{(\cdot)}$ is an indicator function.

This weight function is used for controlling the edge effects in the estimation.

A3 (i) The bandwidth $h = h_n$ satisfies the conditions $\lim_{n\to\infty} h = 0$ and $\liminf_{n\to\infty} nh^{\frac{2(r-1)a+(\rho r-2)}{(a+1)\rho}} > 0$ for some integer $r \geq 3$; (ii) There exists a sequence of positive integers $s_n \to \infty$ such that $s_n = o((nh)^{1/2})$, $ns_n^{-b} \to 0$ and $s_n h^{\frac{2(\rho r-2)}{[2+b(\rho r-2)]}} > 1$ as $n \to \infty$; (iii) $nh^4 = o(1)$ as $n \to \infty$.

A4 Let $\mathbf{f}_0(\cdot) = (f_1(\cdot), ..., f_d(\cdot))^T$ be the vector of the true conditional regression functions, with $f_j(\cdot)$'s defined in Equation (3). For an $\mathbf{f}(\cdot)$, define its Lipschitz norm: For some $\phi > 0$, let $[\phi]$ be the largest integer not greater than ϕ, and define (if it exists)

$$\|\mathbf{f}\|_{\infty,\phi} = \max_{0\leq\kappa\leq[\phi]} \sup_{x\in A} \|\mathbf{f}^{(\kappa)}(x)\| + \sup_{x\neq x';x,x'\in A} \frac{\|\mathbf{f}^{([\phi])}(x) - \mathbf{f}^{([\phi])}(x')\|}{\|x - x'\|^{\phi-[\phi]}}, \tag{20}$$

where $\mathbf{f}^{(\kappa)}(x)$ is the κ-th derivative of $\mathbf{f}(x)$ with respect to x, and $A = \prod_{j=1}^{d}[c_{0j}, c_{1j}]$ with some real values of c_{0j} and c_{1j} satisfying $c_{0j} < c_{1j}$ given in assumption A2. We suppose $\mathbf{f}_0(\cdot)$ with f_j's belongs to the functional space $\mathbf{F}$ with $\phi \geq 2$:

$$\mathbf{F} := \{\mathbf{f} : \text{continuous from } A \text{ to } R^d \text{ with } \|\mathbf{f}\|_{\infty,\phi} \leq c\}, \tag{21}$$

where c is a positive constant. This functional space $\mathbf{F}$ (containing functions $\mathbf{f}$ of which its Lipschitz norm is bounded) is often denoted by $C_c^{\phi}(A)$.

A5 For the local likelihood function (9), define $\Phi(Y_t, z_j) = Y_t - \exp(z_j)/[1 + \exp(z_j)]$, and

$$m(x_j, z_j) = E[\Phi(Y_t, z_j)|x_{jt} = x_j], \tag{22}$$

satisfying $(x_j, z_j) \to m(x_j, z_j) \cdot g_j(x_j)$ is three times continuously differentiable as a function form R^2 to R, where $g_j(x_j)$ is the marginal density of x_{jt}, which is strictly positive and continuous over $A_j = [c_{0j}, c_{1j}]$. We denote the derivative of m with respect to x_j by m_1', and the derivative with respect z_j by m_2', etc.

Remark. (i) Assumption 1 shows a technical standard β-mixing process which is satisfied by many linear and nonlinear time series models under geometric ergodicity (Fan & Yao, 2003; Lu et al., 2007). The edge effect is controlled by Assumption 2, which removes the extreme estimates around the boundaries of X_t, in order to improve the practical performance of the estimation (c.f. Fan, Härdle & Mammen (1998b), Fan, Yao & Cai (2003) and Lu et al. (2007)).

(ii) Assumption 3 is also standard in time series topics (Fan et al., 2003; Lu et al., 2007) and easily satisfied though it looks a bit involved. For example, if we take $h = n^{-c}$ with $1/4 < c < (b-2)/b$ and $s_n = (nh)^{1/k}$ with $2 < k < (1-c)b$, then it follows that $s_n = (nh)^{1/k} = n^{(1-c)/k} \to \infty$, $s_n = o((nh)^{1/2})$, $ns_n^{-b} = n^{1-b(1-c)/k} \to 0$ and $nh^4 = n^{1-4c} = o(1)$ as $n \to \infty$, while $\liminf_{n\to\infty} nh^{b_1} > 0$ if $c < 1/b_1$, where $b_1 \equiv \frac{2(r-1)a+(\rho r-2)}{(a+1)\rho}$. As $ns_n^{-b} \to 0$, we have $s_n \geq n^{1/b}$ as n is sufficiently large, and, letting $b_2 \equiv \frac{2(\rho r-2)}{[2+b(\rho r-2)]}$, hence $s_n h^{b_2} \geq n^{1/b-cb_2} > 1$ if $c < 1/(bb_2) = \frac{[2+b(\rho r-2)]}{2(\rho r-2)b} > 1/2$. Therefore A3(i)-(iii) is satisfied if there is some c such that $1/4 < c < \min\{(b-2)/b, 1/b_1, 1/(bb_2)\}$, which holds true if $b > 8/3$, $b_1 < 4$ and $bb_2 < 4$. Here $b_1 < 4$ is equivalent to $a > \frac{(r-4)\rho-2}{4\rho-2(r-1)}$. Note that $bb_2 < 2$. So A3(i)-(iii) holds true easily. Note that the lim inf in A3(i) that is finite, just greater than 0, is needed – it borrows from Assumption (C7) of Lu et al. (2007).

(iii) Assumptions 4 and 5 give smoothness conditions on the conditional regression and marginal density functions. The Lipschitz norm conditions (Assumption 4) are introduced to give a tighter bound than uniform norm (Nielsen, 2005). For more information on Lipschitz norm, the reader is referred to Van Der Vaart & Wellner (1996).

Theorem 3.1. *(Consistency) Suppose Assumptions A1-A5 hold. Let $\mathfrak{A}$ be a close set in R^{d+1} and $\boldsymbol{\alpha}^{(0)}$ is an interior point of $\mathfrak{A}$ and $\mathbf{f} \in \mathbf{F}$. Then $\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^{(0)} = op(1)$.*

It is to prove the convergence of $\hat{\boldsymbol{\alpha}}$ to $\boldsymbol{\alpha}^{(0)}$ in probability. That is, we would like to show:

$$\forall \delta > 0, P(\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^{(0)}\| > \delta) \to 0,$$

as $n \to \infty$.

Here we follow Lemma 4.1 of Lu et al. (2007), given below, to prove Theorem 3.1.

Proposition 3.1. *(Consistency Lemma) Suppose* $\boldsymbol{\alpha}^{(0)} \in \mathfrak{A}$ *satisfies* $l(\mathbf{f}_0(\cdot), \boldsymbol{\alpha}^{(0)}) = \max_{\boldsymbol{\alpha}\in\mathfrak{A}} l(\mathbf{f}_0(\cdot), \boldsymbol{\alpha})$*, where* $\mathbf{f}_0(\cdot)$ *is the true function vector in Assumption A4,* $\mathfrak{A}$ *is a closed set in* $\mathbb{R}^{d+1}$ *with* $\boldsymbol{\alpha}^{(0)}$ *an interior point of* $\mathfrak{A}$*, and that*

i. $l_n(\hat{\mathbf{f}}(\cdot), \hat{\boldsymbol{\alpha}}) \leq \max_{\boldsymbol{\alpha}\in\mathfrak{A}} l_n(\hat{\mathbf{f}}(\cdot), \boldsymbol{\alpha}) + o_p(1)$.

ii. For all $\delta > 0$*, there exists* $\epsilon(\delta) > 0$ *such that*

$$\inf_{\|\boldsymbol{\alpha}-\boldsymbol{\alpha}^{(0)}\|>\delta} |l(\mathbf{f}_0(\cdot), \boldsymbol{\alpha}) - l(\mathbf{f}_0(\cdot), \boldsymbol{\alpha}^{(0)})| \geq \epsilon(\delta).$$

iii. Uniformly for all $\boldsymbol{\alpha} \in \mathfrak{A}$*,* $l(\mathbf{f}(\cdot), \boldsymbol{\alpha})$ *is continuous with respect to the metric* $\|\cdot\|_{\mathbf{F}}$ *in* $\mathbf{f}(\cdot)$ *at* $\mathbf{f}_0(\cdot)$*, where* $\|\mathbf{f}(\cdot)\|_{\mathbf{F}} = \sup_{x\in A} \|\mathbf{f}(x)\|$ *with* $\|\cdot\|$ *being the Euclidean norm of* R^d.

iv. $\|\hat{\mathbf{f}}(\cdot) - \mathbf{f}_0(\cdot)\|_{\mathbf{F}} = o_p(1)$.

v. For all δ_n *with* $\delta_n = o(1)$*,*

$$\sup_{\boldsymbol{\alpha}\in\mathfrak{A}} \sup_{\|\mathbf{f}(\cdot)-\mathbf{f}_0(\cdot)\|_{\mathbf{F}}\leq\delta_n} |l_n(\mathbf{f}(\cdot), \boldsymbol{\alpha}) - l(\mathbf{f}(\cdot), \boldsymbol{\alpha})| = o_p(1).$$

Then $\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^{(0)} = o_p(1)$.

The proof of Theorem 3.1 is relegated to Appendix A.

For asymptotic normality, we need to introduce some more notation. Let $\widetilde{\chi}_t(\mathbf{f}_0) = (1, f_1(x_{1t})..., f_d(x_{dt}))^T$ with $f_j(x_{jt})$ defined in (3),

$$\mathbf{U} = E[-p_t(1-p_t)\widetilde{\chi}_t(\mathbf{f}_0)\widetilde{\chi}_t(\mathbf{f}_0)^T]w(X_t),$$

and

$$\mathbf{V} = \lim_{n\to\infty} Var\left(\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(Y_t - p_t)\widetilde{\chi}_t(\mathbf{f}_0)w(X_t)\right). \tag{23}$$

Then we have

Theorem 3.2. *(Asymptotic Normality)*

Suppose that the assumptions A1-A5 are satisfied, for $\boldsymbol{\alpha} \in \mathfrak{A}$*, and* $\mathbf{U}$ *is invertible. Then*

$$\sqrt{n}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^{(0)}) \xrightarrow{L} N(0, \mathbf{U}^{-1}\mathbf{V}\mathbf{U}^{-1}), \tag{24}$$

as $n \to \infty$*, where* $\xrightarrow{L}$ *stands for convergence in distribution.*

We remark that owing to time series dependence, $\mathbf{V}$ may not be equal to $\mathbf{U}$ in Theorem 3.2. The proof of Theorem 3.2 is provided in Appendix B.

4. Numerical evidence

In this section, we illustrate the empirical application of our proposed MAMaLoR model by both simulated and real data numerical examples to understand the impact of lagged information on binary-valued time series data forecasting. A Monte-Carlo simulation study is given in the first subsection and an application to financial data of FTSE 100 index is then presented in the second subsection.

4.1. A simulation study

In order to examine the finite sample performance of the method, a Monte-Carlo simulation is made. Bandwidth selection for h in (9) is indeed an important problem but appears quite sensitive to outliers for the Cross-Validation (CV) based on likelihood. So we leave this for further investigation. In the simulation, we applied a simple Cross-Validation by using h.select in R package *sm*, which is actually based on a direct estimation of $p_{jt} = E(Y_t|x_{jt})$.

The model used in this section is given as follows:

$$Y_t = I(x_t > 0),$$

Table 1
Parameters specified in Model (25)

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

$$x_t = \sum_{k=1}^{9} g_{0k}(x_{t-k}) + \epsilon_t,$$

and

$$g_{0k}(x_{t-k}) = a_k x_{t-k} + \delta \exp(-k x_{t-k})/(1+\exp(-k x_{t-k})) + \gamma \cos(x_{t-k} x_{t-1}), \tag{25}$$

where ϵ_t's are $i.i.d.$ following a logistic distribution, generated by $\epsilon_t = \log(e_t/(1-e_t))$ with e_t having a uniform distribution over the interval $(0,1)$. Here we use logistic distribution for the error term so that the resultant model is a logistic time series regression model with the true link function being a logit link function; see (27) below. Our simulation model for x_t is basically similar to that in Li et al. (2015), where the values of a_k, for $k = 1, 2, \ldots, 9$, are given in Table 1. We have taken a_k's such that all the roots of the polynomial, $1 - \sum_{k=1}^{9} a_k \lambda^k$, are outside the unit circle and note that $\sum_{k=1}^{9} g_{0k}(x_k) = \sum_{k=1}^{9} a_k x_k + o(\|x\|)$, as $\|x\| \to \infty$, no matter what finite real values the δ and γ take on, where $\|x\|$ is the Euclidean norm of $x = (x_1, x_2, \cdots, x_9)'$, so there is a geometrically ergodic stationary solution, which is β-mixing with exponentially decaying mixing coefficient, for x_t in (25) (c.f., Lu (1998)). We will have the constants δ and γ taking on values of 0 and 0.5, respectively, with the 4 pairs of which specified in Figures 1 and 2. Note that δ and γ with non-zero values are used to change the model with nonlinear structure or interaction. When $(\delta, \gamma) = (0,0)$, the x_t process in (25) is a purely linear AR model; when $\gamma = 0$ but $\delta \neq 0$, it is an additive AR model, while $\gamma \neq 0$ leads to a model with interaction between $x_{kt} = x_{t-k}$ and $x_{1t} = x_{t-1}$, for $k = 1, 2, \cdots, 9$. The larger the value γ, the larger the deviation of the model from an additive structure for x_t.

By the assumption imposed on model (25), Y_t given $X_t := \{x_{t-1}, \ldots, x_{t-9}\}$ follows a Bernoulli distribution with probability p_t, that is

$$[Y_t | X_t] \sim Bin(1, p_t), \tag{26}$$

where $Bin(\cdot,\cdot)$ stands for a binomial distribution, and the probability p_t is defined as,

$$\begin{aligned} p_t &= P(Y_t = 1 | X_t) = P\left(\sum_{k=1}^{9} g_{0k}(x_{t-k}) + \epsilon_t > 0 | X_t\right) = P\left(\epsilon_t > -\sum_{k=1}^{9} g_{0k}(x_{t-k}) | X_t\right) \\ &= 1 - F\left(-\sum_{k=1}^{9} g_{0k}(x_{t-k})\right) = F\left(\sum_{k=1}^{9} g_{0k}(x_{t-k})\right), \end{aligned} \tag{27}$$

where $F(z) = e^z/(1+e^z)$, for $z \in R^1$, is a logistic cumulative distribution function. In the simulation below, we apply a logistic classification forecasting based on the observations of (Y_t, X_t).

The simulation consists of the data generated with the estimation sample size set to be $n = 500$ and $n = 1000$, respectively, and a testing sample of size of $n_p = 50$ for prediction evaluation. When generating the time series data, in view of a necessary warming up step, we deleted the first 100 observations every time from the $(100 + n + n_p)$ generated sample through the iterations for x_t in (25) with initial values taken to be zero. The simulation is repeated 100 times for each setting.

In this simulation, we let δ and γ take on values in $\{0, 0.5\}$ each time to represent different degrees of nonlinear structures and interactions in (25). For the bandwidth used in our estimation, how to select optimal one for forecasting is still an open question. We just applied the simplest cross validation for the needed bandwidth in simulation. To evaluate the forecasting, we apply the area under the curve (AUC) of receiver operation characteristic (ROC), which is a popular criterion often used to evaluate the performance of prediction for binary variable classification. The larger the AUC, the better the model. The boxplots of the AUC values of 100 repetitions with the testing sample of size $n_p = 50$ for different methods are plotted in Figures 1 and 2 with estimation sample of size $n = 500$ and $n = 1000$,

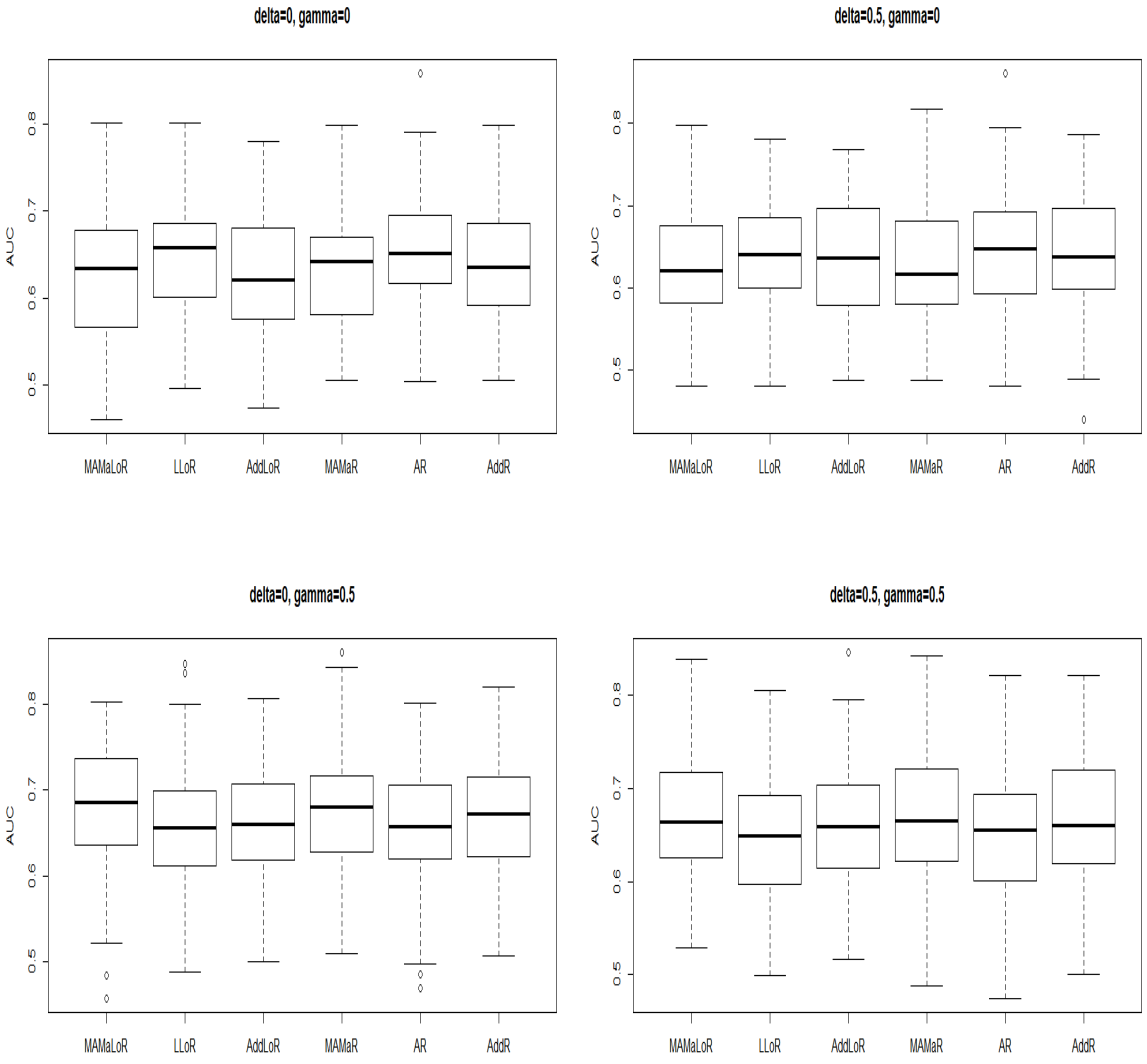

Figure 1: Boxplots of the area under curve (AUC) with 100 repetitions for one-step ahead classification predictions, with $n_p = 50$ observations for testing, of different methods under different true model structures (Top left: linear, Top right: additive, Bottom left & right: nonlinear non-additive) based on $n = 500$ observations for training.

respectively. The methods, in each panel, include "MAMaLoR", "LLoR" and "AddLoR" referring to the maximum likelihood estimation methods based on model averaging marginal nolinear logistic model (proposed in this paper), linear logistic regression model (via GLM) and additive logistic model (via GAM), respectively. In most practical applications with binary classification, we can only observe (Y_t, X_t) with X_t representing the past observations of x_t, rather than x_t itself, but for the real data example with stock price below, we can have the data of x_t, and we have therefore, as a comparison, additionally consider the classification forecasting of Y_t through $Y_t = I(x_t > 0)$ with forecasting of x_t by the methods of "MAMaR", "AR" and "'AddR" representing least squares estimations of nonlinear MAMaR model (Li et al., 2015), pure AR model and additive model for x_t, respectively. Note that the latter three models are used to predict the value of x_t directly and then we convert it into prediction of binary Y_t. Following from Figures 1 and 2, we summarise our findings as follows.

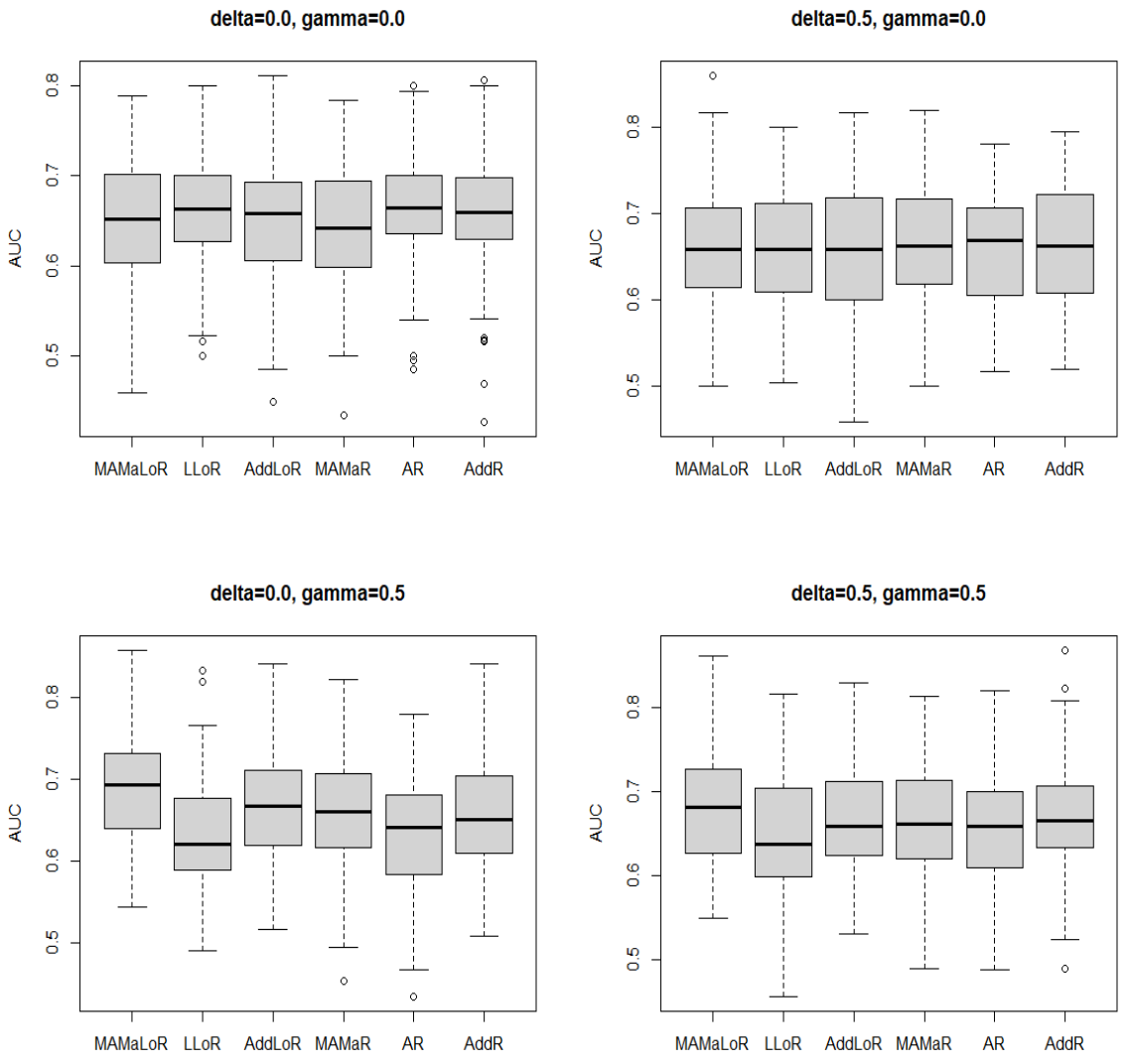

Figure 2: Boxplots of the area under curve (AUC) with 100 repetitions for one-step ahead classification predictions, with $n_p = 50$ observations for testing, of different methods under different true model structures (Top left: linear, Top right: additive, Bottom left & right: nonlinear non-additive) based on $n = 1000$ observations for training.

(i) When the true models are additive (corresponding to $\gamma = 0$) as indicated in the upper panels of Figures 1 and 2, we can see that the performances of our proposed MAMaLoR method, though not the best, are basically comparable to those of the additive logistic (AddLoR) model in classification forecasting in terms of the popular classification performance measure of area under curve (AUC). Here if the true model is linear (corresponding to $\delta = 0$ and $\gamma = 0$), then, as expected, linear logistic (LLoR) model performs the best in classification forecasting. Furthermore, it is interesting to note that the LLoR method even performs better than the AddLoR in forecasting when the true model is nonlinearly additive (corresponding to $\delta = 0.5$ and $\gamma = 0$) with the training sample size being $n = 500$ (shown in the upper right panel of Figure 1); however, as the training sample size increases to $n = 1000$, our proposed MAMaLoR method clearly becomes comparable to both LLoR and AddLoR in performance of forecasting as shown in the upper right panel of Figure 2.

Table 2
Parameters specified in Model (28)

a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}
0.0542	-0.0837	0.0578	-0.1336	-0.0152	-0.0042	-0.0286	0.0102	-0.0174	-0.0302	-0.0629
a_{12}	a_{13}	a_{14}	a_{15}	a_{16}	a_{17}	a_{18}	a_{19}	a_{20}	a_{21}	a_{22}
0.0258	-0.0207	-0.0266	-0.0375	0.0639	-0.0528	0.0615	-0.0508	0.1036	-0.0307	0.0785
a_{23}	a_{24}	a_{25}	a_{26}	a_{27}	a_{28}	a_{29}	a_{30}	a_{31}		
-0.0806	-0.0381	0.0755	0.0096	-0.0257	-0.0273	-0.0717	-0.0229	-0.0309		

(ii) When the true models are not additive (corresponding to $\gamma \neq 0$) as indicated in the bottom panels of Figures 1 and 2, we can clearly see that the performances of our proposed MAMaLoR method are the best among all the six considered methods. Interestingly, our MAMaLoR method performs much better than both LLoR and AddLoR methods in classification forecasting in both cases of $n = 500$ (bottom panel of Figure 1) and $n = 1000$ (bottom panel of Figure 2). Here the LLoR method performs the worst.

(iii) When comparing logistic regression based forecasting methods (MAMaLoR, LLoR, AddLoR) with other indirect least squares (auto)regression based methods (MAMaR, AR, AddR) for classification, both classes of methods are basically correspondingly comparable when the true models are additive. But our MAMaLoR method performs the best if the true models are not additive, as indicated in both bottom panels of Figures 1 and 2, in particular the performance of our MAMaLoR method turns to be more viable when the training sample size n becomes large for time series bigger data.

We now extend the number of lags considered in (25) to 31. We use the following model along with the parameters summarised in Table 2 to generate data for the simulation.

$$Y_t = I(x_t > 0),$$
$$x_t = \sum_{k=1}^{31} g_{0k}(x_{t-k}) + \epsilon_t,$$

and

$$g_{0k}(x_{t-k}) = a_k x_{t-k} + \delta \exp(-k x_{t-k})/(1 + \exp(-k x_{t-k})) + \gamma \cos(x_{t-k} x_{t-1}), \tag{28}$$

where ϵ_t's are $i.i.d.$ following a logistic distribution, generated by $\epsilon_t = \log(e_t/(1 - e_t))$ with e_t having a uniform distribution over the interval $(0, 1)$. We use the parameters estimated by the linear AR model of 31 lags, i.e., AR(31), to the geometric return of Financial data of FTSE 100 Index, which is introduced later in the application section below.

Here (28) has an analogous setting to (25). Similar to (25), Y_t given $X_t := \{x_{t-1}, \ldots, x_{t-31}\}$ follows a Bernoulli distribution with probability p_t. We then conduct the Monte-Carlo simulation with the estimation sample size set to be $n = 1000$ and a testing sample of size of $n_p = 50$ for prediction evaluation.

We focus on the setting of non-additive data structure in (28), where $\delta = 0, \gamma = 0.5$. The results are depicted in Figure 3. It is noted that GAM model has been removed as it costed too much time to converge when facing a high dimension of $d = 31$. The performances of the candidate models are summarised: (i) when the model is not additive ($\gamma \neq 0$), the MAMaLoR model clearly outperforms the other candidate models in the context of prediction power, confirmed by the highest AUC value; (ii) the computational cost of MAMaLoR model is comparable to that for the LLoR and AR models, as it increases only in polynomial time when adding more lags (i.e., enlarge the dimension d).

To conclude it, our proposed MAMaLoR method is flexible to deal with binary-valued time series data with complex nonlinear and interaction structures. It is shown that MAMaLoR model can compete with other popular models in prediction at a lower computational cost. It overcomes the "curse of dimensionality" as one can easily add more predictor variables into the model and the computational time is still in polynomial time . However, when more and more predictor variables are added, we should take care to select the relevant variables for prediction, which is beyond the scope of this paper and left for further study.

4.2. An application: forecasting market moving direction of FTSE 100 index

In this section, we demonstrate practical advantages of our proposed MAMaLoR model by an application to forecasting market moving direction of FTSE 100 Index data. The data set includes close price, cp_t, the maximum price

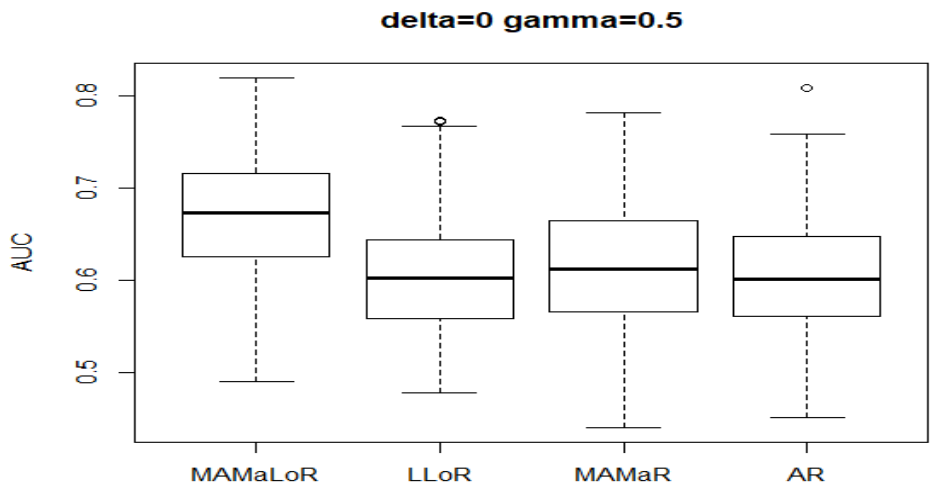

Figure 3: Boxplots of the area under curve (AUC) with 100 repetitions for one-step ahead classification predictions of non-additive data, with $n_p = 50$ observations for testing, for $lag = 31$, based on $n = 1000$ observations for training.

$maxp_t$ and the minimum price $minp_t$ of the day, and the trading volume Vlm_t for each day from 1 May 2013 to 1 May 2018, with 1263 observations. We are concerned with whether the market price is up ($Y_t = 1$) or not ($Y_t = 0$) is determined by the factors of historical data, such as volatility, volume and (geometric) return, which are defined, respectively, by

$$Y_t = \begin{cases} 1 & \text{if } cp_t - cp_{t-1} > 0 \\ 0 & \text{else,} \end{cases} \tag{29}$$

$$v_t = \log(100\frac{(maxp_t - minp_t)}{\frac{1}{2}(maxp_t + minp_t)}),$$

$$V_t = \log(Vlm_t)$$

and

$$G_t = 100log(\frac{cp_t}{cp_{t-1}}).$$

The three series of volatility v_t, log-volume V_t and geometric return G_t are depicted in figure (4). Note that $Y_t = I(G_t > 0)$, with $I(\cdot)$ standing for an indicator function.

In this example, we are interested in the one-step-ahead prediction of the market (price) moving direction Y_t by using the information of a range of lags of all volatility, volume and geometric return to examine if they help to improve the explanation or prediction of market direction. Each lagged variable will be treated as a single predictor and then fed to the model.

To start with, we consider $X_t = (v_{t-j}, V_{t-j}, G_{t-j}, j = 1, 2, 3, 4)$,i.e., a short lag of 4 and $3 * 4 = 12$ variables used in total, to predict Y_t. The number of lags will then be enlarged later to fully exploit the advantage of our proposed MAMaLoR procedure. Though the selection of the lags is important in prediction, we start with this arbitrary selected lag first. The training sample we used is from the 1st observation to the 800th observation. Our evaluation or testing sample for the prediction is the following 200 observations (801 to 1000) right after the training sample. Since Y_t is binary, we are plotting the Receiver Operating Characteristic (ROC) and computing Area Under the Curve (AUC) to compare the performances (see Ballings, Van den Poel, Hespeels & Gryp (2015)).

Figure 4: The time series plot of volatility v_t, log-volume V_t and geometric return G_t defined in (29).

We first estimate the marginal logistic regressions $f_j(\cdot)$'s in (3) for the given lagged volatility, volume and geometric return variables, respectively, with a bandwidth of 0.5 applied for initial investigation. we are comparing our MAMaLoR with the linear logistic (LLoR) and the additive logistic (AddLoR) models in forecasting of Y_t based on the lagged information of X_t. As to the LLoR and AddLoR models, we use, respectively, the GLM in R and the R package (gam) for the binomial family with logistic link, with the $s(.)$ functions that automatically specify a smoothing spline fit for each component of X_t in the GAM model. For ease of statement, we call the LLoR and the AddLoR models the GLM and the GAM below,

Note that in general it is poor to estimate the probability of $P(Y_t = 1|X_t)$ via a purely nonparametric logistic regression for such a high dimensional case with $X_t = (v_{t-j}, V_{t-j}, G_{t-j}, j = 1, 2, 3, 4)$ of dimension $d = 12$ due to curse of dimensionality. We compare the performance of our MAMaLoR model with both GLM and GAM in the forms detailed as follows:

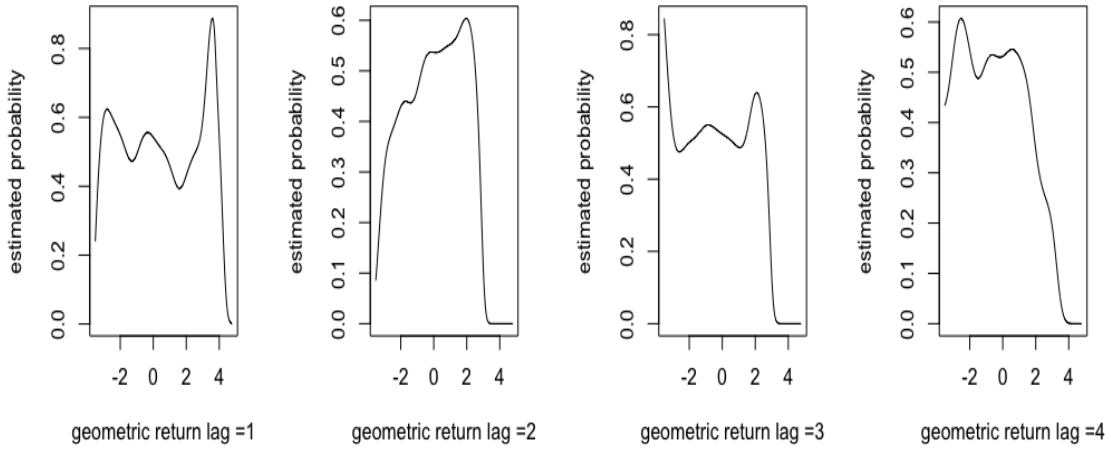

Figure 5: Marginal probability of significant variables in MAMaLoR model, with the quantities for x-axis defined in (29).

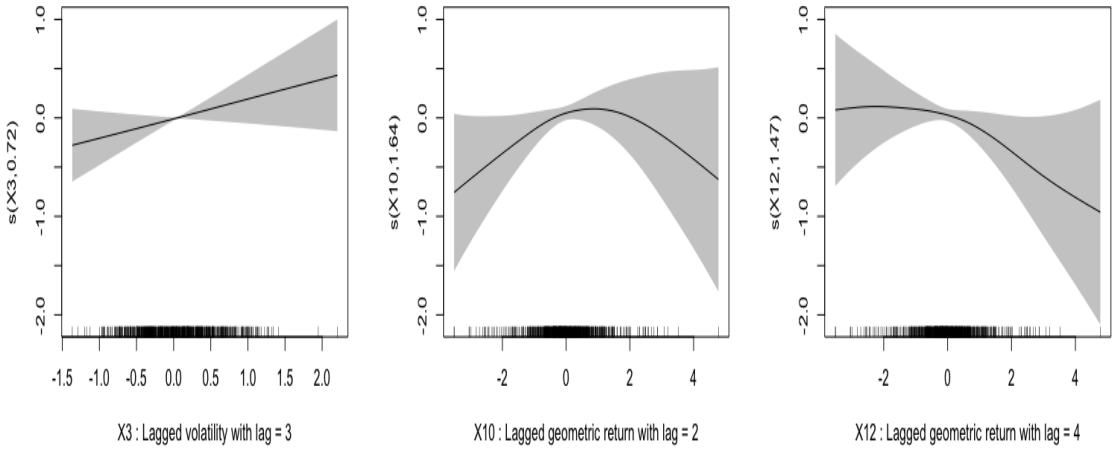

Figure 6: Smooth function for significant variables in GAM model, with the quantities for x-axis defined in (29).

MAMaLoR model:

$$logit(p_t) = \log \frac{p_t}{1-p_t} \approx \alpha_0 + \sum_{j=1}^{4} \alpha_j f_j(v_{t-j}) + \sum_{j=1}^{4} \alpha_{4+j} f_{4+j}(V_{t-j}) + \sum_{j=1}^{4} \alpha_{8+j} f_{8+j}(G_{t-j}), \tag{30}$$

where $f_j(v_{t-j}) = logit(P(Y_t = 1|v_{t-j}))$ for $j = 1, 2, 3, 4$ and $f_{4+j}(V_{t-j})$ and $f_{8+j}(G_{t-j})$ defined similarly are pre-estimated, respectively, as in (10) and then α_j's estimated, detailed in Section 3.1;

Table 3
Summary of MAMaLoR, GLM and GAM model fittings

	MAMaLoR model			GLM (LLoR) model		GAM (AddLoR) model	
	Estimate	Std. Error	Pr(|z|)		Pr(|z|)		P-value
Intercept	-0.7697	0.2001	0.000120 ***	Intercept	0.496	Intercept	0.204
$f_1(v_{t-1})$	-0.1815	1.2192	0.881654	v_{t-1}	0.632	$s(v_{t-1})$	0.7654
$f_2(v_{t-2})$	0.7542	0.6624	0.254884	v_{t-2}	0.460	$s(v_{t-2})$	0.2275
$f_3(v_{t-3})$	0.7087	0.6118	0.246755	v_{t-3}	0.237	$s(v_{t-3})$	0.0586 .
$f_4(v_{t-4})$	0.9240	0.8155	0.257175	v_{t-4}	0.835	$s(v_{t-4})$	0.6357
$f_5(V_{t-1})$	1.0364	1.2139	0.393246	V_{t-1}	0.580	$s(V_{t-1})$	1.0000
$f_6(V_{t-2})$	0.4283	0.6882	0.533729	V_{t-2}	0.306	$s(V_{t-2})$	0.1544
$f_7(V_{t-3})$	0.3587	0.7690	0.640865	V_{t-3}	0.918	$s(V_{t-3})$	0.5899
$f_8(V_{t-4})$	-0.6463	1.0239	0.527874	V_{t-4}	0.852	$s(V_{t-4})$	1.0000
$f_9(G_{t-1})$	1.6064	0.4549	0.000413 ***	G_{t-1}	0.147	$s(G_{t-1})$	0.2821
$f_{10}(G_{t-2})$	1.2537	0.4592	0.006335 **	G_{t-2}	0.214	$s(G_{t-2})$	0.0543 .
$f_{11}(G_{t-3})$	2.1436	0.7808	0.006042 **	G_{t-3}	0.268	$s(G_{t-3})$	0.2892
$f_{12}(G_{t-4})$	1.1419	0.4337	0.008461 **	G_{t-4}	0.121	$s(G_{t-4})$	0.0592 .
		AIC	1062.1	AIC	1118.4	AIC	1095.604
Signif. codes: *** 0.001 ** 0.01 * 0.05 . 0.1							

GLM model:

$$logit(p_t) \approx \alpha_0 + \sum_{j=1}^{4} \alpha_j v_{t-j} + \sum_{j=1}^{4} \alpha_{4+j} V_{t-j} + \sum_{j=1}^{4} \alpha_{8+j} G_{t-j}, \tag{31}$$

where α_j's are estimated by the GLM in R;

GAM model:

$$logit(p_t) \approx \alpha_0 + \sum_{j=1}^{4} g_j(v_{t-j}) + \sum_{j=1}^{4} g_{4+j}(V_{t-j}) + \sum_{j=1}^{4} g_{8+j}(G_{t-j}), \tag{32}$$

where $g_j(\cdot)$'s are unknown functions estimated by GAM in R with the $s(.)$ functions specifying a smoothing spline fit.

The fitting results of these models are summarised in Table 3. Indeed, AIC is widely applied for model selection. As an indicative only, by the AIC values shown in this table, the MAMaLoR with the used bandwidth of $h = 0.5$ seems preferred to the GLM and the GAM. Here the selected bandwidth of $h = 0.5$ is an indicative only for illustration - it appears to work well. Also as shown, none of the GLM coefficients are significant at 5% level of significance, while the GAM result seems to imply that almost all the variables in model (32) are not useful in explaining the market direction Y_t except the components, v_{t-3}, G_{t-2} and G_{t-4}, the additive functions of which are displayed in Figure 6. Differently, our MAMaLoR model appears to show that the market direction Y_t is significantly correlated to the lagged geometric returns from $t-1$ to $t-4$ through marginal local linear logistic (auto)regression estimates together with an intercept (see Figure 5 on the estimated marginal probabilities of $P(Y_t = 1|G_{t-j} = x_j)$ for $j = 1, 2, 3, 4$). From the above analysis, it appears that one may conclude that the true relationship between Y_t and X_t is not linear. In particular, the MAMaLoR model recognizes the relationship between the lags of the geometric return G_t and the market index moving direction Y_t, which appears reasonable according to the way we set them, while the other models fail to provide relevant information.

In addition, we notice from the MAMaLoR result in Table 3 that, though all the lagged volatility and volume variables seem to be removed from our model, it is possible that a longer range of lags of the geometric return would still be significant and help to explain Y_t. We have hence examined to determine the optimal number of lags for geometric return (G_t) in the MAMaLoR model. The AIC value for each fit with different lags of geometric return is plotted in Figure (7). It appears that the MAMaLoR model improves with more lags, though the following lags of G_t after lag of 21 may not help a lot in explaining Y_t with the change of AIC being small from lag = 21 to lag = 31. We have hence considered a lag order of 31 in our MAMaLoR model fitting.

By removing the insignificant lags of G_t in the model, we obtain a new MAMaLoR model fitting result provided in Table 4 with a much smaller AIC value of 968.74 than those in Table 3. We have further compared the AUC values

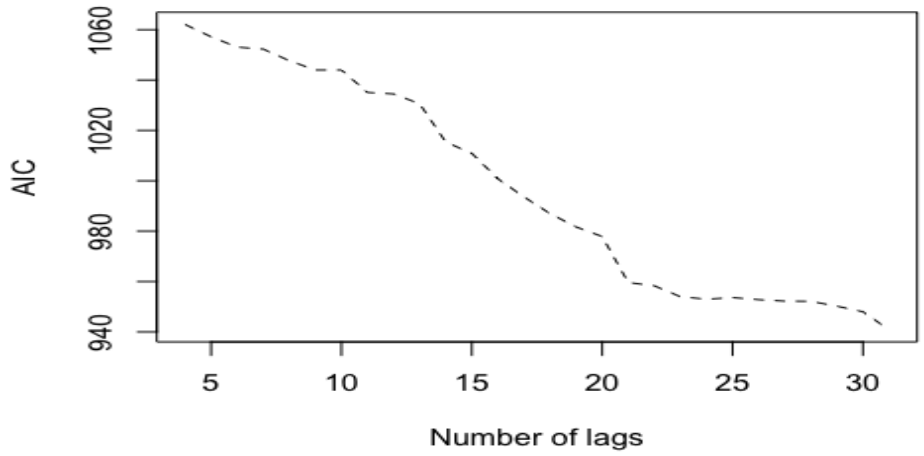

Figure 7: The aic of MAMaLoR model with different number of lagged G_t.

Table 4
MAMaLoR model after lag selection

MAMaLoR model			
	Estimate	Std. Error	Pr(|z|)
Intercept	-1.9983	0.2612	2.01e-14 ***
G_{t-1}	1.7347	0.4848	0.000346 ***
G_{t-3}	2.4553	0.8163	0.002631 **
G_{t-8}	1.0010	0.3262	0.002150 **
G_{t-11}	1.6940	0.6218	0.006439 **
G_{t-13}	1.1286	0.4655	0.015320 *
G_{t-14}	1.1290	0.2811	5.93e-05 ***
G_{t-15}	3.0522	1.1383	0.007332 **
G_{t-16}	1.2039	0.3765	0.001384 **
G_{t-17}	1.5887	0.5373	0.003106 **
G_{t-18}	1.1873	0.5708	0.037528 *
G_{t-21}	1.2115	0.2944	3.87e-05 ***
G_{t-28}	2.1106	0.6844	0.002043 **
G_{t-31}	1.5785	0.8405	0.060367.
AIC		968.74	
Signif. codes: *** 0.001 ** 0.01 * 0.05 . 0.1			

of the forecasting of the market moving direction Y_t based on the significant X_t identified by the above analysis. The group of $X_t = (G_{t-1}, G_{t-3}, G_{t-8}, G_{t-11}, G_{t-13}, G_{t-14}, G_{t-15}, G_{t-16}, G_{t-17}, G_{t-18}, G_{t-21}, G_{t-28}, G_{t-31})$ is identified by the MAMaLoR model given in Table 4. The AUC values with the ROC curves for both the MAMaLoR model with bandwidth selection and the corresponding GLM are investigated.

As is well known, financial return is notoriously difficult to predict, so it is quite understandable that the predictive power of a model on financial return is basically very low with the AUC close to 0.5 for forecasting of price moving direction under an efficient market hypothesis. In this sense, a model that achieves AUC higher than 0.5 for forecasting of the price moving direction is of interest, which indicates some kind of ability in forecasting by the model. The ROC curves together with the AUC values are given in Figure 8. It is clear that the performance of our proposed MAMaLoR is better than that of the GLM model in time series classification prediction, which is promising. Recognising that the

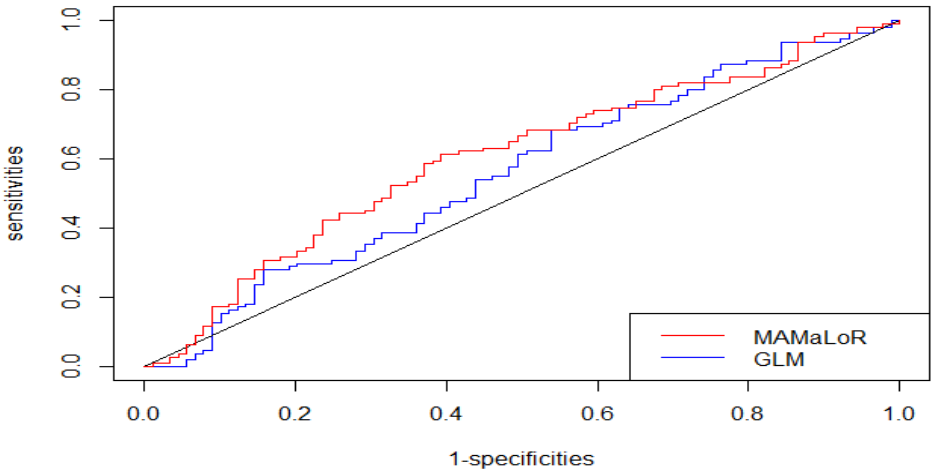

Figure 8: The ROC curves for the MAMaLoR with selected bandwidth (h given in Table 5) and the GLM models. Here the corresponding AUC values for MAMaLoR 0.6041 with h selected, and for GLM it is 0.560, respectively.

Table 5
Bandwidth selected for the 13 significant variables given in Table 4

	G_{t-1}	G_{t-3}	G_{t-8}	G_{t-11}	G_{t-13}	G_{t-14}	G_{t-15}	G_{t-16}	G_{t-17}	G_{t-18}	G_{t-21}	G_{t-28}	G_{t-31}
h selected	0.2287426	0.5396938	0.8646845	0.8266905	1.4065888	1.1340509	1.8326259	0.3591425	1.3146497	0.9988872	1.3914061	1.8377243	1.8388175

market direction may also be influenced by other more factors, there is, henceforth, still a room for our MAMaLoR model to improve its predictability by optimally choosing the lagged information from more explanatory variables.

We comment that the performance of kernel based models, e.g., local linear regression, may depend on the choice of bandwidth. For simplicity, as in the simulation, we used the function h.select available in R package *sm*, which is a direct estimation of $p_{jt} = E(Y_t|x_{jt})$ based on cross validation, to find the bandwidths for the 13 selected predictors given in Table 4. The selected bandwidth h's are summarised in Table 5, used for the MAMaLoR in Figure 8. Again it appears to work well although there is no theoretical guarantee that these h's selected are globally optimal for classification. We leave the investigation of theoretically optimal bandwidth selection to the future work.

5. Conclusion

In this paper, a novel semi-parametric logistic model, namely MAMaLoR, has been proposed to forecast binary time-series classification data with mixing dependence. The consistency and asymptotic normality of the estimator of averaging coefficients are established under mild conditions. A simulation based numerical example is presented to show the strength of our proposed model in forecasting. An application of our MAMaLoR model to forecast of market moving direction of the FTSE100 financial data has further illustrated its power in time series classification forecasting by a comparison with the GAM and GLM models. With more work by careful variable selection, the performance of our proposed model would still improve, which is left for future work. We hope this would contribute to further studies in semiparametric classification models in time series domain, with the future research direction including variable selection and bandwidth selection in high and ultra-high dimension cases.

Acknowledgement

The authors are grateful to the Editor-in-Chief Professor Erricos John Kontoghiorghes, the Associate Editor and two referees for their valuable and constructive comments and suggestion, which have greatly helped to improve the

presentation of this paper.

Appendix

A. Sketch of Proof of Theorem 3.1

Proposition 3.1 (Consistency Lemma), given in Section 3, follows from Lemma 4.1 in Lu et al. (2007). The consistency of $\hat{\boldsymbol{\alpha}}$ can be proved by checking the conditions specified in Proposition 3.1. As $\hat{\boldsymbol{\alpha}}$ and $\boldsymbol{\alpha}^{(0)}$ are the maximizers of $l_n(\hat{\mathbf{f}}(\cdot), \boldsymbol{\alpha})$ and $l(\mathbf{f}_0(\cdot), \boldsymbol{\alpha})$, respectively, (i) and (ii) of Proposition 3.1 hold obviously. (iii) of Proposition 3.1 also holds clearly by the following fact:

$$l(\mathbf{f}(\cdot), \boldsymbol{\alpha}) = E[Y_t \widetilde{\chi}_t(\mathbf{f})^T \boldsymbol{\alpha} - \log(1 + e^{\widetilde{\chi}_t(\mathbf{f})^T \boldsymbol{\alpha}})], \tag{33}$$

where $\widetilde{\chi}_t(\mathbf{f}) = (1, f_1(x_{1t})..., f_d(x_{dt}))^T$ with f_j's being marginal functions that are generally different from those in $\mathbf{f}_0$ given in Assumption A4 at a cost of slight notation confusion.

Then:

$$\begin{aligned}
&\sup_{\boldsymbol{\alpha}\in\mathfrak{A}} |l(\mathbf{f}(\cdot), \boldsymbol{\alpha}) - l(\mathbf{f}_0(\cdot), \boldsymbol{\alpha})| \\
&\leq E|Y_t| \|\widetilde{\chi}_t(\mathbf{f}) - \widetilde{\chi}_t(\mathbf{f}_0)\| \boldsymbol{\alpha} + |\log(1 + e^{\widetilde{\chi}_t(\mathbf{f})^T \boldsymbol{\alpha}}) - \log(1 + e^{\widetilde{\chi}_t(\mathbf{f}_0)^T \boldsymbol{\alpha}})| \\
&\leq \frac{e^{\widetilde{\chi}_t(\mathbf{f})^T \boldsymbol{\alpha}}}{1 + e^{\widetilde{\chi}_t(\mathbf{f})^T \boldsymbol{\alpha}}} \|\widetilde{\chi}_t(\mathbf{f}) - \widetilde{\chi}_t(\mathbf{f}_0)\| \|\boldsymbol{\alpha}\| \\
&\leq C\|\mathbf{f} - \mathbf{f}_0\|_{\mathbf{F}},
\end{aligned} \tag{34}$$

where C is a generic constant.

Now, to prove (iv) of Proposition 3.1, we show that the estimator $\hat{f}_j(.)$ replacing $f_j(.)$ function in the model averaging step is uniformly consistent. The proof for the local fitting technique is given as follows. It is similar to that of Nielsen (2005) under $i.i.d.$ data, but we are concerned with time series data process of β-mixing as defined in Subsection 3.2.

The non-linear logistic regression can be formulated as follows:

$$\text{logit}(p_j(x_{jt})) = \log(\frac{p_j(x_{jt})}{1 - p_j(x_{jt})}) = f_j(x_{jt}), \tag{35}$$

where $p_j(x_{jt}) = P(Y_t = 1|x_{jt})$ and $1 - p_j(x_{jt}) = P(Y_t = 0|x_{jt})$, and $f_j(\cdot)$ is a nonparametric function from R to R.

Given the local log likelihood function in (9), we have the following types of estimation equations:

$$\Omega_n^{(1)}(\boldsymbol{\beta}, x_j, h) = \frac{1}{n}\frac{\partial \ell}{\partial \beta_1} = \frac{1}{n}\sum_{t=1}^{n}\left[Y_t - \frac{\exp(\beta_1 + \beta_2^T(x_{jt} - x_j))}{1 + \exp(\beta_1 + \beta_2^T(x_{jt} - x_j))}\right] K_h(x_{jt} - x_j) = 0, \tag{36}$$

$$\Omega_n^{(2)}(\boldsymbol{\beta}, x_j, h) = \frac{1}{nh}\frac{\partial \ell}{\partial \beta_2} = \frac{1}{n}\sum_{t=1}^{n}\left[Y_t - \frac{\exp(\beta_1 + \beta_2^T(x_{jt} - x_j))}{1 + \exp(\beta_1 + \beta_2^T(x_{jt} - x_j))}\right] \frac{x_{jt} - x_j}{h} K_h(x_{jt} - x_j) = 0. \tag{37}$$

Intuitively, if $\Omega_n(\boldsymbol{\beta}, x_j, h) = (\Omega_n^{(1)}(\boldsymbol{\beta}, x_j, h), \Omega_n^{(2)}(\boldsymbol{\beta}, x_j, h))^T$ is uniformly close to $E[\Omega_n(\boldsymbol{\beta}, x_j, h)]$ in $x_j \in A_j = [c_{j0}, c_{1j}]$, then $\hat{\boldsymbol{\beta}}$ should be close to the solution of $E[\Omega_n(\boldsymbol{\beta}, x_j, h)] = 0$, and is a consistent estimator of $\boldsymbol{\beta}_0$. We first check $\boldsymbol{\beta}_0$ is close to the solution to $E[\Omega_n(\boldsymbol{\beta}, x_j, h)] = 0$ with our local maximum likelihood estimation under model (35):

$$\begin{aligned}
E[\Omega_n^{(1)}(\boldsymbol{\beta}, x_j, h)] &= E\left[\frac{1}{n}\sum_{t=1}^{n}\left(Y_t - \frac{\exp(\beta_1 + \beta_2(x_{jt} - x_j))}{1 + \exp(\beta_1 + \beta_2(x_{jt} - x_j))}\right) K_h(x_{jt} - x_j)\right] \\
&= E\left\{\frac{1}{n}\sum_{t=1}^{n} E\left[\left(Y_t - \frac{\exp(\beta_1 + \beta_2)(x_{jt} - x_j)}{1 + \exp(\beta_1 + \beta_2(x_{jt} - x_j))}\right) K_h(x_{jt} - x_j)|x_{jt}\right]\right\}
\end{aligned}$$

$$= E\left[\frac{1}{n}\sum_{t=1}^{n}\left(E[Y_t|x_{jt}] - \frac{\exp(\beta_1+\beta_2(x_{jt}-x_{j0}))}{1+\exp(\beta_1+\beta_2(x_{jt}-x_j))}\right)K_h(x_{jt}-x_j)\right]$$
$$= E\left[\frac{1}{n}\sum_{t=1}^{n}\left(\frac{\exp(f_j(x_{jt}))}{1+\exp(f_j(x_{jt}))} - \frac{\exp(\beta_1+\beta_2(x_{jt}-x_j))}{1+\exp(\beta_1+\beta_2(x_{jt}-x_j))}\right)K_h(x_{jt}-x_j)\right],$$

where note that $E[Y_t|x_{jt}] = \frac{\exp(f_j(x_{jt}))}{1+\exp(f_j(x_{jt}))}$.

Let $\tilde{f}(z_j) = \frac{e^{z_j}}{1+e^{z_j}}$. Then by Taylor expansion together with assumptions A4 and A2 we find:

$$E\left[\Omega_n^{(1)}(\boldsymbol{\beta}, x_j, h)\right] = E\left[\frac{1}{n}\sum_{t=1}^{n}\left(\tilde{f}(f_j(x_{jt})) - \tilde{f}(\beta_1+\beta_2(x_{jt}-x_j))\right)K_h(x_{jt}-x_j)\right]$$
$$= (1+o(1))[\tilde{f}(f_j(x_{jt})) - \tilde{f}(\beta_1)]g_j(x_j),$$

where $o(1)$ is uniform in $x \in A$ owing to Assumption A4, and g_j is the marginal probability density function of x_{jt}. In fact, if we denote $\Phi(Y_t, z_j) = Y_t - \exp(z_j)/[1+\exp(z_j)]$ as in Assumption A5, then

$$E[\Omega_n^{(1)}(\boldsymbol{\beta}, x_j, h)] = E[\Phi(Y_t; \beta_1+\beta_2(x_{jt}-x_j))K_h(x_{jt}-x_j)]$$
$$= E[m(x_j; \beta_1+\beta_2(x_{jt}-x_j))K_h(x_{jt}-x_j)]$$
$$= m(x_j, \beta_1)g_j(x_j) + O(h^2), \tag{38}$$

where, corresponding to our local logistic regression, $m(x_j, \beta_1) = \tilde{f}(f_j(x_j)) - \tilde{f}(\beta_1)$, and the O-term does not depend on $x \in A$ nor on $\beta_1 = f_j(x_j)$ which is the j-th component of $\mathbf{f}_0(\cdot)$ owing to Assumption A4.

Similarly,

$$E[\Omega_n^{(2)}(\boldsymbol{\beta}, x_j, h)] = E\left[\Phi(Y_i; \beta_1+\beta_2(x_{jt}-x_j))\frac{x_{jt}-x_j}{h}K_h(x_{jt}-x_j)\right]$$
$$= h(\beta_2 m_2'(x_j, \beta_1)) + m_1'(x_j, \beta_1))g_j(x_j) + hm(x_j, \beta_1)g_j'(x_j) + O(h^3),$$

where, corresponding to our local logistic regression model, $m_1'(x_j, \beta_1) = \tilde{f}'(f_j(x_j))f_j'(x_j) = f_j'(x_j)\frac{e^{f_j(x_j)}}{(1+e^{f_j(x_j)})^2}$ and $m_2'(x_j, \beta_1) = -\tilde{f}'(\beta_1) = -\frac{e^{\beta_1}}{(1+e^{\beta_1})^2}$, with $\tilde{f}'(z_j) = e^{z_j}/(1+e^{z_j})^2$ as defined above, and the O-term is uniform with respect to $x \in A$.

Thus we get:

$$E[\Omega_n^{(1)}(\boldsymbol{\beta}, x_j, h)] = \Omega_0^{(1)}(\boldsymbol{\beta}, x_j) + O(h^2), \tag{39}$$

and

$$E[\Omega_n^{(2)}(\boldsymbol{\beta}, x_j, h)] = h\Omega_0^{(2)}(\boldsymbol{\beta}, x_j) + O(h^3), \tag{40}$$

where

$$\Omega_0^{(1)}(\boldsymbol{\beta}, x_j) = m(x_j, \beta_1)g_j(x_j), \tag{41}$$

$$\Omega_0^{(2)}(\boldsymbol{\beta}, x_j) = (\beta_2 m_2'(x_j, \beta_1) + m_1'(x_j, \beta_1))g_j(x_j) + m(x_j, \beta_1)g_j'(x_j). \tag{42}$$

Denote by $\boldsymbol{\beta}_0 = (\beta_{01}, \beta_{02})$ the solution to $\boldsymbol{\Omega}_0(\boldsymbol{\beta}, x_j) = 0$, where $\boldsymbol{\Omega}_0(\boldsymbol{\beta}, x_j) = (\Omega_0^{(1)}(\boldsymbol{\beta}, x_j), \Omega_0^{(2)}(\boldsymbol{\beta}, x_j))^T$. Then we have:

$$\begin{cases} m(x_j, \beta_{01}) = 0 \\ \beta_{02}(x) = -\frac{m_1'(x_j, \beta_{01})}{m_2'(x_j, \beta_{02})}, \end{cases} \tag{43}$$

which is actually unique correspondingly to our local linear logistic regression (9) with $\beta_{01} = f_j(x_j)$ and $\beta_{02} = f_j'(x_j)$.

For $\Omega_0^{(i)}(\boldsymbol{\beta}, x_j)$, $i = 1, 2$, we further know from the above that $\Omega_0^{(i)}(\boldsymbol{\beta}, x_j)$ is continuous in $\boldsymbol{\beta} \in \mathbf{F}$ (in Lipschitz norm) and $x \in A$ (in Euclidean norm) owing to Assumption A4. Therefore, for any $\varepsilon > 0$, there exists $\delta > 0$ such that

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_\infty > \delta \Rightarrow \max_{i=1,2} |\Omega_0^{(i)}(\hat{\boldsymbol{\beta}}, x_j)| > \varepsilon, \quad for \;\; x \in A. \tag{44}$$

Therefore for the uniform consistency of $\hat{\boldsymbol{\beta}}$ to $\boldsymbol{\beta}_0$ in probability, by (44), it suffices to show $\max_{i=1,2} \sup_{x\in A} |\Omega_0^{(i)}(\hat{\boldsymbol{\beta}}, x_j)|$ $= \max_{i=1,2} \sup_{x\in A} |\Omega_0^{(i)}(\hat{\boldsymbol{\beta}}, x_j) - \Omega_0^{(i)}(\boldsymbol{\beta}_0, x_j)| \to 0$ in probability as $n \to \infty$. This follows from

$$\max_{i=1,2} \sup_{\|\boldsymbol{\beta}\|_{\mathbf{F}} \le C} \sup_{x_j \in A_j} |\Omega_n^{(i)}(\boldsymbol{\beta}, x_j) - \Omega_0^{(i)}(\boldsymbol{\beta}, x_j)| \to 0,$$

as $n \to \infty$, which is easily proved under Assumptions A1–A4 (c.f., Lu et al. (2007)) with details omitted. The proof of (iv) of Proposition 3.1 is done.

To check (v) of Proposition 3.1, let $\delta_n = o(1)$ and $\|\mathbf{f} - \mathbf{f}_0\|_{\mathbf{F}} \le \delta_n$. Then we have:

$$\begin{aligned} l_n(\mathbf{f}(\cdot), \boldsymbol{\alpha}) - l(\mathbf{f}(\cdot), \boldsymbol{\alpha}) &= \{l_n(\mathbf{f}(\cdot), \boldsymbol{\alpha}) - l_n(\mathbf{f}_0(\cdot), \boldsymbol{\alpha})\} + \{l_n(\mathbf{f}_0(\cdot), \boldsymbol{\alpha}) - l(\mathbf{f}_0(\cdot), \boldsymbol{\alpha})\} + \{l(\mathbf{f}_0(\cdot), \boldsymbol{\alpha}) - l(\mathbf{f}(\cdot), \boldsymbol{\alpha})\} \\ &= I + II + III. \end{aligned} \tag{45}$$

Uniformly, for $\boldsymbol{\alpha} \in \mathfrak{A}$ and $\mathbf{f}$ satisfying $\|\mathbf{f} - \mathbf{f}_0\|_{\mathbf{F}} \le \delta_n$, I, II and III can be proved to tend to zero. It is easy to show that III tending to zero follows from equation (34) and II tending to zero is easily proved by the law of large number together with $\mathfrak{A}$ being a compact set. Note that III is the expected value of I. That I tends 0 can be proved similarly. Hence we know that $I + II + III$ tends to zero.

By completing the checking of the conditions of Proposition 3.1 (Consistency Lemma), the proof of Theorem 3.1 is completed.

B. Sketch of Proof of Theorem 3.2

Now we will derive the asymptotic normality. Note that $l_n(\mathbf{f}(\cdot), \boldsymbol{\alpha})$ and $l(\mathbf{f}(\cdot), \boldsymbol{\alpha})$ are differentiable with respect to $\boldsymbol{\alpha}$. By applying simple algebraic operations, we can obtain and denote the derivatives as follows:

$$\begin{aligned} l_n'(\mathbf{f}(\cdot), \boldsymbol{\alpha}) &= \frac{1}{n}\sum_{t=1}^{n}[(Y_t - p_t(\mathbf{f}, \boldsymbol{\alpha}))\widetilde{\chi}_t(\mathbf{f})]w(X_t), \\ l'(\mathbf{f}(\cdot), \boldsymbol{\alpha}) &= E[(Y_t - p_t(\mathbf{f}, \boldsymbol{\alpha}))\widetilde{\chi}_t(\mathbf{f})]w(X_t), \\ l_n''(\mathbf{f}(\cdot), \boldsymbol{\alpha}) &= -\frac{1}{n}\sum_{t=1}^{n} p_t(\mathbf{f}, \boldsymbol{\alpha})(1 - p_t(\mathbf{f}, \boldsymbol{\alpha}))\widetilde{\chi}_t(\mathbf{f})\widetilde{\chi}_t(\mathbf{f})^T w(X_t), \end{aligned}$$

and

$$\begin{aligned} l''(\mathbf{f}(\cdot), \boldsymbol{\alpha}) &= \frac{\partial l'(\mathbf{f}(\cdot), \boldsymbol{\alpha})}{\partial \alpha} = \frac{\partial E[(Y_t - \frac{e^{\alpha\widetilde{\chi}_t(\mathbf{f})}}{1+e^{\alpha\widetilde{\chi}_t(\mathbf{f})}})\widetilde{\chi}_t(\mathbf{f})]w(X_t)}{\partial \alpha} \\ &= E[-\frac{e^{\alpha\widetilde{\chi}_t(\mathbf{f})}}{1 + e^{\alpha\widetilde{\chi}_t(\mathbf{f})}} \cdot \frac{1}{1 + e^{\alpha\widetilde{\chi}_t(\mathbf{f})}} \cdot \widetilde{\chi}_t(\mathbf{f}) \cdot \widetilde{\chi}_t(\mathbf{f})^T]w(X_t) \\ &= E[-p_t(\mathbf{f}, \boldsymbol{\alpha})(1 - p_t(\mathbf{f}, \boldsymbol{\alpha}))\widetilde{\chi}_t(\mathbf{f})\widetilde{\chi}_t(\mathbf{f})^T]w(X_t), \end{aligned}$$

where $\widetilde{\chi}_t(\mathbf{f}) = (1, f_1(x_{1t})..., f_d(x_{dt}))^T$.

We apply the Taylor expansion:

$$\begin{aligned} 0 = l_n'(\hat{\mathbf{f}}(\cdot), \hat{\boldsymbol{\alpha}}) &= l_n'(\hat{\mathbf{f}}(\cdot), \boldsymbol{\alpha}^{(0)}) + l_n''(\hat{\mathbf{f}}(\cdot), \boldsymbol{\alpha}^{(0)} + \xi(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^{(0)}))(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^{(0)}), \\ &\text{and} \\ \sqrt{n}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^{(0)}) &= -[l_n''(\hat{\mathbf{f}}(\cdot), \boldsymbol{\alpha}^{(0)} + \xi(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^{(0)}))]^{-1}\sqrt{n} l_n'(\hat{\mathbf{f}}(\cdot), \boldsymbol{\alpha}^{(0)}), \end{aligned}$$

where $|\xi| < 1$. Then we have, together with the consistency of $\hat{\boldsymbol{\alpha}}$ to $\boldsymbol{\alpha}^{(0)}$,

$$\sqrt{n}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^{(0)}) = -(1 + op(1))[l_n^{''}(\mathbf{f}_0(\cdot), \boldsymbol{\alpha}^{(0)})]^{-1}\sqrt{n}[l_n^{'}(\mathbf{f}_0(\cdot), \boldsymbol{\alpha}^{(0)}) + O(h^2)], \tag{46}$$

by noting that

$$\begin{aligned} l_n^{'}(\hat{\mathbf{f}}, \boldsymbol{\alpha}^{(0)}) - l_n^{'}(\mathbf{f}_0, \boldsymbol{\alpha}^{(0)}) &= (1 + o_P(1))\frac{1}{n}\sum_{t=1}^{n}[(Y_t - p_t)(\widetilde{\chi}_t(\hat{\mathbf{f}}) - \widetilde{\chi}_t(\mathbf{f}_0))w(X_t)] \\ &= O_P(h^2), \end{aligned} \tag{47}$$

owing to the uniform consistency of $\hat{\mathbf{f}}$ to $\mathbf{f}_0$ and $E[\hat{\mathbf{f}}] - \mathbf{f}_0 = O(h^2)$ as we have proved.

Note that

$$l_n^{''}(\mathbf{f}_0, \boldsymbol{\alpha}^{(0)}) = -\frac{1}{n}\sum_{t=1}^{n} p_t(\mathbf{f}_0, \boldsymbol{\alpha}^{(0)})(1 - p_t(\mathbf{f}_0, \boldsymbol{\alpha}^{(0)}))\widetilde{\chi}_t(\mathbf{f}_0)\widetilde{\chi}_t(\mathbf{f}_0)^T w(X_t). \tag{48}$$

By law of large number, we have

$$l_n^{''}(\mathbf{f}_0, \boldsymbol{\alpha}^{(0)}) \to l^{''}(\mathbf{f}_0, \boldsymbol{\alpha}^{(0)}) = \mathbf{U} = E[-p_t(1 - p_t)\widetilde{\chi}_t(\mathbf{f}_0)\widetilde{\chi}_t(\mathbf{f}_0)^T]w(X_t). \tag{49}$$

By central limit theorem,

$$\sqrt{n}l_n^{'}(\mathbf{f}_0, \boldsymbol{\alpha}^{(0)}) \to N(0, \mathbf{V}), \tag{50}$$

where

$$\mathbf{V} = \lim_{n\to\infty} Var(\frac{1}{\sqrt{n}}(Y_t - p_t)\widetilde{\chi}_t(\mathbf{f}_0)w(X_t)). \tag{51}$$

Thus the asymptotic variance matrix

$$Var(\hat{\boldsymbol{\alpha}}|\mathbf{f}(\cdot)) = \mathbf{U}^{-1}\mathbf{V}\mathbf{U}^{-1}. \tag{52}$$

The asymptotic normality of $\hat{\boldsymbol{\alpha}}$ hence follows.

References

Al-Sulami, D., Jiang, Z., Lu, Z., & Zhu, J. (2017). Estimation for semiparametric nonlinear regression of irregularly located spatial time-series data. *Econometrics and Statistics*, *2*, 22–35.

Ballings, M., Van den Poel, D., Hespeels, N., & Gryp, R. (2015). Evaluating multiple classifiers for stock price direction prediction. *Expert Systems with Applications*, *42*, 7046–7056.

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

Buckley, D., & Bulger, D. (2012). Trends and weekly and seasonal cycles in the rate of errors in the clinical management of hospitalized patients. *Chronobiology International*, *29*, 947–954.

Chen, J., Li, D., Linton, O., & Lu, Z. (2016). Semiparametric dynamic portfolio choice with multiple conditioning variables. *Journal of Econometrics*, *194*, 309–318.

Chen, J., Li, D., Linton, O., & Lu, Z. (2018). Semiparametric ultra-high dimensional model averaging of nonlinear dynamic time series. *Journal of the American Statistical Association*, *113*, 919–932.

Cox, D. R., & Snell, E. J. (1989). *Analysis of binary data* volume 32. CRC press.

Davis, R. A., Dunsmuir, W. T., & Wang, Y. (1999). Modeling time series of count data. *Statistics Textbooks and Monographs*, *158*, 63–114.

Davis, R. A., Holan, S. H., Lund, R., & Ravishanker, N. (2016). *Handbook of discrete-valued time series*. CRC Press.

Davis, R. A., & Wu, R. (2009). A negative binomial model for time series of counts. *Biometrika*, *96*, 735–749.

Doukhan, P., Massart, P., & Rio, E. (1995). Invariance principles for absolutely regular empirical processes. In *Annales de l'IHP Probabilités et statistiques* (pp. 393–427). volume 31.

Fan, J., Farmen, M., & Gijbels, I. (1998a). Local maximum likelihood estimation and inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *60*, 591–608.

Fan, J., & Gijbels, I. (1995). Adaptive order polynomial fitting: bandwidth robustification and bias reduction. *Journal of Computational and Graphical Statistics*, *4*, 213–227.

Fan, J., Härdle, W., & Mammen, E. (1998b). Direct estimation of low-dimensional components in additive models. *The Annals of Statistics*, *26*, 943–971.

Fan, J., & Yao, Q. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika*, *85*, 645–660.

Fan, J., & Yao, Q. (2003). *Nonlinear time series: nonparametric and parametric methods*. Springer Science & Business Media.

Fan, J., Yao, Q., & Cai, Z. (2003). Adaptive varying-coefficient linear models. *Journal of the Royal Statistical Society: series B (statistical methodology)*, *65*, 57–80.

Ferland, R., Latour, A., & Oraichi, D. (2006). Integer-valued garch process. *Journal of Time Series Analysis*, *27*, 923–942.

Gao, J. (2007). *Nonlinear time series: semiparametric and nonparametric methods*. CRC Press.

Hastie, T., & Tibshirani, R. (1987). Generalized additive models: some applications. *Journal of the American Statistical Association*, *82*, 371–386.

Hofert, M., Prasad, A., & Zhu, M. (2021). Multivariate time-series modeling with generative neural networks. *Econometrics and Statistics*, . doi:https://doi.org/10.1016/j.ecosta.2021.10.011.

Jacobs, P. A., & Lewis, P. A. (1978). Discrete time series generated by mixtures. i: Correlational and runs properties. *Journal of the Royal Statistical Society. Series B (Methodological)*, (pp. 94–105).

Jones, M., Davies, S., & Park, B. (1994). Versions of kernel-type regression estimators. *Journal of the American Statistical Association*, *89*, 825–832.

Lahiri, K., & Yang, L. (2016). A non-linear forecast combination procedure for binary outcomes. *Studies in Nonlinear Dynamics & Econometrics*, *20*, 421–440.

Li, D., Linton, O., & Lu, Z. (2015). A flexible semiparametric forecasting model for time series. *Journal of Econometrics*, *187*, 345–357.

Liesenfeld, R., Nolte, I., & Pohlmeier, W. (2006). Modelling financial transaction price movements: a dynamic integer count data model. *Empirical Economics*, *30*, 795–825.

Lu, Z. (1998). On the geometric ergodicity of a non-linear autoregressive model with an autoregressive conditional heteroscedastic term. *Statistica Sinica*, *8*, 1205–1217.

Lu, Z., Tjøstheim, D., & Yao, Q. (2007). Adaptive varying-coefficient linear models for stochastic processes: asymptotic theory. *Statistica Sinica*, *17*, 177–198.

McKenzie, E. (1985). Some simple models for discrete variate time series. *JAWRA Journal of the American Water Resources Association*, *21*, 645–650.

Nielsen, S. F. (2005). Local linear estimating equations: Uniform consistency and rate of convergence. *Nonparametric Statistics*, *17*, 493–511.

de Oliveira Maia, G., Barreto-Souza, W., de Souza Bastos, F., & Ombao, H. (2021). Semiparametric time series models driven by latent factor. *International Journal of Forecasting*, .

Ryabko, D., & Mary, J. (2013). A binary-classification-based metric between time-series distributions and its use in statistical and learning problems. *The Journal of Machine Learning Research*, *14*, 2837–2856.

Rydberg, T. H., & Shephard, N. (2003). Dynamics of trade-by-trade price movements: decomposition and models. *Journal of Financial Econometrics*, *1*, 2–25.

Seifert, B., & Gasser, T. (1996). Finite-sample variance of local polynomials: analysis and solutions. *Journal of the American Statistical Association*, *91*, 267–275.

Shephard, N. (1995). *Generalized linear autoregressions.*. Technical Report Nuffield College, Oxford.

Terasvirta, T., Tjostheim, D., Granger, C. W. et al. (2010). Modelling nonlinear economic time series. *OUP Catalogue*, .

Tibshirani, R., & Hastie, T. (1987). Local likelihood estimation. *Journal of the American Statistical Association*, *82*, 559–567.

Tong, H. (1990). *Non-linear time series: a dynamical system approach*. Oxford University Press.

Turner, R., Hayen, A., Dunsmuir, W., & Finch, C. F. (2011). Air temperature and the incidence of fall-related hip fracture hospitalisations in older people. *Osteoporosis International*, *22*, 1183–1189.

Van Der Vaart, A. W., & Wellner, J. A. (1996). Weak convergence. In *Weak convergence and empirical processes* (pp. 16–28). Springer.

Varin, C., Reid, N., & Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, (pp. 5–42).

Waller, L. A., Carlin, B. P., Xia, H., & Gelfand, A. E. (1997). Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical association*, *92*, 607–617.

Zhang, X., Yu, D., Zou, G., & Liang, H. (2016). Optimal model averaging estimation for generalized linear models and generalized linear mixed-effects models. *Journal of the American Statistical Association*, *111*, 1775–1790.