UNIVERSITY OF SOUTHAMPTON

FACULTY OF SOCIAL, HUMAN AND MATHEMATICAL SCIENCES

Mathematical Sciences

Generalised Dynamic Nonlinear Time Series Regression and Forecasting: Theory with Applications

by

Rong Peng

Thesis submitted for the degree of Doctor of Philosophy

October 2021

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SOCIAL, HUMAN AND MATHEMATICAL SCIENCES
Mathematical Sciences

Doctor of Philosophy

GENERALISED DYNAMIC NONLINEAR TIME SERIES REGRESSION AND FORECASTING: THEORY WITH APPLICATIONS

by Rong Peng

This thesis aims to develop a series of nonlinear time series models for analysing count data, especially to overcome the "curse of dimensionality" for high and ultra-high dimensions. This is of particular needs for big data analysis in applications to discrete-valued outcome events, such as financial market direction, infected patients number in epidemiology and etc., where the nature of data is often unknown.

In contrast to time series for continuous responses, where numerous related studies are available, literature paid scant attention to discrete-valued time series estimation and forecasting. Existing studies are developed based on the extension of classic AutoRegressive Moving Averge model (ARMA). To better capture the relationship between response and exogenous variables, we have proposed a semi-parametric procedure called the " Generalised Model Averaging MArginal nonlinear Regressions (GMAMaR) and showed the uniform consistency for local maximum likelihood estimation of one dimensional non-parametric local linear estimation. The asymptotic properties of the procedure are established under mild conditions on the time series observations that are of β-mixing property. This model has overcame the "curse of dimensionality" by taking the advantage of cheap computational cost of low dimensional estimation and the idea of model averaging to approximate the true estimates.

In particular, to deal with the popular binary classification problem, we study a special case of logistic regression, namely "Model Averaging MArginal nonlinear

LOgistic Regressions (MAMaLOR). This is the case where binary outcome is considered. The performance of our proposed model is superior when compared to conventional method with numerical examples.

We notice another problem when facing big data that only a few of them are truly useful in explaining the responses out of hundreds and thousands exogenous variables. Thus, we propose a penalise maximum likelihood estimation for variable selection combined with our developed model by utilising adaptive LASSO as a tool. A new computational procedure is also suggested to solve the proposed penalised likelihood estimation. By extracting important information from data, the performance of our proposed methods is improved significantly both in estimation and in prediction.

Last but not least, with the on-going event of COVID-19 in the UK, we further consider the spatial effects along with temporal dependency. The idea is thus to extend time series analysis to the domain of spatio-temporal modelling. We utilise proposed model to investigate impacts of micro variables of the implementation of lockdown on the daily number of confirmed cases. The results are consistent with the consensus of epidemiology studies, and deeper understandings of how to adapt and prioritise the policies in the combat of epidemic are also provided.

To conclude, the proposed series of nonlinear time series models show great potential in the context of discrete-valued events. While providing a more accurate estimation and prediction, the models also offer a better interpretability and deeper understanding of the relationships between response and potential factors. We hope to demonstrate that this thesis thus contribute to the development of this area, and could be further extended to the area of sptio-temporal and other areas of applications.

Contents

List of Figures

List of Tables

Declaration of Authorship

I, Rong Peng , declare that the thesis entitled *Generalised Dynamic Nonlinear Time Series Regression and Forecasting: Theory with Applications* and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- none of this work has been published before submission

Signed:.....

Date:.....

Acknowledgements

My deepest gratitude goes first and foremost respect to my main supervisor, Professor Zudi Lu, who gave me constant encouragement and support throughout the study of my Ph.D. His excellent knowledge in statistics and inspiring passion of research have motivated me to be an independent researcher in the area of time series analysis. I am appreciate his guidance that contributes enormously to this thesis.

I would also like to express my sincere thanks to Dr. Ramin Okhrati (Second supervisor) who has now jointed University College London, Professor Hou-duo Qi, Professor Wei Liu and Professor Stefanie Biedermann (Independent assessors for annual reviews). Their valuable help and suggestion provide me various interesting ideas about improving my work.

I am also grateful to the great research environment with my colleagues of the School of Mathematical Sciences, where I have expanded my knowledge in statistics to a wider scope of topics through the weekly seminars organised within S3RI.

Last but not least, I would like to express my heartfelt gratitude to my beloved parents and little brother for their endless love and support all through my life. I thank with love to my boyfriend, especially for his accompany and support through this hard period of COVID-19.

To my beloved family, friends and colleagues

Chapter 1

Introduction

1.1 Research background

With the rapid development of large scale data analysis, time-series data now plays a crucial role in today's business. Financial companies are dealing with high frequency trades, stocks and options; Banks and insurance companies need to decide the credit allowance for different applicant with their past information; Even suppliers and manufactures are using past data to predict the future demands. All these practical needs are calling for more precise and concise tools, e.g., models with better interpretability, to provide insights of the data rather than a black-box for the user to better understand the nature of it.

Time-series models have been extensively studied in literature, where most of them are for continuous responses. In the contrast, literature on discrete-valued time-series estimation and forecasting is still very limited. However, such data is common in practice (Brown, 2004). For instance, the binary type of time-series data is often seen in credit scoring, predicting the possibility of natural hazard and the success rate of human activities (campaigns, sport games and etc.). It is also assumed in epidemiology that the infect rate of pandemic follows a poisson or negative-binomial distribution. Researchers in these areas need to study the data to understand the on-going events and thus help to make better decisions, e.g., to predict the impact of epidemic and to help reduce its damage. The importance of these applications therefore calls a further research in this area.

Traditionally, e.g., in generalised linear regression (GLM)(McCullagh and Nelder, 1989), modelling discrete-valued data and continuous-valued data requires only different link functions. However, in the case of time series analysis where the dependence of data is present, such techniques developed based on the assumption of independent and identically distributed (i.i.d) cannot be applied directly. As to the classic time series models, e.g., ARIMA(Box et al., 2015) and GARCH(Bollerslev, 1986)(Taylor, 2008), the results are not guaranteed to be integers due to the lack of such constraints when applied to count data. Moreover, since they are originally developed for autoregressive procedures, including exogenous variables, e.g., in ARMAX model, is therefore not straightforward and requires treatment for the aim of interpretation.

The first ever discrete-valued time series model can be tracked back to Jacobs and Lewis (1978). They have proposed the DARMA (Discrete mixed AutoRegressive-Moving Average) processes, in which the correlation structure of the process is determined by parameters and the marginal distributions. However its long term performance is not as good as expected. McKenzie (1985) proposed the INARMA (INteger-valued AutoRegressive-Moving Average), which is still extensively used in today. Based on the work of Shephard (1995), Generalised Linear Autoregressive Moving Average model (GLARMA) has been applied in many different fields. For example, Rydberg and Shephard (2003) and Liesenfeld et al. (2006) in financial modelling, Turner et al. (2011) and Buckley and Bulger (2012) in epidemiological assessments and clinical management individually. GLARMA model assumes a state process depending linearly on covariates and non-linearly on past values. The observation is independent and has an exponential distribution. So for time series with long time period or large number of individuals, it is comparatively easily to fit than other parameter-driven models. Davis et al. (1999) provide a review of varieties to modelling discrete-time series. For more relevant information, please refer to Davis et al. (2016), where a comprehensive review has been provided.

Consider an example of credit scoring that if the bank needs to make a decision on the acceptance of loans for a person based on his past information, it is natural to assume the person's financial ability or the probability of defaulting has dependence of its past values and hundreds or thousands of other information, e.g., career, earnings, age, marriage status and etc. In the case where a new applicant is present, i.e., we don't have his or her past credits, the evaluation must be done based on the other information, and thus the interpretation of the relationships

between the exogenous covariates and defaulting rate is important. Moreover, due to the unknown nature of data, strong assumptions such as linear relationship are hard to be appreciated. Hence, a novel model with better interpretability and robustness for discrete time-series data is in a timely need.

Due to the development of modern techniques, we are now able to collect large-scale data that are of high and ultra-high dimensions. When applying data-driven techniques, e.g., non-parametric models, the computational capacity would limit the number of dimensionality to be considered. This is known as the "curse of dimensionality", which suggests that they are not applicable to high dimension data sample due to high computational costs.

Intuitively, to overcome such difficulty, one would consider to reduce the dimensionality. Recently, a novel semi-parametric model for continuous data proposed by Li et al. (2015), namely the Model Average MArginal Regression (MAMAR), presents an idea that one can first extract one dimensional information and then combine them together as a kind of model averaging. Since the computational cost for low dimension estimation is cheap, the "curse of dimensionality" is thus overcame.

In this thesis, we therefore follow this idea to develop our discrete-valued time series model for big data. In addition, as it is nearly impossible to select the "true" variables that are related to the response according to human experience, variable selection techniques, e.g., LASSO(Tibshirani, 1996), will be considered to extract important information. That is, a penalty term is added to the model and the coefficients of non-correlated variables would thus be forced to (near) zero.

Last but not least, in the case where the mixing time series data are collected from multiple regions, the dependency of data may be not only in the manner of time but also of space. For instance, it is known in epidemiology that the infect rates in different locations are different (Avery et al., 2020). It is, of course, subjected to the local population, medical resources and etc., but also impacted by the development of epidemic nearby. To better capture this type of spatial impacts along with the nature of time dependency, a spatio-temporal model for discrete-valued big data is also needed.

1.2 Research aims

In this section, we will summarise the research questions studied in this thesis. In particular, we consider these questions in the context of discrete-valued time series data of possibly high and ultra-high dimensions. To represent the dependency structure of time series data, β-mixing condition is adopted throughout this thesis. We will further clarify these assumptions in the following chapters where the potential question may arise.

The research questions are given as follows:

1. How to develop the uniform consistency of local likelihood estimation valid with discrete-valued response time series data?
2. How to overcome the "curse of dimensionality" of data-driven models?
3. How to select the important factors among hundreds and thousands variables, e.g., in high and ultra-high dimensions?
4. How to capture the neighbouring effects, i.e., the spatial impacts, of time series data collected at different locations?

Local linear regression has been developed for data-driven analysis that relaxes the linear assumption (Fan and Yao, 2003). The computational cost is known to be cheap as it is one dimensional. However, when present in the context of time series, the asymptotic properties of it need to be re-confirmed, as the i.i.d assumption, e.g., made in Fan et al. (1998a), is no longer valid. Thus the first question here is to confirm the uniform consistency of local likelihood estimation for time series data.

With the results of the first question, we are now able to overcome the "curse of dimensionality" following the idea of Li et al. (2015). That is, we can treat each one dimensional information extracted as a single "model", and use the popular model averaging method to approximate the true estimates. In particular, we are considering the general form of exponential family for all kinds of discrete-valued time series data.

Once the above question is answered, to expand the practical applicability of the developed model, it is important to extract only the important information from

many variables. Conventional techniques involve penalising the original model in different ways, and therefore force the coefficients of non-correlated variables to be zero. In this thesis, we are concerning the asymptotic properties of such penalty methods combined with our model.

The last question further extends the area of application to the spatio-temporal domain. This is of particular importance to deal with emerging events such as COVID-19. Such global events provide us rich data across different regions. It is thus in a timely need to study the corresponding neighbouring effects along with the time series dependency for predictions.

1.3 Outline

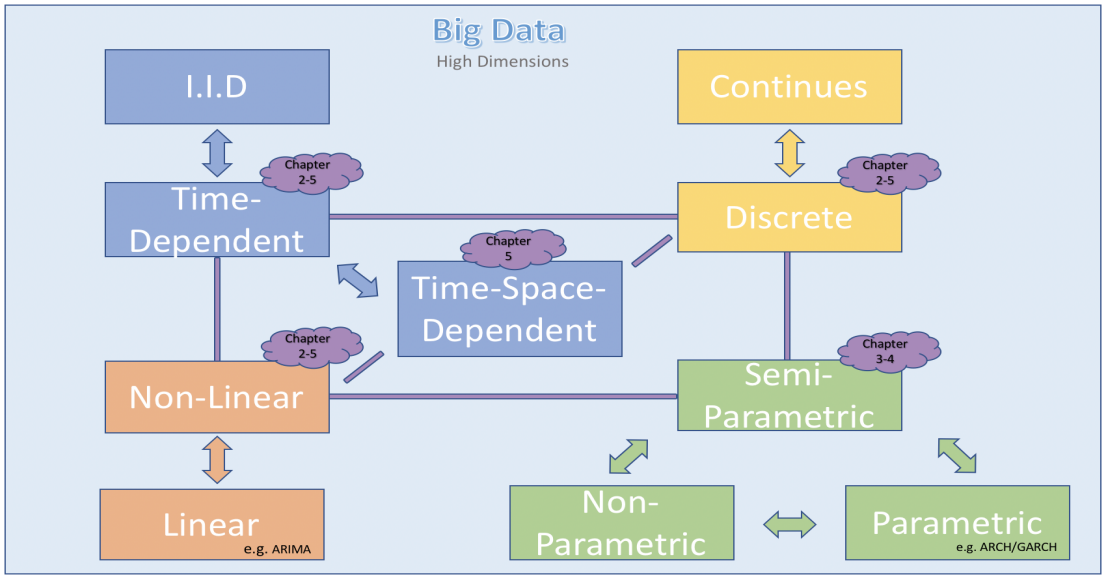

Figure 1: Outline of the Thesis

The structure of this thesis will be structured as follows:

Chapter 2 will present a non-parametric local likelihood estimation, namely the local linear maximum likelihood (LLML), based on β-mixing time series data. We then provide the proof of its asymptotic consistency. A Monte-Carlo simulation and the applications to COVID-19 and FTSE100 Index data are given to show the strength of our approach. A bandwidth selection criterion is also discussed and applied to improve the performance of this approach.

Based on the work in Chapter 2, we then propose a semi-parametric model in Chapter 3, namely the Model Averaging nonlinear Marginal Logistic Regressions (MAMaLoR), following the ideas of Li et al. (2015). This is a special case where the popular binary classification problem is considered in the context of time series data. Asymptotic results are established for our proposed procedure. Numerical results of simulation and application to FTSE100 index data further confirm the ability of our proposed model compared to conventional methods, when dealing with large number of dimensions. This chapter has been submitted to peer review journal for publication.

Next, in Chapter 4, we generalise the semi-parametric model proposed in Chapter 3 to the exponential family. In addition, the variable selection technique adapted LASSO (Zou, 2006) is applied to our generalised model. We show the asymptotic properties under mild conditions, where the computational algorithm is designed for the purpose of application. Numerical performances to the application of FTSE100 index data are compared to both conventional method and popular machine learning technique.

Chapter 5 is an application to the COVID-19 data in the UK. A simple yet powerful spatio-temporal analysis is present to deal with the emerging needs of infection number prediction. That is, both the temporal lags and the neighbouring effects are considered in the model. In particular, we are also concerning exogenous factors such as people's self-awareness of protection and the power of lockdown regards reducing the infect rate. The numerical example indicates the importance of capturing these spatial effects when dealing with such type of data.

Conclusions, challenges and future works are summarised at the end of the thesis in Chapter 6.

Chapter 2

Uniform Consistency for Local Maximum Likelihood Estimation of Time Series Non-Parametric Regression by Local Linear Estimating Equations

Local linear kernel fitting is a popular non-parametric technique for modelling nonlinear time series data. Investigations into it, although extensively made for continuous-valued case, are still rare for the time series that are discrete-valued. In this chapter, we propose and develop the uniform consistency of local linear maximum likelihood (LLML) fitting for time series regression allowing response to be discrete-valued under β-mixing dependence condition. Specifically, the uniform consistency of LLML estimators is established under time series conditional exponential family distributions with aid of a beta-mixing empirical process through local estimating equations. Performances of the proposed method are demonstrated by a Monte-Carlo simulation study and the applications to COVID-19 data and financial time series data FTSE100 . There is a huge potential for the developed theory contributing to further development of discrete-valued semiparametric time series models.

2.1 Introduction

The research of local linear regression is of wide interest in statistical and econometric nonlinear and non-parametric modelling (c.f., Fan and Gijbels (1996), Fan and Yao (2003), Li and Racine (2007), Lu and Linton (2007)). This is because in practice people often have no prior knowledge about the relationship between variables, and especially in the age of big data. Thus, non-parametric models, and especially semiparametric models that combine non-parametric and parametric methods, are particularly of interest to deal with such situation of nonlinear time series analysis; see e.g., Gao (2007) and Terasvirta et al. (2010).

Though in literature continuous-valued response is often assumed, discrete outcomes are common in practice, e.g., in finance, insurance, biology and etc. Specifically, we are interested in the discrete-valued time series datasets, which, in particular, can be expressed in the form of conditional exponential family distributions. For example, the Poisson distribution is widely applied in applications such as in queuing theory, e.g., to express the number of people joining the queue, and in particular in modelling COVID-19 time series data such as the series of daily increase number of virus infected cases. Binomial distribution (or Categorical distribution in a more general sense), e.g., of financial time series data, is another example that plays an important role in areas of classification such as disease diagnosing, default rate checking, and etc. Within the discrete-valued time series models, parametric linear or nonlinear autoregression technique is very popular. The reader is referred to Davis et al. (1999), Fokianos et al. (2009) and Davis et al. (2016) for a comprehensive review on the related developments.

Differently from those parametric models which suffer from model misspecification, in this chapter we propose analysing time series regression in a non-parametric manner for discrete-valued response under a conditional exponential family. In this sense, maximum likelihood method is preferred over ordinary mean least square method. The idea of adopting maximum likelihood method in local fitting can be traced back to Tibshirani and Hastie (1987), where they have applied it to the generalised linear models and proportional hazards models for independent data. Later Fan et al. (1998a) have discussed the good properties of it in local polynomial fitting. Related research also includes Carroll et al. (1997), among others, where they have done a series of research work on local estimation.

However, when applied to time series, the independence assumption often assumed in literature is violated with temporal dependency, characterising of which is also known in terms of "mixing". Mixing conditions, as briefly discussed in Wong et al. (2020), are established in literature as a way to extending results from i.i.d cases to dependent structure (c.f., Bradley (2005), Lu (2001) and Lu and Linton (2007)). In particular, β-mixing, which is often discussed in machine learning (Mcdonald et al., 2011), defines the β coefficient at lag n to be the l_1 distance from independence in probability (c.f., Definition 2.1 in Section 2.2). The β-mixing property also implies the α-mixing condition as it is stronger and with a faster decay rate. For a more detailed discussion of β-mixing conditions, the reader is referred to Doukhan et al. (1995)[Section2.4].

Our focus in this chapter is thus to establish the asymptotic properties of the local linear maximum likelihood (LLML) fitting for time series non-parametric regression allowing for discrete-valued response under β-mixing condition. As is well known, the uniform consistency results of such non-parametric kernel-based estimators are widely useful in further developments such as semiparametric modelling (c.f., Nielsen (2005), Hansen (2008) and Kristensen (2009)). Investigations into the method, although extensively made for continuous-valued time series (c.f., Liebscher (1996) , Masry (1996), Bosq (2012), Fan and Yao (2003), Hansen (2008) and Kristensen (2009), Li et al. (2012), and the references therein), are still rare for the time series that are discrete-valued. In this chapter, we develop the uniform consistency of local linear maximum likelihood (LLML) fitting under β-mixing dependence condition. Specifically, the uniform consistency of LLML estimators under time series conditional exponential family distributions is established. Differently from the local least squares based estimation with available analytical solution in the literature (c.f., Li et al. 2012), study of the LLML estimator becomes much harder as it lacks an analytical solution, which need more efforts by a β-mixing empirical process theory to cope with (c.f., Lu et al. (2007)) in this chapter. Performances of the proposed method are demonstrated by a Monte-Carlo simulation study and the applications to COVID-19 and FTSE100 Index data . There is a huge potential for the developed theory contributing to further development of discrete-valued semiparametric time series models.

The rest of this chapter is structured as follows. We will introduce the local linear estimating model in Section 2, followed by the establishment of its uniform consistency discussed in Section 3. In Section 4, the numerical examples including

a Monte-Carlo simulation and the applications to COVID-19 and FTSE100 index data will be demonstrated before the conclusion in Section 5.

2.2 Time Series Local Linear Model

We consider a general regression model with (Y_t, X_t) being the β-mixing time series process, where Y_t allows to be discrete valued, and X_t denotes the d-dimensional covariate series. Formally, the β-mixing property can be explicitly expressed to measure dependence as follows:

Definition 2.1. *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 (also-called σ-algebra) of $\{Z_s, s \geq t+n\}$.

Here the strict stationary time series means that the joint probability of Z_t do not change in time. For further details, the readers are referred to Hamilton (2020).

Assume that Y_t has a conditional distribution in the exponential family given the past information up to time $t-1$ expressed in X_t. Then the generic form of density function of the conditional exponential family can be expressed as:

$$m_Y(y;\theta_t) = \Theta(y)\exp(y\theta_t - \phi(\theta_t)), \tag{2.1}$$

where $\Theta(.)$ and $\phi(.)$ are known functions for a particular distribution family, and θ_t is the canonical parameter depending on the given information in X_t, which can also be expressed by a link function $\eta(\mu_t)$. Here μ_t is the conditional mean $\mu_t = E(Y_t|X_t)$ that is to be estimated, which connects the covariate vector X_t, satisfying $\mu_t = E(Y_t|X_t) = \phi'(\theta_t)$, where $\phi'(\cdot)$ stands for the derivative of $\phi(\cdot)$. So ${\phi'}^{-1}(\cdot)$ is a canonic link function, which is known for a specific distribution, where ${\phi'}^{-1}$ stands for the inverse function of ϕ'. We will hence consider a known link function $\eta = {\phi'}^{-1}$ by which we express the regression as follows:

$$\eta(\mu_t) = \theta_t = f(X_t), \tag{2.2}$$

with $f(\cdot)$ the unknown function that we need to estimate. Therefore this problem of non-parametric estimation is essentially semi-parametric in the sense that non-parametric function f and conditional exponential family for Y_t given the information expressed in X_t apply.

Then given the observations $\{(Y_t, X_t), t = 1, 2, \cdots, n\}$ of the size n, the local log conditional likelihood for the Y_t's (given initial information) is thus given by

$$\ell_{h,x}(\boldsymbol{\mu}; Y) = \sum_{t=1}^{n} \log m_{Y_t}(Y_t, \boldsymbol{\theta}_t) K_h(X_t - x), \tag{2.3}$$

where $K_h(\cdot) = h^{-d} K(\cdot/h)$ with $K(\cdot)$ a kernel function on $\mathbb{R}^d$, and $h > 0$ is a bandwidth satisfying $h = h_n \to 0$ as $n \to \infty$. Note that we denote the dimensions considered d, where $d = 1$ in our special case of one-dimensional local linear regression.

Since the relationship between Y_t and X_t is often unknown, non-parametric smoothers can be used to estimate the conditional mean by estimating equations obtained by setting the partial differentiations of (2.3) being zero,

$$\frac{1}{n} \sum_{t=1}^{n} \Lambda(Y_t, \theta_t) K_h(X_t - x) = 0, \tag{2.4}$$

where $\Lambda()$ is an appropriately defined function denoting the distance between Y_t and θ_t. For instance, if Y_t is binary-valued, then $\Lambda(Y_t, \theta_t) = Y_t - \phi'(\theta_t)$. The model in population can then be expressed as:

$$E[\Lambda(Y_t, \theta_t)|X_t] = 0. \tag{2.5}$$

Suppose $f(x)$ has $(p+1)$-th continuous derivative at any given point x. Then for data points X_t in the neighbourhood of x, we can approximate $f(X_t)$ via Taylor expansion by polynomial of degree p:

$$\begin{aligned} f(X_t) &\approx f(x) + f'(x)(X_t - x) + ... + \frac{f^{(p)}(x)}{p!}(X_t - x)^p \\ &\equiv \mathbf{x}_t^T \boldsymbol{\beta}, \quad |X_t - x| \leq h, \end{aligned} \tag{2.6}$$

where $\mathbf{x}_t = (1, (X_t - x), ..., (X_t - x)^p)^T$, with the superscript T demoting transpose, and $\boldsymbol{\beta} = (\beta_0, ..., \beta_p)$ with $\beta_j = f^{(j)}(x)/j!$, and $f^{(j)}(x)$ is the j-th order derivative of $f(x)$ w.r.t. x.

In a general sense, the larger degree of polynomial would give a smoother estimator but at the cost of stronger assumptions with more local parameters to estimate. In this regard, local linear fitting is usually preferred, i.e., $p = 1$. (c.f., Fan et al. (1998a)).

Thus under the first order partial derivative,

$$\begin{aligned} f(X_t) &\approx f(x) + f'(x)^T(X_t - x) \\ &\equiv \beta_1 + \beta_2^T(X_t - x), \quad if\ |(X_t - x)| \leq h, \end{aligned} \tag{2.7}$$

where $f'(x)$ is the derivative of $f(x)$ w.r.t. x, and $\boldsymbol{\beta} = (\beta_1, \beta_2^T)^T \in \mathbb{R}^{1+d}$ is a vector of local coefficients at x.

By solving the local maximum likelihood estimation above (see Fan et al. (1998a)), which is easy as it could be seen as a locally weighted linear regression, we then get the estimation at x as the intercept $\hat{f}(x)$ in the equation (2.7). Since x is chosen arbitrary, we now let x go through each point in X_t and hence get the estimated conditional mean $\hat{\mu}_t = \eta^{-1}(\hat{f}(X_t))$ with $\eta^{-1}()$ standing for the inverse function of the link function $\eta(\cdot)$.

2.3 Uniform consistency

In this section, we will derive the uniform consistency of the local fitting estimator $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2^T)^T = (\hat{f}(x), (\hat{f}'(x))^T)^T$ to $\boldsymbol{\beta}_0 = (f(x), (f'(x))^T)^T$ with respect to $x \in A$, a closed subset of R^d. It is based on general local estimating equations.

Our estimator $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_n$ is defined as the solution to :

$$\Omega_n(\boldsymbol{\beta}, x, h) = \begin{pmatrix} \Omega_n^{(1)}(\boldsymbol{\beta}, x, h) \\ \Omega_n^{(2)}(\boldsymbol{\beta}, x, h) \end{pmatrix} = 0, \tag{2.8}$$

where

$$\Omega_n^{(1)}(\boldsymbol{\beta}, x, h) = \frac{1}{n}\sum_{t=1}^{n} \Lambda(Y_t; \beta_1 + \beta_2^T(X_t - x))K_h(X_t - x), \tag{2.9}$$
$$\Omega_n^{(2)}(\boldsymbol{\beta}, x, h) = \frac{1}{n}\sum_{t=1}^{n} \Lambda(Y_t; \beta_1 + \beta_2^T(X_t - x)) \cdot [\frac{X_t - x}{h}]K_h(X_t - x),$$

with $\boldsymbol{\beta} = (\beta_1, \beta_2)$ and the bandwidth h.

For greater generality, we allow $\hat{f}(x)$ to be an approximate solution to the equation so that $\Omega_n(\hat{\boldsymbol{\beta}}_n, x, h)$ goes to zero in probability at a rate to be specified later. For independent and identically distributed (i.i.d.) data, the convergence of the estimators was established by Nielsen (2005). However, for our concerned β-mixing time series, we give the theorems with proofs shown in this Section 2.3.

Before jumping into the results, we need the following

2.3.1 Assumptions

A1 (i) The process (Y_t, X_t), with Y_t of a conditional distribution in the exponential family given X_t, is strictly stationary β-mixing with the mixing coefficient $\beta(t) = O(t^{-b})$ for some $b > max(2(\rho r + 1)/(\rho r - 2), (r + a)/(1 - 2/\rho))$ with $a \geq (\rho r - 2)r/(2 + \rho r - 4r)$; (ii) the joint probability density function $g_{X_{t-1},\cdots,X_{t-s}}(x_1, \cdots, x_s)$ is bounded uniformly for any $t_0 < t_1 < \cdots < t_s$ and $0 \leq s \leq 2(r - 1)$; (iii) $E|\Lambda(Y_t, f(X_t))|^{\rho r} < \infty$, $E|X_t|^{\rho r} < \infty$ for some real number $\rho > 4 - 2/r$, where $r \geq 1$ is some positive integer.

A2 The kernel $K(\cdot)$ is a bounded and symmetric density function on $\mathbf{R}$ with bounded support S_K. Furthermore, $|K(z) - K(x)| \leq C|z - x|$ for $z, x \in S_K$ and some $0 < C < \infty$.

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 $n \to \infty$;

A4 For any concerned function f in 2.2 , we define its Lipschitz norm: For some $\varpi > 0$, let $[\varpi]$ be the largest integer not greater than ϖ, and define (if it

exists)

$$\|f\|_{\infty,\varpi} = \max_{0\leq\kappa\leq[\varpi]} \sup_{x\in A}\|f^{(\kappa)}(x)\| + \sup_{x\neq x';x,x'\in A} \frac{\|f^{([\varpi])}(x) - f^{([\varpi])}(x')\|}{\|x - x'\|^{\varpi-[\varpi]}}, \quad (2.10)$$

where $f^{(\kappa)}(x)$ stands for the κ-th derivative of $f(x)$ with respect to x. We suppose f belongs to a functional space $\mathbf{F}$ with some $\varpi \geq 2$ and $c > 0$:

$$\mathbf{F} := \{f : f \textit{ is a continuous function from } A \textit{ to } R \textit{ with } \|f\|_{\infty,\varpi} \leq c\}, \quad (2.11)$$

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 $C_c^{\varpi}(A)$.

A5 Assume that $E[\Lambda(Y_t, z)^2] < \infty$ for all $z \in R$. Let

$$\Phi(x, z) = E[\Lambda(Y_t, z)|X_t = x]. \quad (2.12)$$

- (i) $(x, z) \to \Phi(x, z) \cdot g(x)$ is three times continuously differentiable as a function form $\mathbb{R}^2$ to $\mathbb{R}$, where $g(x)$ is the marginal density of x, which is strictly positive and continuous over A. We denote the derivative of Φ with respect to x by $\dot{\Phi}_x$, and the derivative with respect z by $\dot{\Phi}_z$, etc.
- (ii) For any fixed Y_t the function $z \to \Lambda(Y_t, z)$ is Lipschitz on a compact set. For any compact $C \subset R$ there is a function $\Omega^*(Y_t)$ (depending on C) such that

$$|\Lambda(Y_t, z) - \Lambda(Y_t, \tilde{z})| \leq \Omega^*(Y_t) \cdot |z - \tilde{z}| \quad \textit{for all} \quad z, \tilde{z} \in C, \quad (2.13)$$

  where $E[(\Omega^*(Y_t))^{2r}(1 + |X_t|^{2r})] < \infty$ with r given in assumption A1.

Remark 2.2. *Assumption 1 shows a standard β-mixing process which is satisfied by many linear and nonlinear time series models (Fan and Yao, 2003; Lu et al., 2007). The kernel is guaranteed to be bounded by Assumption 2, which is commonly seen in this type of problem (Hardle et al., 1993; Xia and Li, 1999). Assumption 3 is also standard in time series topics (Fan et al., 2003; Lu et al., 2007) 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).*

The Lipschitz norm conditions (Assumption 4 and 5) are introduced to give a tighter bound than uniform norm (Nielsen, 2005). Note that we are concerned with function $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$ *and* $\boldsymbol{\beta}_0$ *as a function of* x *in* $\mathbf{F}$. *Under A4 the Lipschitz continuous norm, it is stronger than the uniform norm for a function in* $\mathbf{F}$, *i.e. the Lipschitz norm of* ϖ*:*

$$\|\boldsymbol{\beta}\|_{\mathbf{F}} = \|\boldsymbol{\beta}\|_{\infty} = \max_{i=1,2} sup_{x\in A}|\beta_i(x)| \leq \|\boldsymbol{\beta}\|_{\infty,\varpi}, \tag{2.14}$$

Thus, consistency in Lipschitz norm implies uniform consistency. Assumption A5 was introduced for a general case of local estimating equations (c.f., Nielsen (2005)). Recalling that $f(X_t) = \eta(\mu_t)$ *under canonical link function* η *and (2.9), we have* $\Lambda(Y_t, z) = Y_t - \phi'(z)$, $\Phi(X_t, z) = E[\Lambda(Y_t, z)|X_t = x] = E(Y_t|X_t = x) - \phi'(z) = \phi'(f(X_t)) - \phi'(z)$. *Clearly* $\Phi(x, f(X_t)) = 0$. *Here assumption A5 holds automatically under assumption A1.*

2.3.2 Theorems

We first need to study the properties of $\Omega_n^{(1)}$ and $\Omega_n^{(2)}$ in expectation.

Theorem 2.3. *Suppose the assumptions A1-A4 with model 2.2 are satisfied. Then*

$$E[\Omega_n(\boldsymbol{\beta}, x, h)|X_t] = (1 + o(1))\text{diag}(1, h^d)\Omega_0(\boldsymbol{\beta}, x),$$

where $o(1)$ *is uniformly with respect to* $x \in A$ *and* $\boldsymbol{\beta} \in \mathbf{F}$, *and* $\Omega_0(\boldsymbol{\beta}, x) = (\Omega_0^{(1)}(\boldsymbol{\beta}, x), (\Omega_0^{(d)}(\boldsymbol{\beta}, x))^T)^T$, *with* $\Omega_0^{(1)}(\boldsymbol{\beta}, x) = \Phi(x, \beta_1)g(x)$ *and*

$$\Omega_0^{(2)}(\boldsymbol{\beta}, x) = (\beta_2\dot{\Phi}_z(x, \beta_1) + \dot{\Phi}_x(x, \beta_1))g(x) + \Phi(x, \beta_1)g'(x).$$

The true value of the local parameter $\boldsymbol{\beta}_0 = (f(x), (\mathbf{f}'(x))^T)^T$ *is the solution to*

$$E[\Omega_n(\boldsymbol{\beta}, x, h)|X_t] = 0.$$

Further, $E[\Omega_n(\boldsymbol{\beta}, x, h)|X_t] = 0$ *has the unique solution at* $\boldsymbol{\beta}_0$.

Proof. We only outline the proof as it is similar to the derivation in Section 2 of Nielsen (2005).

First, we note that the solution of $\Omega_n(f(x), x, h) = 0$ is also the solution to

$$M_n(\boldsymbol{\beta}, h) = \sup_{x\in A}|\Omega_n(\boldsymbol{\beta}, x, h)| = 0 \tag{2.15}$$

Now consider the solution point $\boldsymbol{\beta}_0$ of $M_n(\boldsymbol{\beta}, h) = 0$ over Lipschitz continuous function $\boldsymbol{\beta}(x)$ (define on A) with $\|\boldsymbol{\beta}\|_{\infty,\phi} \leq c$ and $c > 0$. Note that by differentiability of the $\boldsymbol{\beta}_0 = \boldsymbol{\beta}_0(x) = (f(x), (f'(x))^T)^T$ and the boundedness of A, such a c exits.

Intuitively, if $\Omega_n(\boldsymbol{\beta}, x, h)$ is uniformly close to $E[\Omega_n(\boldsymbol{\beta}, x, h)]$. Then $\hat{\boldsymbol{\beta}}$ should be close to the solution of $E[\Omega_n(\boldsymbol{\beta}, x, h)] = 0$, and is a consistent estimator of $\boldsymbol{\beta}_0$. We first check $\boldsymbol{\beta}_0$ is the solution to $E[\Omega_n(\boldsymbol{\beta}, x, h)] = 0$ with our local estimating equations for the local exponential family model estimated by local maximum likelihood estimation under model (2.2):

$$\begin{aligned}
E[\Omega_n^{(1)}(\boldsymbol{\beta}, x, h)] &= E[\frac{1}{n}\sum_{t=1}^{n}[Y_t - \phi'(\beta_1 + \beta_2(X_t - x))]h^{-d}K_h(X_t - x)] \\
&= E[E[\frac{1}{n}\sum_{t=1}^{n}[Y_t - \phi'(\beta_1 + \beta_2(X_t - x))]]h^{-d}K_h(X_t - x)|X_t] \\
&= E[\frac{1}{n}\sum_{t=1}^{n}[E[Y_t|X_t] - \phi'(\beta_1 + \beta_2(X_t - x))]h^{-d}K_h(X_t - x)] \\
&= E[\frac{1}{n}\sum_{t=1}^{n}[\phi'(f(X_t) - \phi'(\beta_1 + \beta_2(X_t - x))]h^{-d}K_h(X_t - x)],
\end{aligned}$$

where $E[Y_t|X_t] = \phi'(f(X_t))$ follows from (2.2).

Let $\tilde{f}(z_j) = \phi'(z_j)$, and by Taylor expansion together with assumptions A4 and A2 we find:

$$\begin{aligned}
E[\Omega_n^{(1)}(\boldsymbol{\beta}, x, h)] &= E[\frac{1}{n}\sum_{t=1}^{n}[\tilde{f}(f(X_t)) - \tilde{f}(\beta_1 + \beta_2(X_t - x))]h^{-d}K_h(X_t - x)] \\
&= (1 + o(1))[\tilde{f}(f(x)) - \tilde{f}(\beta_1)]g(x) \qquad (2.16)
\end{aligned}$$

where $o(1)$ is uniformly in $x \in A$ owing to assumption A4.

Although we are mainly interested in the generalised local regression model in Section 2.2, where $\Lambda(y_j, z_j) = y - \phi'(z_j)$ as indicated above, but for a general $\Lambda(y_j, z_j)$ under assumption A5, we can still establish (2.16) as in

Nielsen (2005):

$$\begin{aligned} E[\Omega_n^{(1)}(\boldsymbol{\beta}, x, h)] &= E[\Lambda(Y_t; \beta_1 + \beta_2(X_t - x))h^{-d}K_h(X_t - x)] \\ &= E[\Phi(X_t; \beta_1 + \beta_2(X_t - x))h^{-d}K_h(X_t - x)] \\ &= \Phi(x, \beta_1)g(x) + O(h^2), \end{aligned}$$

where the O-term does not depend on x nor on $\|\beta(x)\|_\infty \leq C$, and corresponding to the local exponential family regression in Section 2.2, $\Phi(x, \beta_1) = \tilde{f}(f(x)) - \tilde{f}(\beta_1)$.

Similarly, as in Nielsen (2005),

$$\begin{aligned} E[\Omega_n^{(2)}(\boldsymbol{\beta}, x, h)] &= E[\Lambda(Y_i; \beta_1 + \beta_2(X_t - x))\frac{X_t - x}{h}h^{-d}K_h(X_t - x)] \\ &= h(\beta_2\dot{\Phi}_z(x, \beta_1)) + \dot{\Phi}_x(x, \beta_1))g(x) + h\Phi(x, \beta_1)g'(x) + O(h^3), \end{aligned}$$

where corresponding to the local exponential family regression in Section 2.2, $\dot{\Phi}_x(x, \beta_1) = \tilde{f}'(f(x))f'(x) = f'(x)\phi''(f(x))$ and $\dot{\Phi}_z(x, \beta_1) = -\tilde{f}'(\beta_1) = -\phi''(\beta_1)$, with $\tilde{f}'(z_j) = \phi''(z_j)$ as defined above.

Thus we get:

$$E[\Omega_n^{(1)}(\boldsymbol{\beta}, x, h)] = \Omega_0^{(1)}(\boldsymbol{\beta}, x) + O(h^2) \tag{2.17}$$

and

$$E[\Omega_n^{(2)}(\boldsymbol{\beta}, x, h)] = h\Omega_0^{(2)}(\boldsymbol{\beta}, x) + O(h^3) \tag{2.18}$$

where

$$\begin{aligned} \Omega_0^{(1)}(\beta, x) &= \Phi(x, \beta_1)g(x) \\ \Omega_0^{(2)}(\beta, x) &= (\beta_z\Phi'_z(x, \beta_1) + \dot{\Phi}_x(x, \beta_1))g(x) + \Phi(x, \beta_1)g'(x). \end{aligned}$$

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

$$\begin{cases} \Phi(x, \beta_{01}) = 0 \\ \beta_{02}(x) = -\frac{\dot{\Phi}_x(x, \beta_{01})}{\dot{\Phi}_z(x, \beta_{02})}, \end{cases} \tag{2.19}$$

which is actually unique correspondingly to our local general linear regression in Section 2.2, with $\beta_{01} = f(x)$ and $\beta_{02} = f'(x)$ (in this special case of one-dimensional estimation, i.e., $d = 1$, both β_{01} and β_{02} are scalar; see (2.7)).

The proof of Theorem 2.3 is done. □

We turn to the uniform consistency of $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_n$ in probability. For $\Omega_0^{(i)}(\boldsymbol{\beta}, x)$, $i = 1, 2$, we further know from the above that $\Omega_0^{(i)}(\boldsymbol{\beta}, x)$ is continuous in $\boldsymbol{\beta} \in \mathbf{F}$ (in Lipschitz norm) and $x \in A$ (in Euclidean norm). Therefore, for any $\varepsilon > 0$, there exists $\delta > 0$ such that

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_\infty > \delta \Rightarrow \max_{i=1,2} |\Omega_0^{(i)}(\hat{\boldsymbol{\beta}}, x)| > \varepsilon, \quad for \quad x \in A. \tag{2.20}$$

By (2.20), it suffices to show $\max_{i=1,2} \sup_{x\in A} |\Omega_0^{(i)}(\hat{\boldsymbol{\beta}}, x)| = \max_{i=1,2} \sup_{x\in A} |\Omega_0^{(i)}(\hat{\boldsymbol{\beta}}, x) - \Omega_0^{(i)}(\boldsymbol{\beta}_0, x)| \to 0$ in probability as $n \to \infty$. We have the uniform consistency as follows.

Theorem 2.4. *Suppose the assumptions are satisfied, then* $\mathrm{diag}(1, h^{-d})\Omega_n(\boldsymbol{\beta}, x)$ *converges uniformly in probability to* $\Omega_0(\boldsymbol{\beta}, x)$ *with respect to* $\boldsymbol{\beta} \in \mathbf{F}, x \in A$, *and further if* $nh^2 \to \infty$, *then* $\hat{\boldsymbol{\beta}}_n(x) \to \boldsymbol{\beta}_0(x)$ *uniformly for* $x \in A$.

Proof. We first need the fact that $\tilde{\Omega}_n = \mathrm{diag}(1, h^{-d})\Omega_n$ converges uniformly in probability to Ω_0, the proof for which is standard (c.f., Lemma A.1 of Lu et al. (2007)) udder the given assumptions.

Now we notice that

$$\begin{aligned} \Omega_0^{(i)}(\hat{\boldsymbol{\beta}}, x) - \Omega_0^{(i)}(\boldsymbol{\beta}_0, x) =& (\Omega_0^{(i)}(\hat{\boldsymbol{\beta}}, x) - \tilde{\Omega}_n^{(i)}(\hat{\boldsymbol{\beta}}, x)) \\ & - (\Omega_0^{(i)}(\boldsymbol{\beta}_0, x) - \tilde{\Omega}_n^{(i)}(\boldsymbol{\beta}_0, x)) \\ & + (\tilde{\Omega}_n^{(i)}(\hat{\boldsymbol{\beta}}, x) - \tilde{\Omega}_n^{(i)}(\boldsymbol{\beta}_0, x)). \end{aligned} \tag{2.21}$$

The first two terms on the RHS of (2.21) converge to zero uniformly by the uniform convergence of $\tilde{\Omega}_n$ just mentioned above. So we only need to show the third term on the RHS of (2.21) converges to zero uniformly, which is shown below.

Let $D_n = \mathrm{diag}(1, h)$ and define the empirical process:

$$\begin{aligned} G_n(\boldsymbol{\beta}, x, h) &= \frac{1}{\sqrt{n}} \sum_{t=1}^{n} (\Lambda^*(Z_t, \boldsymbol{\beta}, x, h) - E[\Lambda^*(Z_t, \boldsymbol{\beta}, x, h)]) \\ &= \sqrt{n}h(\Omega_n(\boldsymbol{\beta}, x) - E\Omega_n(\boldsymbol{\beta}, x)) \end{aligned} \tag{2.22}$$

with

$$\Lambda^*(Z_t, \boldsymbol{\beta}, x, h) = \Lambda\Big(Y_t, \beta_1 + \beta_2(X_t - x)\Big) \cdot K_h(X_t - x) \begin{bmatrix} 1 \\ \vdots \\ (\frac{X_t - x}{h})^d \end{bmatrix} \tag{2.23}$$

The two components of $G_n(\boldsymbol{\beta}, x, h)$ are denoted by $G_n^{(i)}(\boldsymbol{\beta}, x, h)$, $i = 1, 2$. Then $\tilde{\Omega}_n(\boldsymbol{\beta}, x) - E\tilde{\Omega}_n(\boldsymbol{\beta}, x) = D_n^{-1}(\Omega_n(\boldsymbol{\beta}, x) - E\Omega_n(\boldsymbol{\beta}, x))$, which is equal to $(n^{1/2}h)^{-1}D_n^{-1}G_n(\boldsymbol{\beta}, x, h)$.

In this proof, we need to determine when $G_n(\boldsymbol{\beta}, x, h)$ converges uniformly in distribution to a bivariate Gaussian process. This can be done as follows in two steps.

Firstly, by the usual Slutsky's skill, it is easy to show the convergence in distribution of $G_n(\boldsymbol{\beta}, x, h)$ to a Gaussian distribution at any finite number of pairs of $(\boldsymbol{\beta}, x)$'s. Secondly, to show the weak convergence in process, we will need to show the stochastic equicontinuity of $\{G_n^{(i)}(\boldsymbol{\beta}, x, h) : \boldsymbol{\beta} \in \mathbf{F}, x \in A\}$, that is, for every $\epsilon > 0$ and $\varphi > 0$, there is a $\delta > 0$ such that:

$$\limsup_{n\to\infty} P\Big(\sup_{\boldsymbol{\beta}\in\mathbf{F}, x\in A} \sup_{(\boldsymbol{\beta}', x')\in B((\boldsymbol{\beta}, x), \delta)} |G_n^{(i)}(\boldsymbol{\beta}'(\cdot), x', h) - G_n^{(i)}(\boldsymbol{\beta}(\cdot), x, h)| > \epsilon\Big) < \varphi. \tag{2.24}$$

Here $B(\vartheta, \delta)$ represents a ball in the parameter space, centred at $\vartheta = (\boldsymbol{\beta}, x)$ and whose radius depends on δ. For this we need a lemma owing to Doukhan et al. (1995).

Lemma 2.5. *To prove the stochastic equicontinuity of the empirical process we need to check the following conditions (Doukhan et. al, 1995, page 405)*

(a) $\{Z_t = (Y_t, X_t) : t \geq 1\}$ *is a stationary absolutely regular sequence with mixing coefficient* $\beta(s) = O(s^{-b})$ *for some* $b > r/(r-1)$, *and* $r > 1$.

(b) $E_p[\{\widetilde{\Omega}^*\}^{2r}(Z_t)] < \infty$, *where* $r > 1$ *in (a), and* $\widetilde{\Omega}^*(Z_t)$ *is the envelope of* $\mathcal{M} = \{\Lambda^*(\cdot, \beta, x, h) : \beta \in \mathbf{F}, x \in A\}$, *that is* $|\Lambda^*(\cdot, \beta \in \mathbf{F}, x \in A, h)| \leq \widetilde{\Omega}^*(\cdot)$ *for any* $\beta \in \mathbf{F}, x \in A$

(c) $\forall \varepsilon > 0$, $\log N_2(\epsilon, \mathcal{M}) = O(\varepsilon^{-2\eta})$ *for some* $\varphi > 0$, *with* $b(1-\varphi) > r/(r-1)$ *for* b *and* r *as in (a), where* $N_2(\varepsilon, \mathcal{M})$ *is the* L_2*-bracketing cover number of* $\mathcal{M}$ *in (b)*

Now the following proof is to check if the conditions above are met.

(a) holds by the Assumption A1.

(b) can be validated as we have $\|\boldsymbol{\beta}\| \leq \|\boldsymbol{\beta}_0\|+1 = C$. For $\boldsymbol{\beta}, \widetilde{\boldsymbol{\beta}} \in \mathbf{F}$ with Lipschitz norm $\|\boldsymbol{\beta}\|, \|\widetilde{\boldsymbol{\beta}}\| \leq C$ and $x, \widetilde{x} \in A$ we have

$$\begin{aligned}
|\beta_i(x) - \widetilde{\beta}_i(\widetilde{x})| &\leq |\beta_i(x) - \widetilde{\beta}_i(x)| + |\widetilde{\beta}_i(x) - \widetilde{\beta}_i(\widetilde{x})| \\
&\leq \sup_{x \in A} |\beta_i(x) - \widetilde{\beta}_i(x)| + |x - \widetilde{x}| \cdot \sup_{x, x' \in A} \frac{|\widetilde{\beta}_i(x) - \widetilde{\beta}(x')|}{|x - x'|} \\
&\leq \|\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\| + 2C|x - \widetilde{x}|.
\end{aligned}$$

Similarly, by Lipschitz norm

$$\begin{aligned}
|\beta_2(x)x - \widetilde{\beta}_2(\widetilde{x})\widetilde{x}| &\leq |\beta_2(x)| \cdot |x - \widetilde{x}| + |\widetilde{x}| \cdot |\beta_2(x) - \widetilde{\beta}_2| \\
&\leq C|x - \widetilde{x}| + sup_{x \in A}|x| \cdot (\|\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\| + 2C|x - \widetilde{x}|)
\end{aligned}$$

and

$$\{(y, z) \longrightarrow \Lambda(y; \beta_1(x) + \beta_2(x)(z - x))K(\frac{z - x}{h}) : x \in A, h > 0, \beta \in \mathbf{F}, \|\boldsymbol{\beta}\|_{\mathbf{F}} < C\}$$

with the envelope (c.f. Nielsen (2005), page 497)

$$(|\Lambda(y, 0)| + (C_1 + C_2|z|)\Omega^*(y)) \cdot \sup_{u \in S_K} K(u),$$

where $\Omega^*(y)$ is defined in assumption A5(ii), and S_K is the support of $K(\cdot)$. Similarly,

$$\{(y, z) \longrightarrow \Lambda(y; \beta_1(x) + \beta_2(x)(z - x))\frac{z - x}{h}K(\frac{z - x}{h}) : x \in A, h > 0, \beta \in \mathbf{F}, \|\boldsymbol{\beta}\|_{\mathbf{F}} < C\}$$

with the envelope

$$(|\Lambda(y, 0)| + (1 + sup_{x \in A}(|x| + |z|)\Omega^*(y)|) \cdot \sup_{u \in S_K} |u|K(u)$$

Hence (b) holds by conditions A1 and A2.(van der Vaart and Wellner 1996, Section 2.7.4).

(c) The proof can be as done in Lu et al. (2007) (page S26), so we only have a simple idea given here. As $\mathbf{F} = C_c^{\phi}(A)$ with $\phi \geq 2$, for $\forall \varepsilon > 0$, we can cover $\mathbf{F}$ by finite number, say N_1, of balls of radius ε with centres $\beta_i, i = 1, .., N_1$,

in $\mathbf{F}$, say, $\mathbf{F}_i, i = 1, ..., N_1$, such that: $\forall \beta \in \mathbf{F}, \exists \beta_i$, such that

$$\|\beta - \beta_i\| \leq \frac{\varepsilon}{2C}.$$

By van der Vaart and Wellner (1996, Theorem 2.7.1), it is known that $N_1 = N(\varepsilon, \mathbf{F}, \|\cdot\|_\infty)$ satisfies $\log N(\varepsilon, \mathbf{F}, \|\cdot\|_\infty) \leq C\varepsilon^{-1/2}$. Similarly, A is a closed subset in R, for $\forall \varepsilon > 0$, we can cover A by finite number, $N_2 = C\varepsilon^{-1}$, of balls of radius ε with centres $x_i, i = 1, .., N_2$, in A, say, $A_j, j = 1, ..., N_2$, such that: $\forall x \in A, \exists x_i$, such that

$$\|x - x_j\| \leq \frac{\varepsilon}{2C}.$$

As $\Lambda^*(Z_t, \boldsymbol{\beta}, x, h) =$

$$\Lambda\Big(Y_t, \beta_1 + \beta_2(X_t - x)\Big) \cdot K_h(X_t - x) \begin{bmatrix} 1 \\ \vdots \\ (\frac{X_t - x}{h})^d \end{bmatrix}$$

is a continuous function of $(\boldsymbol{\beta}, x)$, $\Lambda^*(\cdot, \boldsymbol{\beta}, x, h)$ can be approximated by, say, $\Lambda^*(\cdot, \boldsymbol{\beta}_{i^*}, x_{j^*}, h)$ for some i^* and j^* for any $\boldsymbol{\beta} \in \mathbf{F}, x \in A$. Therefore we can cover $\mathcal{M} = \{\Lambda^*(\cdot, \boldsymbol{\beta}, x, h) : \beta \in \mathbf{F}, x \in A\}$ by $N_2(\varepsilon, M) \leq N_1 \times N_2$ suitably defined balls as specified in (c) (van der Vaart and Wellner,1996, Theorem 2.7.1). The details are omitted here (c.f., Lu et al. (2007)).

Thus $\{G_n(\boldsymbol{\beta}, x, h) : \boldsymbol{\beta} \in \mathbf{F}, x \in A\}$ converges in distribution in process. Hence,

$$\sup_{\|\boldsymbol{\beta}(x)\| \leq C, x \in A} |G_n^{(i)}(\beta, x, h)| = Op(1), i = 1, 2 \tag{2.25}$$

By Equations (2.17) and (2.18) we have

$$\begin{aligned}
&\sup_{\|\boldsymbol{\beta}(x)\| \leq C, x \in A} |\Omega_n^{(1)}(\boldsymbol{\beta}(x), x, h) - \Omega_0^{(1)}(\boldsymbol{\beta}(x), x)| \\
&\leq \frac{1}{\sqrt{n}h} \sup_{\|\boldsymbol{\beta}(x)\| \leq C, x \in A, h > 0} |G_n^{(i)}(\boldsymbol{\beta}(x), x, h)| \\
&+ \sup_{\|\boldsymbol{\beta}(x)\| \leq C, x \in A, h > 0} |E[\Omega_n^{(1)}(\boldsymbol{\beta}(x), x, h) - \Omega_0^{(1)}(\boldsymbol{\beta}(x), x)]| \\
&= Op(1/(\sqrt{n}h)) + O(h^2) \xrightarrow{\mathrm{P}} 0;
\end{aligned}$$

and

$$\sup_{\|\boldsymbol{\beta}(x)\|\leq C, x\in A} |\frac{1}{h}\Omega_n^{(2)}(\boldsymbol{\beta}(x), x, h) - \Omega_0^{(2)}(\boldsymbol{\beta}(x), x)|$$
$$= Op(\frac{1}{\sqrt{n}h}) + O(h^2) \xrightarrow{\mathrm{P}} 0.$$

Thus by (2.20) and (2.21) with $nh^2 \to \infty$, $\|\hat{\boldsymbol{\beta}}(.) - \boldsymbol{\beta}_0(.)\|_{\mathbf{F}} = o_p(1)$ is proved.

□

Based on Theorem 2.4, we can simply have $\hat{f}(x)$ is uniformly consistent to $f(x)$ over $x \in A$, a closed subset of R^d. This is a very useful theoretical result. For example, in practice, we are interested in $\mu_t = E(Y_t|X_t) = \eta^{-1}(f(X_t))$ (following from (2.2), with η^{-1} the inverse function of a known link η) for prediction of Y_t, which can therefore be estimated by $\hat{\mu}_t = \hat{E}(Y_t|X_t) = \eta^{-1}(\hat{f}(X_t))$. We can thus have the consistency as follows. Under the assumptions of Theorem 2.4 with a continuous link function η, we have

$$\sup_{X_t\in A} |\hat{\mu}_t - \mu_t| \to 0$$

in probability as $n \to \infty$. In practice, we can take the close set $A \subset R^d$ very large so that the observed values of X_t belong to it. This guarantees that our predicted value $\hat{\mu}_t$, i.e., $\hat{Y}_t$, is uniformly consistent to the theoretically optimal predictor μ_t as the training sample size n tends to infinity.

2.4 Numerical examples

In this section, a Monte-Carlo simulation is first present to show the advantage of this method. The response Y_t generated is assumed to follow a binomial distribution given X_t. This is the case that our proposed method works as a binary classification, which can be applied to a wide range of applications in practice. Then we give an application to the COVID-19 data of which the daily confirmed number of new cases are estimated and predicted. A poisson distribution is assumed for the response variable, which is commonly adopted in epidemiology studies. Another application to FTSE100 Index data, where the market direction is assumed to follow the binomial

distribution, is presented at the end of this section. We hope to demonstrate that the proposed local linear method is robust for the exponential family.

2.4.1 Simulation

Let the mixing time series data of size n be generated by

$$\begin{aligned} X_t &= \cos(2X_{t-1}) + \epsilon_t \\ Y_t &= I(X_t > 0), \end{aligned} \tag{2.26}$$

where $\epsilon_t \sim i.i.dN(0, \sigma^2)$. For the sake of simplicity, here we choose $\sigma^2 = 1$. According to the assumption, Y_t given X_t follows a binomial distribution with probability p_t. Hence we have

$$[Y_t|X_t] \sim Bin(1, p_t),$$

where $p_t = p(X_t)$ is defined as:

$$\begin{aligned} p(x) &= P(Y_t = 1|X_{t-1} = x) \\ &= P(cos(2X_{t-1}) + \epsilon_t > 0|X_{t-1} = x) \\ &= P(\frac{\epsilon_t}{\sigma} > \frac{-cos(2x)}{\sigma}) \\ &= 1 - \mathbf{\Phi}(-\frac{cos(2x)}{\sigma}) = \mathbf{\Phi}(\frac{cos(2x)}{\sigma}), \end{aligned} \tag{2.27}$$

where $\mathbf{\Phi}$ is the cumulative distribution function (CDF) of the standard normal distribution The corresponding log odds can be obtained from:

$$f(X_t) = log\frac{p_t}{1 - p_t}, \tag{2.28}$$

Now we can re-write the log likelihood function as:

$$\log L = \sum_{i=1}^{n}[log(p_i) \cdot Y_i + log(1 - p_i)(1 - Y_i)]K(\frac{X_i - x}{h}), \tag{2.29}$$

where $K(\frac{X_i - x}{h})$ is the Epanechnikov kernel with standard formulation

$$K(u) = \frac{3}{4}(1 - u^2)I_{[-1,1]}(u), \tag{2.30}$$

where it is commonly adopted in literature; see Fan et al. (1998a), and the range $[-1, 1]$ is used here to generate a sequence of points within it to estimate.

It is known that, the bandwidth h selected for kernel would have large impact on its performance (Fan et al., 1998a). Different criterion would also lead to different optimal h. In this chapter, we are going to use cross validation based on log likelihood to select the best h within given data sample. Note that the $\log L$

$$logL = \sum_{i=1}^{n} [log(\frac{1}{1+e^{-(f+f'(X_i-x))}}) \cdot Y_i + \log(1 - \frac{1}{1+e^{-(f+f'(X_i-x))}}) \cdot (1-Y_i)] K(\frac{X_i - x}{h}). \tag{2.31}$$

The idea is to remove the ith point of X_t and Y_t each time for $i \in (1, 2, ..., n)$. With the new data $Y_{[-i]}$ and $X_{[-i]}$, we can estimate $\hat{f}_{[-i]}(X_i)$ using our local exponential family model and then estimate the probability. The cross validation function is thus maximised with the optimal bandwidth to be selected:

$$\hat{p}_i^{h_{[-i]}} = \frac{1}{1 + e^{-\hat{f}^h_{[-i]}(X_i)}}, \tag{2.32}$$

$$CV(h) = \sum_{i=1}^{n} [log(\hat{p}_i^{h_{[-i]}}) Y_i + \log(1 - \hat{p}_i^{h_{[-i]}})(1 - Y_i)]. \tag{2.33}$$

Similarly, for other exponential family distribution, e.g., Poisson distribution, the log likelihood function (2.29) need be re-written appropriately and the cross validation is defined correspondingly. This is omitted to save space.

The performance of our proposed method is then examined on the fixed points of the set $[-1, 1]$ with a grid of 0.01. To evaluate the quality of estimation, here we give a criterion, namely Squared Estimation Error(SEE), defined by

$$SEE = \frac{1}{n_{est}} \sum_{j=1}^{n_{est}} (\hat{f}(x_j) - f(x_j))^2, \tag{2.34}$$

where $f(x_j) = log(\frac{p_j}{1-p_j})$ with $p_j = p(x_j)$ and $p(x)$ defined in (2.27). Here x_j's are the points of the partition of $[-1, 1]$ into small intervals of length 0.01 with $n_{est} = 201$.

Table 2.1 summarises the statistics of bandwidth selected of three different cases $n = 200$, 400 and 800 with 100 replications. Together with Figure 1, it clearly shows that with the increase of sample size n, the local exponential

Table 2.1: Statistics of optimal bandwidth selected for 3 different sizes of sample (n=200, 400 and 800) with 100 repetitions

	Min	1st Qu.	Median	Mean	3rd Qu.	Max.
n=200	0.3603	0.6758	0.7885	0.8020	0.9512	1.4409
n=400	0.2891	0.5578	0.6572	0.6517	0.7656	0.9089
n=800	0.1806	0.4663	0.5465	0.5240	0.6108	0.7029

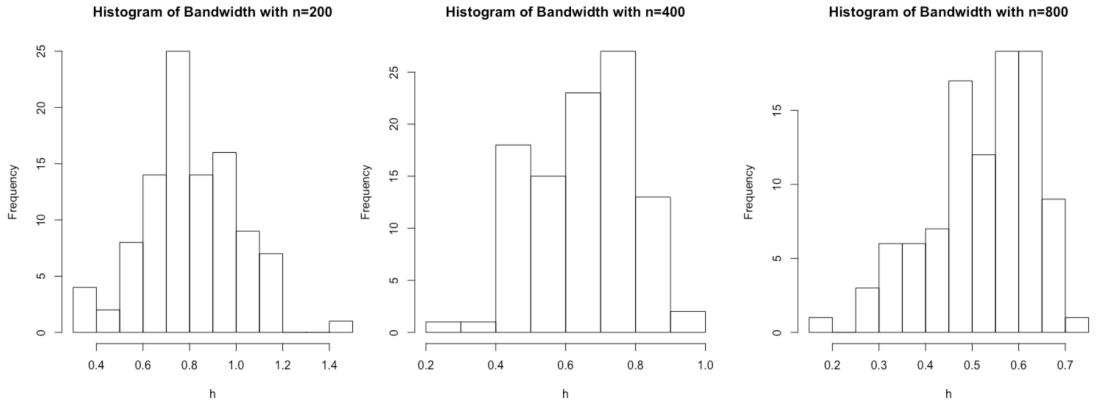

Figure 1: Bandwidth selected for sample size $n = 200$, $n = 400$ and $n = 800$ with 100 repetitions

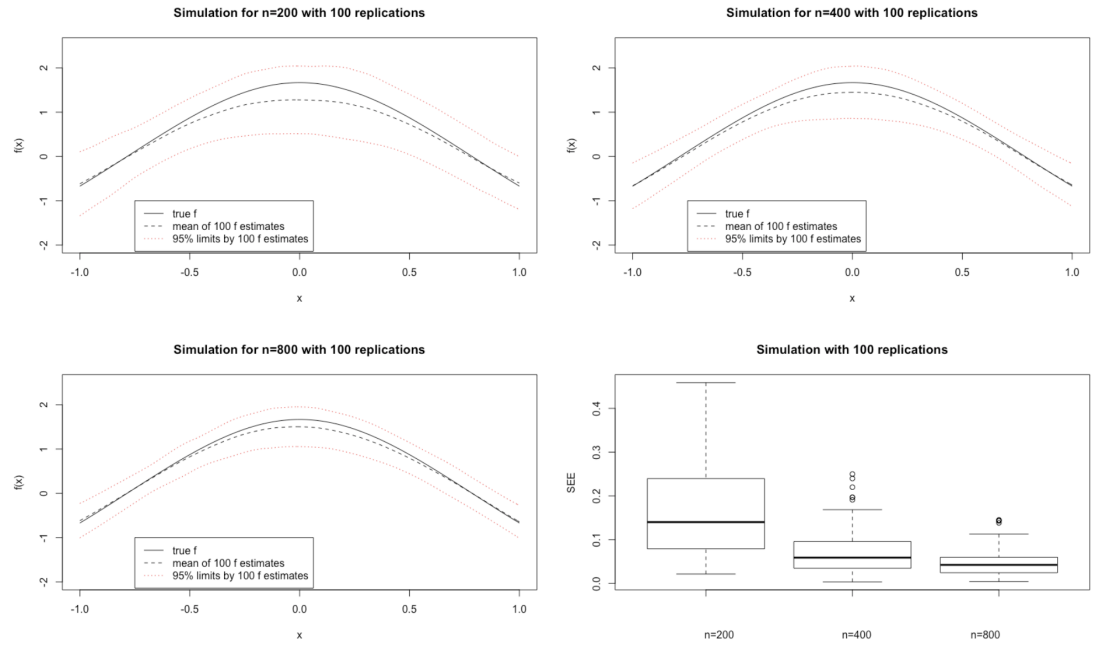

Figure 2: Estimation results of sample size $n = 200$, $n = 400$ and $n = 800$ with 100 repetitions

family model would require a smaller kernel to capture the insight of data over time, which is consistent to the expectation. The estimation results, as depicted in Figure 2, further confirms that with larger number of sample size n, the estimation would converge to the real value. It also indicates the difficulty in estimating the curves by a small range of local observations due to the fact that there might be a sequence of all $Y_t = 1$s or $Y_t = 0$s. The box-plot of SEE indicates that all three cases perform well with small errors and few outliers. However, larger sample size would further increase the estimation accuracy as suggested by the narrower 95% confidence level range and smaller SEE mean in the case of $n = 800$.

In summary, the performance of our proposed model combined with the bandwidth selection technique is quite well in estimation when the actual data has mixing structure.

2.4.2 An illustrative application to the COVID-19 daily increase in UK

In this subsection, we will introduce a simple application of the local Poisson estimation in healthcare forecasting. We have collected roughly 9 months data of COVID-19 daily increase number (UKGovernment, 2021). The data covers the time period from 16th-Jan-2020 to 1st-Sep-2020 in UK, consisting of 230 observations in total. We will estimate the daily increase number Y_t, given some known information X_t. Owing to curse of dimensionality for non-parametric estimation of $\mu_t = E(Y_t|X_t)$ with the dimension d of X_t being large, we only give a simple illustration, with X_t being taken with $d = 1$. For more practical scenario of high dimension d, some kind of semiparametric models will be necessary, which is left for research elsewhere.

Here, as a demonstration, we consider two cases for X_t. In Case 1, the past value Y_{t-i}, say $i = 7$, for X_t is considered on one-week lag effect, where we will see Y_t as discrete-valued for count number, but simply put $X_t = Y_{t-7}$ as continuous-valued so that our method can be applied in this chapter. In Case 2, alternatively, we will take X_t for the log of UK Daily News Index, which can be seen as continuous-valued more naturally. This UK Daily News Index is also known as newspaper-based Economic Policy Uncertainty (EPU) Index (EconomicPolicyUncertainty, 2021), which is considered as people may be interested in how COVID-19 is connected to our daily life in many different

aspects. The data is divided into two samples. The training sample contains the first 200 observations to fit the model. The predicting sample contains the rest 30 observations to validate the ability of prediction.

Suppose that $Y_t|X_t \sim Poisson(\lambda_t)$ (as it is reported to be roughly symmetric and bell-shaped in epidemiology studies, see also Farr (1840)). We can estimate the log conditional mean of Y_t given X_t, that is $\log \lambda_t = f(X_t)$, using the proposed method. Here λ_t can be interpreted as the expected daily increase rate of COVID-19.

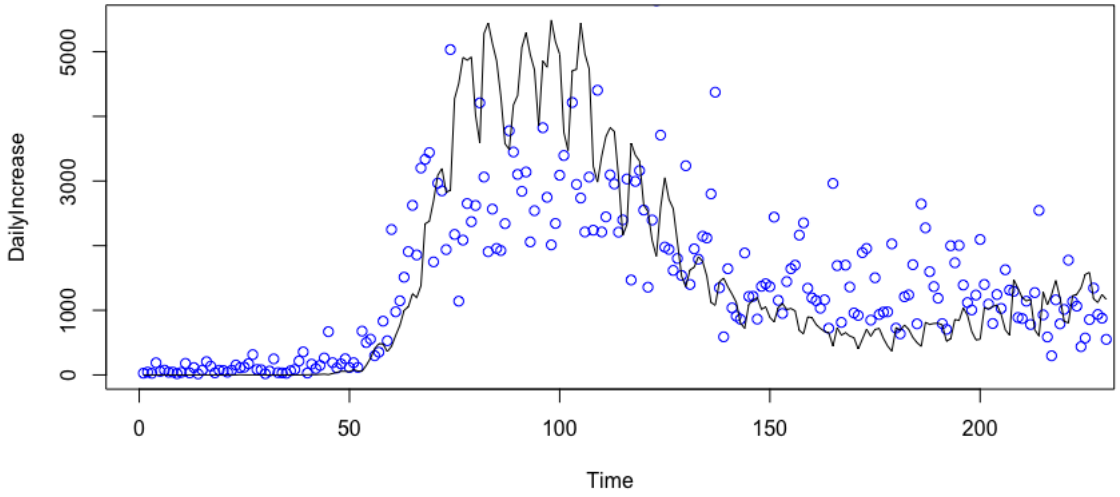

Figure 3: Estimated Daily Increase (Blue dots) based on EPU Index versus Actual Daily Increase (Black line)

We first look at Case 2, with the estimation of $\lambda_t = \exp\{f(X_t)\}$ based on EPU Index. The estimations of λ_t at each time t are depicted in Figure (3). It indicates that there is a very weak (and maybe even weaker) correlation between it and the daily increase number, as the Index itself covers a rather too wide aspects. For example, after the daily increase Y_t has been controlled, e.g., during the quarantine, we still have news with regard to policies and vaccine. The Brexit is also an important factor that may impact EPU Index better than the daily increase number. We also examined the estimation based on the lags of $\log(EPU)$, with similar outcomes omitted here. As a consequence, the estimation based on the logarithm of EPU fails to provide the accurate estimation nor the prediction.

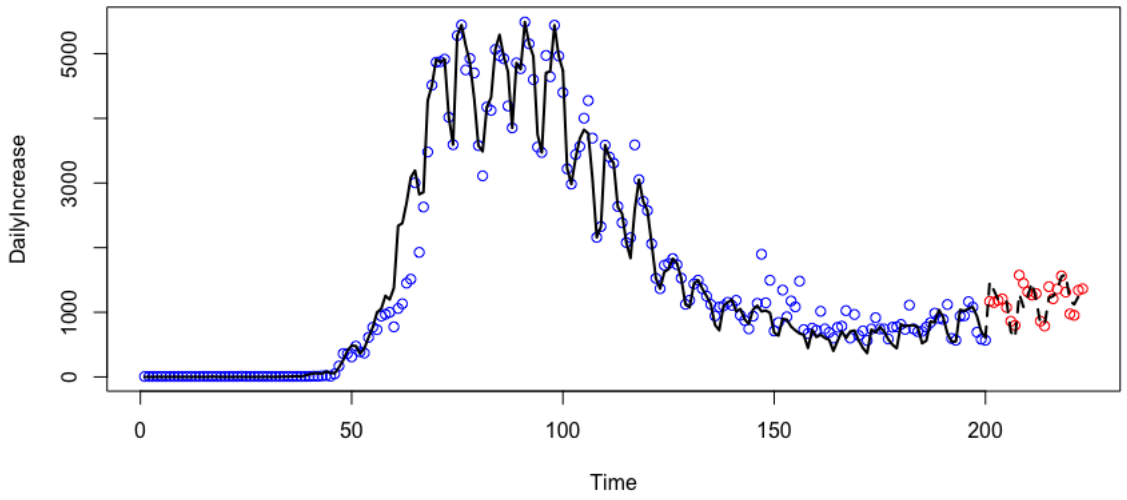

Figure 4: Estimated Daily Increase (blue dots) and Predicted Daily Increase (red dots) base on past information Y_{t-7} versus Actual Daily Increase (black line)

We now look at Case 1. The usage of past information is widely tested in the domain of time series. In this example, we find that the daily increase number Y_t has a week pattern. By applying our model to $X_t = Y_{t-7}$ and estimate λ_t at each t, which is provided in Figure (4). It shows that the lagged value $X_t = Y_{t-7}$ can provide the much better information and thus results, including both estimation and prediction. Such weekly pattern may be a result of the incubation period and diagnosis as it is now known that it takes on average 5 days (range 1-11 days and the maximum is 14 days) for the patient to show symptoms and then it may take some time for the patient to be treated and confirmed by NHS; see also Lauer et al. (2020).

To further benchmark the performance of our model, we fit the data also into a GLM model with Poisson family based on the same information. The results of the estimated λ_t over the prediction sample period are plotted with red dots for GLM in Figure (5), where the predictions by our proposed local linear method are coloured in blue, with actual observations in black. It is therefore obvious that allowing the relationship to be nonlinear by our method shows its value.

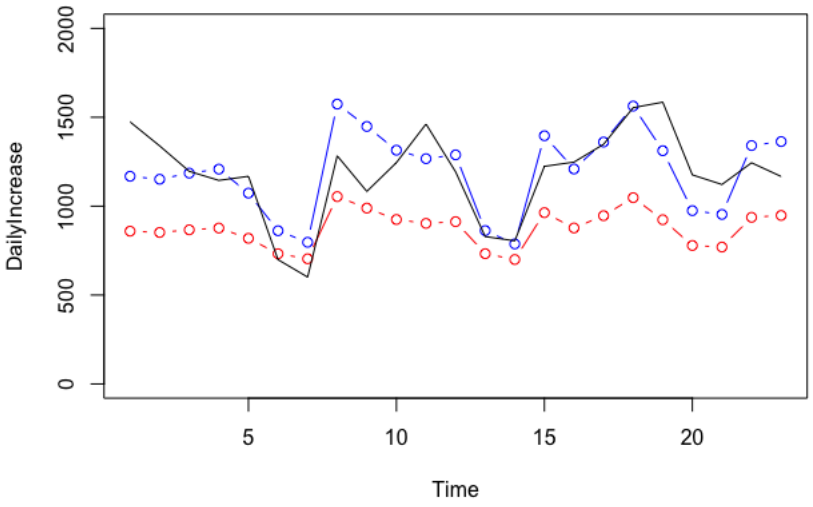

Figure 5: Predicted Daily Increase by Local Linear Regression (blue dots) and Generalised Linear Regression (red dots) based on past information $X_t = Y_{t-7}$, versus Actual Daily Increase (black line)

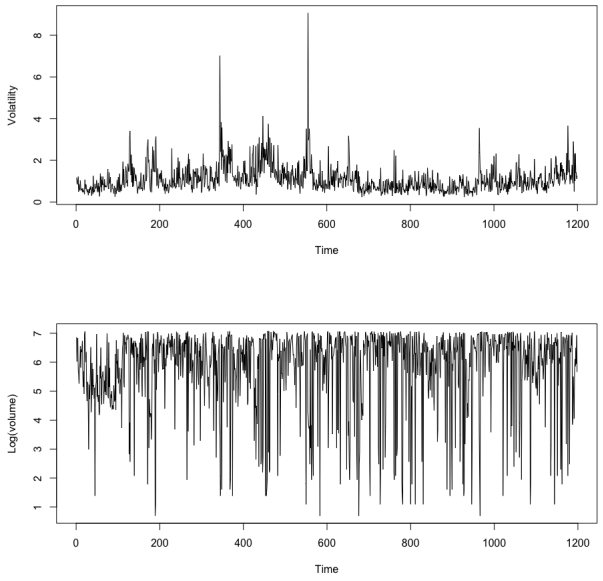

Figure 6: The time series plot of volatility and log(volume)

2.4.3 An application: forecasting FTSE100 index

In this subsection, we will introduce an application of the local logistic estimating in financial data where 5 years-long data set of FTSE100 index is used. The data includes the close price cp_t, open price op_t, high price $maxp_t$, low price $minp_t$ and volume Vlm_t within the time period from 14 April 2014

to 13 May 2019 consisting 1283 observations in total. We will estimate the marginal distribution of market evolution Y_t, whether the market price go up ($Y_t = 1$) or not ($Y_t = 0$), given the log volume V_t , volatility v_t and geometric return G_t respectively, by

$$Y_t = \begin{cases} 1 & \text{if } cp_t - cp_{t-1} > 0 \\ 0 & \text{else,} \end{cases} \tag{2.35}$$

$$\begin{aligned} v_t &= \log(100\frac{(maxp_t - minp_t)}{\frac{1}{2}(maxp_t + minp_t)}), \\ V_t &= \log(Vlm_t), \\ G_t &= 100log(\frac{cp_t}{cp_{t-1}}). \end{aligned} \tag{2.36}$$

The three series of volatility v_t, log-volume V_t and geometric return G_t are depicted in figure (6). Note that $Y_t = I(G_t > 0)$, with $I(\cdot)$ standing for an indicator function.

By assuming $Y_t|\log(v_t)(\text{or } V_t, G_t) \sim binomial(1, p_t)$, we can estimate the marginal log odds $f(\cdot) = \log \frac{p_t}{1-p_t}$ using our method. To further give an boundary to the estimation, we use the bootstrap confidence interval with 100 repetitions. The idea is to regenerate $\dot{Y}_t$ using the binomial distribution with the estimated $\hat{p_t}$. Now substitute $\dot{Y}_t$ into our estimations, we can have 100 estimations for each $f(\cdot)$. Thus provide a standard deviation that can be used to draw the confidence level.

Figure 7 shows the estimated $f(\cdot)$ for different sample sizes of three cases. Note that, we assume there is dependency in the data hence the last v_t in each sample will be omitted as well as the first Y_t. Thus we are estimating the market change today based on the trading volume (or volatility & geometric return) of yesterday. The estimated marginal log odds $f(\cdot)$ are similar but with small adjustments of the slops.

Note that the bootstrap 95 percentage confidence level have provided the possible regions of $f(\cdot)$. With the increase of sample size N, the confidence bounds tend to be narrower for both cases, which suggests a smaller error term. According to the confidence interval, we notice that for volume and

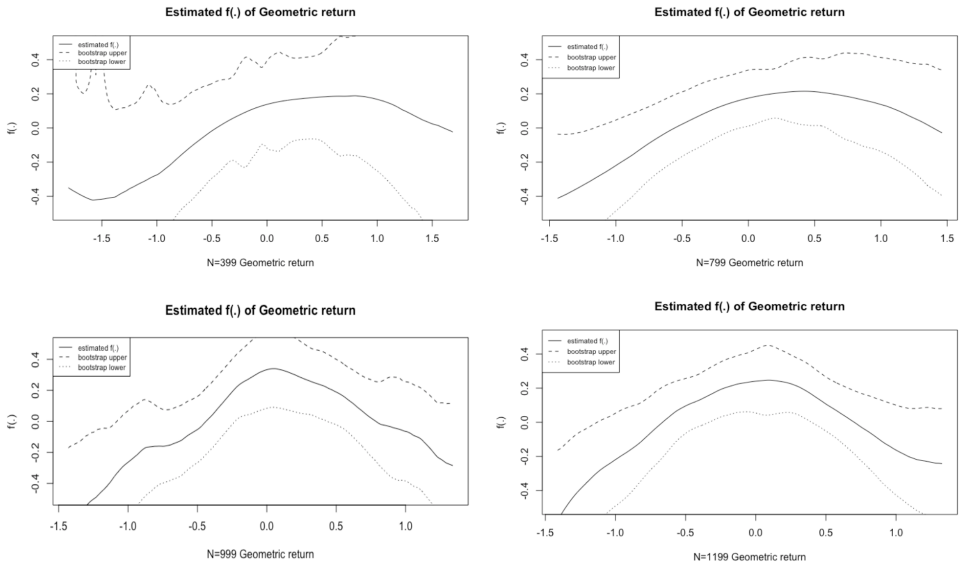

Figure 7: Estimation with different sample size

volatility, constant zero is always contained in the 95% intervals, which suggests that there is no evidence to make any conclusion based on these two variables and they are not statistically significant. This is to say, among the three different variables, only geometric return seems to have significant relationship to the market directions. It also indicates the difficulty of estimating the marginal probability of stock performance based on high frequency financial data.

However, the forecasting results in Figure 8 show that the performance of our method is good, compared to the benchmark of the liner regression model. Here the accuracy is the ratio of correct forecasts divided by the total forecasts made $(1 - \frac{\sum|\hat{Y}_t - Y_t|}{N_{forecast}})$ and we set any forecast $\hat{Y}_t = 1$ if $\hat{p}_t > 0.5$. The performance of our proposed non-parametric method clearly shows the advantage of adopting nonlinear relationships in the estimation and forecasting of such financial data. Linear model can only make pure guessing given any of the three covariates, while our local linear method suggests that the geometric return can, to some degree, help explaining the future market direction.

One shall also note that normal time series techniques such as ARIMA cannot be adopted when there is binary variable, which further confirms the advantage of our method. Our results also suggest that, estimating the stock performance based on a single variable is very difficult. Hence the estimation

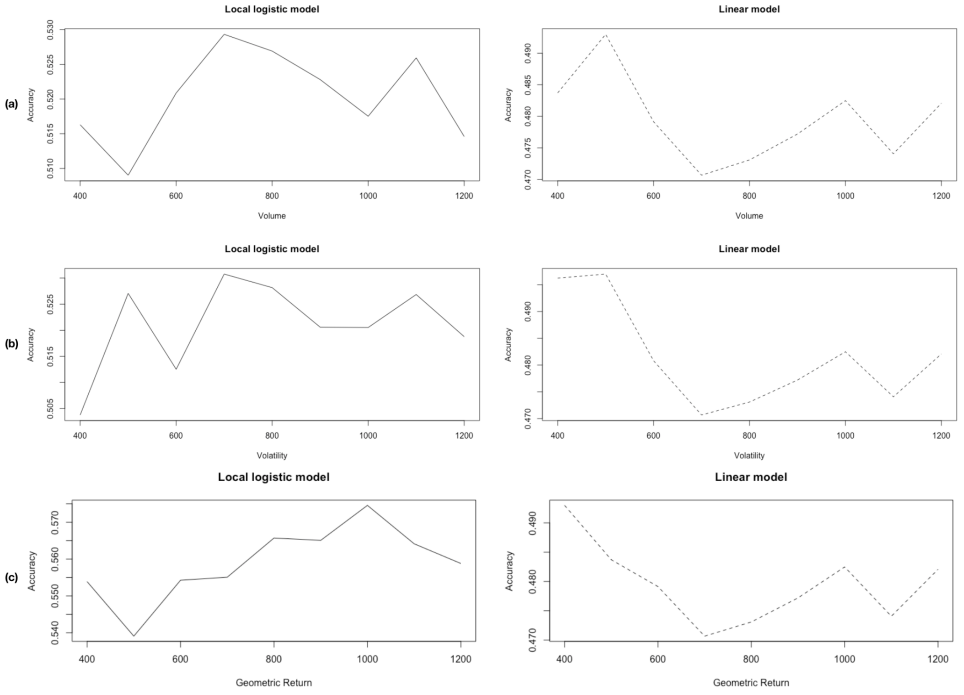

Figure 8: (a) is the accuracy of log(volume) for local logistic model and linear model estimations ;(b) is the accuracy of volatility for local logistic model and linear model estimations ; (c) is the accuracy of geometric return for local logistic model and linear model estimations.

based on it could be only a little bit better than pure guessing. However, similar to the other non-parametric methods, the curse of dimensionality would make the computational cost of adding independent variables directly to be very large. This calls the future study based on our local logistic model to overcome such computational difficulty.

In summary, the performance of our proposed generalised local linear method shows great potential in dealing with discrete-valued time series. The application of empirical data further indicates that such method can well capture the nonlinear relationship between response and covariate. Future usage of it in the areas of discrete-valued time series analysis and forecasting is therefore warranted.

2.5 Conclusion

In this chapter, we have introduced a generalised local linear fitting of discrete-valued time series under mixing conditions. Theoretical results including the uniform consistency property and corresponding proofs are presented. A simulation study of binomial distributed time series is used to illustrate the performance of our method. In addition, the applications to COVID-19 and FTSE100 Index data are examined. Results of these numerical examples show the great power and potential of our method. We thus believe it can contribute to the further development of discrete-valued time series estimation and forecasting in the future.

The investigation of non-parametric smoother for time series data is still an active area that can be applied to many disciplines. The results in this chapter thus can further contributes to the studies related to count time series data. Based on the result of this chapter, further research to establish optimal uniform convergence rates for time series data need more efforts to make, which is left for future research.

Chapter 3

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

Binary classification is an important issue in many applications but mostly studied for independent data in the literature. In this chapter, we investigate binary time series classification by proposing a semiparametric procedure named a "Model Averaging nonlinear MArginal LOgistic Regressions" (MAMaLoR) for binary time series data based on the time series information of covariates and their lags. 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. We have hence suggested a time series conditional likelihood method for estimating the optimal averaging weights together with local maximum likelihood estimations of the non-parametric 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. With cheap computational cost of low-dimensional estimation, our procedure 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 FTSE100 index data.

3.1 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 et al. (1970)). Further development of nonlinear and non-parametric analysis of that kind of time series data can be found in Tong (1990), Fan and Yao (2003), Gao (2007) and Terasvirta 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). Alternatively, Li et al. (2015) have recently introduced a novel procedure for forecasting the unknown future by conditional time series regression with high-dimensional time series lagged data, 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 et al. (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 chapter 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 and Snell (1989)). Our aim in this chapter is thus 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 covariates, 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 chapter. In fact, binary classification has been thought of as one of the most important problems in machine learning and statistics (c.f., Ryabko and Mary (2013)).

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 and 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 et al. (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 and Shephard (2003) and Liesenfeld et al. (2006) in financial modelling and Turner et al. (2011) and Buckley and Bulger (2012) in epidemiological assessments and clinical management. Similarly, an Integer-valued GARCH model (INGARCH) has been proposed by Ferland et al. (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 and Wu (2009) (Experiment 2 on Page 743) and de Oliveira Maia et al. (2021) (Subsection 2.3). For a comprehensive review on the related developments, the reader is referred to Davis et al. (1999) and Davis et al. (2016).

Though linearity is widely adopted in the literature, it may often be too strong to be appreciated when dealing with unknown data. 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. The assumptions made on linear parametric relationship may be incorrect if we don't have prior knowledge about the true relationship between the lagged covariates and the response. Differently from the parametric discrete-valued time series models above, in this chapter, we will therefore suggest utilising non-parametric method where 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 non-parametric approach in the literature is local fitting or kernel smoother of unknown functions (c.f., Fan et al. (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 non-parametric local linear fitting of the low-dimensional conditional marginal logistic regressions. The idea of maximum likelihood local fitting can be traced back to Tibshirani and Hastie (1987) and Fan and Gijbels (1995) for independent and identically distributed (i.i.d.) data, and Fan and 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 chapter, 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 et al. (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 and 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 and 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 3.4.2 below) We will also show in the data examples that our MAMaLoR procedure is not only easy to implement, but also work better in classification forecasting than GAM.

The structure of the rest of the chapter 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.

3.2 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 covariates and their lags 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:

$$p_t = P(Y_t = 1|I_{t-1}). \tag{3.1}$$

Because of the curse of dimensionality, it is well known that a direct non-parametric 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:

$$p_{jt} = P(Y_t = 1|x_{jt}),\ j = 1, ..., d. \tag{3.2}$$

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 non-parametric logistic regression is

$$logit(p_{jt}) \equiv \log \frac{p_{jt}}{1 - p_{jt}} = f_j(x_{jt}), \tag{3.3}$$

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

$$p_{jt} = F(f_j(x_{jt})). \tag{3.4}$$

Our second step is to combine the marginal logistic regressions together with a constant to approximate our concerned p_t in (3.12) by using the idea of model average as follows:

$$\begin{aligned} logit(p_t) &\approx \alpha_0 + \alpha_1 logit(p_{1t}) + ... + \alpha_d logit(p_{dt}) \\ &= \alpha_0 + \alpha_1 f_1(x_{1t}) + ... + \alpha_d f_d(x_{dt}) \equiv f_t^{MA}, \end{aligned} \tag{3.5}$$

where $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, ..., \alpha_d)$ is the vector of unknown coefficients. Indeed the direct motivation for equation (3.5) 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 (3.5) 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 (3.5) 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 (3.5) can alternatively be expressed as

$$F^{-1}(p_t) = \log(\frac{p_t}{1-p_t}) \approx \alpha_0 + \sum_{j=1}^{d} \alpha_j F^{-1}(E(Y_t|x_{jt})), \tag{3.6}$$

where $E(Y_t|x_{jt}) = P(Y_t = 1|x_{jt}) = p_{jt}$. Therefore our (3.6) 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 is $E\{E(Y_t|I_{t-1}) - \alpha_0 - \sum_{j=1}^{d} \alpha_j E(Y_t|x_{jt})\}^2$ 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 (3.5) and (3.6) 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 (3.5),

$$KL(p_{yt}, q_{yt}) = E_{p_{yt}}\{\log(p_{yt}/q_{yt})\}, \tag{3.7}$$

which is minimised with respect to $\boldsymbol{\alpha}$; we denote this minimiser by $\boldsymbol{\alpha}^{(0)}$. Note that (3.5) or (3.6) 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 et al. (2016)). We hence need to estimate the minimiser by maximum likelihood estimation below.

We make some comments before ending this section. Firstly, in this chapter we focus on the MAMaLoR procedure as given in (3.5) or (3.6) 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 chapter. Here our MAMaLoR approximation given in (3.5) or (3.6) shares a similar model form as a special case of GAM (Hastie and 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 (3.5) 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 and Yang (2016). In Lahiri and 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 chapter) 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 chapter. 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.

3.3 Estimation and Properties

3.3.1 Estimation

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

To estimate the weight coefficients in (3.5), as $f_j(x_{jt})$'s are unknown, we need to estimate them first. Here nonparametric smoother is used to estimate the marginal probability $p_{jt} = E(Y_t = 1|x_{jt})$ through that given in (3.3). We suggest applying maximum likelihood local linear fitting (c.f., Fan et al. (1998a)) for estimation of $f_j(\cdot)$ in (3.3) 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

$$\begin{aligned} f_j(x_{jt}) &\approx f_j(x_{j0}) + f_j'(x_{j0})(x_{jt} - x_{j0}) \\ &\equiv \beta_1 + \beta_2(x_{jt} - x_{j0}), \end{aligned} \tag{3.8}$$

if $|x_{jt} - x_{j0}| \leq h$, 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 (3.3) and (3.8) by:

$$\begin{aligned} \ell(\boldsymbol{\beta}, x_{j0}, h) = \sum_{t=1}^{n} [Y_t(\beta_1 + \beta_2(x_{jt} - x_{j0})) \\ - \log(1 + \exp(\beta_1 + \beta_2(x_{jt} - x_{j0})))] K_h(x_{jt} - x_{j0}), \end{aligned} \tag{3.9}$$

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

$$\begin{bmatrix} \hat{f}_j(x_{j0}) \\ \hat{f}_j'(x_{j0}) \end{bmatrix} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \arg\max_{\beta_1, \beta_2} \ell(\boldsymbol{\beta}, x_{j0}, h). \tag{3.10}$$

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 (3.8). Since x_{j0} is chosen arbitrary, we now let x_{j0} go through each point in x_{jt} and hence get the estimated marginal probability $\hat{p}_{jt} = F(\hat{f}_j(x_{jt}))$, where we recall $F(y) = e^y/(1 + e^y)$.

Now we can try to estimate the coefficients in (3.5) 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 (3.7) 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 (3.5), we can define the (approximate) conditional likelihood function as follows:

$$\begin{aligned} L(\boldsymbol{\alpha}) &= \prod_{t=1}^{n} P(Y_t = y_t | I_{t-1}; \boldsymbol{\alpha}) \\ &= \prod_{t=1}^{n} (p_t(\boldsymbol{\alpha}))^{y_t} (1 - p_t(\boldsymbol{\alpha}))^{1-y_t}, \end{aligned} \tag{3.11}$$

where

$$p_t(\boldsymbol{\alpha}) = \frac{e^{\alpha_0 + \sum_{j=1}^{d} \alpha_j f_j(x_{jt})}}{1 + e^{\alpha_0 + \sum_{j=1}^{d} \alpha_j f_j(x_{jt})}}. \tag{3.12}$$

Note that (3.11) can be also viewed as a kind of composite likelihood; see Varin et al. (2011). Then taking nature log of the equation (3.11) together with (3.12), 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

$$\hat{l}(\boldsymbol{\alpha}) = \frac{1}{n} \sum_{t=1}^{n} \left[y_t(\alpha_0 + \sum_{j=1}^{d} \alpha_j \hat{f}_j(x_{jt})) - \log(1 + e^{\alpha_0 + \sum_{j=1}^{d} \alpha_j \hat{f}_j(x_{jt})}) \right]. \tag{3.13}$$

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 (3.13), and define the following modified log-likelihood function:

$$\begin{aligned} l_n(\boldsymbol{\alpha}) = l_n(\hat{\mathbf{f}}(\cdot), \boldsymbol{\alpha}) = \frac{1}{n} \sum_{t=1}^{n} \Bigg[& \left\{ Y_t \left(\alpha_0 + \sum_{j=1}^{d} \alpha_j \hat{f}_j(x_{jt}) \right) \right\} \\ & - \log \left\{ 1 + \exp \left(\alpha_0 + \sum_{j=1}^{d} \alpha_j \hat{f}_j(x_{jt}) \right) \right\} w(X_t), \end{aligned} \tag{3.14}$$

which asymptotically corresponds to the population log-likelihood function:

$$l(\mathbf{f}(\cdot), \boldsymbol{\alpha}) = E \left[\left\{ Y_t \left(\alpha_0 + \sum_{j=1}^{d} \alpha_j f_j(x_{jt}) \right) \right\} - \log \left\{ 1 + \exp \left(\alpha_0 + \sum_{j=1}^{d} \alpha_j f_j(x_{jt}) \right) \right\} \right] w(X_t), \tag{3.15}$$

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 $w(X_t) = \prod_{j=1}^{d} \mathbf{I}_{(c_{0j} \leq x_{jt} \leq c_{1j})}$ 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 et al. (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 (3.14) with respect to α_j:

$$\frac{\partial l_n(\boldsymbol{\alpha})}{\partial \alpha_j} = \frac{1}{n} \sum_{t=1}^{n} [y_t \hat{f}_j(x_{jt}) - \hat{p}_t \hat{f}_j(x_{jt})] w(X_t), \tag{3.16}$$

where

$$\hat{p}_t = \hat{p}_t(\boldsymbol{\alpha}) = \frac{e^{\alpha_0 + \sum_{j=1}^{d} \alpha_j \hat{f}_j(x_{jt})}}{1 + e^{\alpha_0 + \sum_{j=1}^{d} \alpha_j \hat{f}_j(x_{jt})}}. \tag{3.17}$$

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

$$\frac{\partial^2 l_n(\boldsymbol{\alpha})}{\partial \alpha_j \partial \alpha_k} = -\frac{1}{n} \sum_{t=1}^{n} \hat{f}_j(x_{jt}) \hat{p}_t (1 - \hat{p}_t) \hat{f}_k(x_{kt}) w(X_t). \tag{3.18}$$

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

From the computational perspective, note that equation (3.5) 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.

3.3.2 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

$$p_t(\mathbf{f}(\cdot), \boldsymbol{\alpha}) = \frac{e^{\alpha_0 + \sum_{j=1}^{d} \alpha_j f_j(x_{jt})}}{1 + e^{\alpha_0 + \sum_{j=1}^{d} \alpha_j f_j(x_{jt})}}. \tag{3.19}$$

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 given in definition 2.1.

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 (3.3). For an $\mathbf{f}(\cdot)$, define its Lipschitz norm: For some $\varpi > 0$, let $[\varpi]$ be the largest integer not

greater than ϖ, and define (if it exists)

$$\|\mathbf{f}\|_{\infty,\varpi} = \max_{0\le\kappa\le[\varpi]} \sup_{x\in A} \|\mathbf{f}^{(\kappa)}(x)\| + \sup_{x\neq x';x,x'\in A} \frac{\|\mathbf{f}^{([\varpi])}(x) - \mathbf{f}^{([\varpi])}(x')\|}{\|x - x'\|^{\varpi-[\varpi]}}, \tag{3.20}$$

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 $\varpi \ge 2$:

$$\mathbf{F} := \{\mathbf{f} : \text{continuous from } A \text{ to } R^d \text{ with } \|\mathbf{f}\|_{\infty,\varpi} \le c\}, \tag{3.21}$$

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 $C_c^{\varpi}(A)$.

A5 For the local likelihood function (3.9), define $\Lambda(Y_t, z_j) = Y_t - \exp(z_j)/[1+\exp(z_j)]$, and

$$\Phi(x_j, z_j) = E[\Lambda(Y_t, z_j)|x_{jt} = x_j], \tag{3.22}$$

satisfying $(x_j, z_j) \to \Phi(x_j, z_j) \cdot g_j(x_j)$ 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 $\dot{\Phi}_x$, and the derivative with respect z_j by $\dot{\Phi}_z$, etc.

Remark. (i) Assumption 1 shows a technical standard β-mixing process which is satisfied by many linear and non-linear time series models under geometric ergodicity (Fan and 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 et al. (1998b), Fan et al. (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 and 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* $\mathfrak{A}$, $\mathbf{f} \in \mathbf{F}$ *and* $nh^4 = o(1)$. *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.

Lemma 3.2. *(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 given:

Proof. Proposition 3.2 (Consistency Lemma) 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 3.2. 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) hold obviously. (iii) also holds clearly by the following fact:

$$l(\mathbf{f}(\cdot),\boldsymbol{\alpha}) = E[Y_t\widetilde{\chi}_t(\mathbf{f})^T\boldsymbol{\alpha} - \log(1+e^{\widetilde{\chi}_t(\mathbf{f})^T\boldsymbol{\alpha}})], \tag{3.23}$$

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:

$$\begin{aligned}
&\sup_{\boldsymbol{\alpha}\in\mathfrak{A}}|l(\mathbf{f}(\cdot),\boldsymbol{\alpha}) - l(\mathbf{f}_0(\cdot),\boldsymbol{\alpha})| \\
&\leq E|Y_t|\|\widetilde{\chi}_t(\mathbf{f}) - \widetilde{\chi}_t(\mathbf{f}_0)\|\boldsymbol{\alpha} + |\log(1+e^{\widetilde{\chi}_t(\mathbf{f})^T\boldsymbol{\alpha}}) - \log(1+e^{\widetilde{\chi}_t(\mathbf{f}_0)^T\boldsymbol{\alpha}})| \\
&\leq \frac{e^{\widetilde{\chi}_t(\mathbf{f})^T\boldsymbol{\alpha}}}{1+e^{\widetilde{\chi}_t(\mathbf{f})^T\boldsymbol{\alpha}}}\|\widetilde{\chi}_t(\mathbf{f}) - \widetilde{\chi}_t(\mathbf{f}_0)\|\|\boldsymbol{\alpha}\| \\
&\leq C\|\mathbf{f} - \mathbf{f}_0\|_{\mathbf{F}},
\end{aligned} \tag{3.24}$$

where C is a generic constant.

Now, to proved (iv), 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 nonlinear logistic regression can be formulated as follows:

$$\text{logit}(p_j(x_{jt})) = \log(\frac{p_j(x_{jt})}{1-p_j(x_{jt})}) = f_j(x_{jt}), \tag{3.25}$$

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 non-parametric function from R to R.

Given the local log likelihood function in (3.9), we have the following types of estimation equations:

$$\Omega_n^{(1)}(\boldsymbol{\beta}, x_j, h) = \frac{1}{n}\frac{\partial \ell}{\partial \beta_1} = \frac{1}{n}\sum_{t=1}^{n}[Y_t - \frac{\exp(\beta_1 + \beta_2^T(x_{jt} - x_j))}{1 + \exp(\beta_1 + \beta_2^T(x_{jt} - x_j))}]K_h(x_{jt} - x_j) = 0, \tag{3.26}$$

$$\begin{aligned}\Omega_n^{(2)}(\boldsymbol{\beta}, x_j, h) &= \frac{1}{nh}\frac{\partial \ell}{\partial \beta_2} \\ &= \frac{1}{n}\sum_{t=1}^{n}[Y_t - \frac{\exp(\beta_1 + \beta_2^T(x_{jt} - x_j))}{1 + \exp(\beta_1 + \beta_2^T(x_{jt} - x_j))}]\frac{x_{jt} - x_j}{h}K_h(x_{jt} - x_j) = 0.\end{aligned} \tag{3.27}$$

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 (3.25):

$$\begin{aligned}E[\Omega_n^{(1)}(\boldsymbol{\beta}, x_j, h)] &= E[\frac{1}{n}\sum_{t=1}^{n}[Y_t - \frac{\exp(\beta_1 + \beta_2(x_{jt} - x_j))}{1 + \exp(\beta_1 + \beta_2(x_{jt} - x_j))}]K_h(x_{jt} - x_j)] \\ &= E[E[\frac{1}{n}\sum_{t=1}^{n}[Y_t - \frac{\exp(\beta_1 + \beta_2)(x_{jt} - x_j))}{1 + \exp(\beta_1 + \beta_2(x_{jt} - x_j))}]]K_h(x_{jt} - x_j)|X_t] \\ &= E[\frac{1}{n}\sum_{t=1}^{n}[E[Y_t|x_{jt}] - \frac{\exp(\beta_1 + \beta_2(x_{jt} - x_{j0}))}{1 + \exp(\beta_1 + \beta_2(x_{jt} - x_j))}]K_h(x_{jt} - x_j)] \\ &= E[\frac{1}{n}\sum_{t=1}^{n}[\frac{\exp(f_j(x_{jt}))}{1 + \exp(f_j(x_{jt}))} - \frac{\exp(\beta_1 + \beta_2(x_{jt} - x_j))}{1 + \exp(\beta_1 + \beta_2(x_{jt} - x_j))}]K_h(x_{jt} - x_j)],\end{aligned}$$

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

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

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

where $o(1)$ is uniformly 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 $\Lambda(Y_t, z_j) = Y_t - \exp(z_j)/[1 + \exp(z_j)]$ as in Assumption A5, then

$$\begin{aligned} E[\Omega_n^{(1)}(\boldsymbol{\beta}, x_j, h)] &= E[\Lambda(Y_t; \beta_1 + \beta_2(x_{jt} - x_j))K_h(x_{jt} - x_j)] \\ &= E[\Phi(x_j; \beta_1 + \beta_2(x_{jt} - x_j))K_h(x_{jt} - x_j)] \\ &= \Phi(x_j, \beta_1)g_j(x_j) + O(h^2), \end{aligned} \tag{3.28}$$

where corresponding to our local logistic regression, $\Phi(x_j, \beta_1) = \tilde{f}(f_j(x_j)) - \tilde{f}(\beta_1)$, 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,

$$E[\Omega_n^{(2)}(\boldsymbol{\beta}, x_j, h)] = E[\Lambda(Y_i; \beta_1 + \beta_2(x_{jt} - x_j))\frac{x_{jt} - x_j}{h}K_h(x_{jt} - x_j)]$$
$$= h(\beta_2\dot{\Phi}_z(x_j, \beta_1)) + \dot{\Phi}_x(x_j, \beta_1))g_j(x_j) + h\Phi(x_j, \beta_1)g_j'(x_j) + O(h^3),$$

where, corresponding to our local logistic regression model, $\dot{\Phi}_x(x_j, \beta_1) = \tilde{f}'(f_j(x_j))f_j'(x_j)$
$= f_j'(x_j)\frac{e^{f_j(x_j)}}{(1+e^{f_j(x_j)})^2}$ and $\dot{\Phi}_z(x_j, \beta_1) = -\tilde{f}'(\beta_1) = -\frac{e^{\beta_1}}{(1+e^{\beta_1})^2}$, 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:

$$E[\Omega_n^{(1)}(\boldsymbol{\beta}, x_j, h)] = \Omega_0^{(1)}(\boldsymbol{\beta}, x_j) + O(h^2) \tag{3.29}$$

and

$$E[\Omega_n^{(2)}(\boldsymbol{\beta}, x_j, h)] = h\Omega_0^{(2)}(\boldsymbol{\beta}, x_j) + O(h^3) \tag{3.30}$$

where

$$\Omega_0^{(1)}(\boldsymbol{\beta}, x_j) = \Phi(x_j, \beta_1)g_j(x_j), \tag{3.31}$$

$$\Omega_0^{(2)}(\boldsymbol{\beta}, x_j) = (\beta_2\dot{\Phi}_z(x_j, \beta_1) + \dot{\Phi}_x(x_j, \beta_1))g_j(x_j) + \Phi(x_j, \beta_1)g_j'(x_j). \tag{3.32}$$

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:

$$\begin{cases} \Phi(x_j, \beta_{01}) = 0 \\ \beta_{02}(x) = -\frac{\dot{\Phi}_x(x_j, \beta_{01})}{\dot{\Phi}_z(x_j, \beta_{02})}, \end{cases} \tag{3.33}$$

which is actually unique correspondingly to our local linear logistic regression (3.9) 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

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_\infty > \delta \Rightarrow \max_{i=1,2} |\Omega_0^{(i)}(\hat{\boldsymbol{\beta}}, x_j)| > \varepsilon, \quad for \quad x \in A. \tag{3.34}$$

Therefore for the uniform consistency of $\hat{\boldsymbol{\beta}}$ to $\boldsymbol{\beta}_0$ in probability, by (3.34), 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, where C is a generic constant which may be large. The proof of (iv) is done.

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

$$\begin{aligned} l_n(\mathbf{f}(\cdot), \boldsymbol{\alpha}) - l(\mathbf{f}(\cdot), \boldsymbol{\alpha}) &= \{l_n(\mathbf{f}(\cdot), \boldsymbol{\alpha}) - l_n(\mathbf{f}_0(\cdot), \boldsymbol{\alpha})\} + \{l_n(\mathbf{f}_0(\cdot), \boldsymbol{\alpha}) - l(\mathbf{f}_0(\cdot), \boldsymbol{\alpha})\} \\ &\quad + \{l(\mathbf{f}_0(\cdot), \boldsymbol{\alpha}) - l(\mathbf{f}(\cdot), \boldsymbol{\alpha})\} \\ &= I + II + III. \end{aligned} \tag{3.35}$$

Uniformly, for $\boldsymbol{\alpha} \in \mathfrak{A}$ and f satisfying $\|\mathbf{f} - \mathbf{f}_0\|_{\mathbf{F}} \le \delta_n$, I,II and III can be proved to tend to zero. The proof is easy that we can show here III follows equation (3.24) and II is easily proved by 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 $I + II + III$ tend to zero.

By completing the checking of the conditions of Proposition 3.2 (Consistency Lemma), the proof of Theorem 2.1 is completed.

□

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 (3.3),

$$\begin{aligned}\mathbf{U} &= E[p_t(1-p_t)]\widetilde{\chi}_t(\mathbf{f}_0)\widetilde{\chi}_t(\mathbf{f}_0)^T w(X_t),\\ \mathbf{V} &= \lim_{n\to\infty} Var\left(\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(Y_t - p_t)\widetilde{\chi}_t(\mathbf{f}_0)w(X_t)\right).\end{aligned} \tag{3.36}$$

Then we have

Theorem 3.3. *(Asymptotic Normality)*

Suppose that the assumptions A1-A5 are satisfied, for $\boldsymbol{\alpha} \in \mathfrak{A}$*, and* $\mathbf{U}$ *is positive define. If* $nh^4 = o(1)$*, then*

$$\sqrt{n}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) \xrightarrow{L} N(0, \mathbf{U}^{-1}\mathbf{V}\mathbf{U}^{-1}) \tag{3.37}$$

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 3.3. The proof of Theorem 3.3 is provided in:

Proof. 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:

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

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

To obtain the bias, now we can apply the Taylor Expansion:

$$0 = l_n^{'}(\hat{\mathbf{f}}(\cdot), \hat{\boldsymbol{\alpha}}) = l_n^{'}(\hat{\mathbf{f}}(\cdot), \boldsymbol{\alpha}_0) + l_n^{''}(\hat{\mathbf{f}}(\cdot), \boldsymbol{\alpha}_0 + \xi(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0))(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0),$$
$$\sqrt{n}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) = -[l_n^{''}(\hat{\mathbf{f}}(\cdot), \boldsymbol{\alpha}_0 + \xi(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0))]^{-1}\sqrt{n}l_n^{'}(\hat{\mathbf{f}}(\cdot), \boldsymbol{\alpha}_0), \tag{3.38}$$

where $|\xi| < 1$.

Then we get them together with the consistency of $\hat{\boldsymbol{\alpha}}$ to $\boldsymbol{\alpha}_0$.

$$\sqrt{n}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) = -(1 + op(1))[l_n^{''}(\mathbf{f}_0(\cdot), \boldsymbol{\alpha}_0)]^{-1}\sqrt{n}[l_n^{'}(\mathbf{f}_0(\cdot), \boldsymbol{\alpha}_0) + O(h^2)], \tag{3.39}$$

by noting that

$$\begin{aligned} l_n^{'}(\hat{\mathbf{f}}, \boldsymbol{\alpha}_0) - l_n^{'}(\mathbf{f}_0, \boldsymbol{\alpha}_0)) &= (1 + o_P(1))\frac{1}{n}\sum_{t=1}^{n}(Y_t - p_t)(\widetilde{\chi}_t(\hat{\mathbf{f}}) - \widetilde{\chi}_t(\mathbf{f}_0))w(X_t)] \\ &= O_P(h^2) \end{aligned} \tag{3.40}$$

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

$$l_n^{''}(\mathbf{f}_0, \boldsymbol{\alpha}_0) = -\frac{1}{n}\sum_{t=1}^{n} p_t(\mathbf{f}_0, \boldsymbol{\alpha}_0)(1 - p_t(\mathbf{f}_0, \boldsymbol{\alpha}_0))\widetilde{\chi}_t(\mathbf{f}_0)\widetilde{\chi}_t(\mathbf{f}_0)^T w(X_t). \tag{3.41}$$

By law of large number, we have

$$l_n^{''}(\mathbf{f}_0, \boldsymbol{\alpha}_0) \to l^{''}(\mathbf{f}_0, \boldsymbol{\alpha}_0) = \mathbf{U} = E[p_t(1 - p_t)]\widetilde{\chi}_t(\mathbf{f}_0)\widetilde{\chi}_t(\mathbf{f}_0)^T w(X_t). \tag{3.42}$$

By central limit theorem,

$$\sqrt{n}l_n^{'}(\mathbf{f}_0, \boldsymbol{\alpha}_0) \to N(0, \mathbf{V}) \tag{3.43}$$

where

$$\mathbf{V} = \lim_{n\to\infty} Var(\frac{1}{\sqrt{n}}(Y_t - p_t)\widetilde{\chi}_t(\mathbf{f}_0)w(X_t)). \tag{3.44}$$

Thus the asymptotic variance matrix

$$Var(\hat{\boldsymbol{\alpha}}|\mathbf{f}(\cdot)) = \mathbf{U}^{-1}\mathbf{V}\mathbf{U}^{-1}. \tag{3.45}$$

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

□

3.4 Numerical evidence

In this section, we illustrate the empirical application of our proposed MA-MaLoR 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.

3.4.1 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 (3.9) 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:

$$\begin{aligned} Y_t &= I(x_t > 0), \\ x_t &= \sum_{k=1}^{9} g_{0k}(x_{t-k}) + \epsilon_t, \\ g_{0k}(x_{t-k}) &= a_k x_{t-k} + \delta \exp(-k x_{t-k})/(1+\exp(-k x_{t-k})) + \gamma \cos(x_{t-k}x_{t-1}), \end{aligned} \tag{3.46}$$

Table 3.1: Parameters specified in Model (3.46)

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

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 (3.48) 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 3.1. 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 (3.46) (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 we can adjust δ and γ with non-zero values to change the model with different degree of nonlinear structure or interaction. When $(\delta, \gamma) = (0, 0)$, the x_t process in (3.46) 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 (3.46), Y_t given $X_t := \{x_{t-1}, \ldots, x_{t-9}\}$ follows a Bernoulli distribution with probability p_t, that is

$$[Y_t | X_t] \sim Bin(1, p_t), \tag{3.47}$$

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

$$p_t = P(Y_t = 1|X_t) = P\left(\sum_{k=1}^{9} g_{0k}(x_{t-k}) + \epsilon_t > 0|X_t\right) = P\left(\epsilon_t > -\sum_{k=1}^{9} g_{0k}(x_{t-k})|X_t\right)$$
$$= 1 - F\left(-\sum_{k=1}^{9} g_{0k}(x_{t-k})\right) = F\left(\sum_{k=1}^{9} g_{0k}(x_{t-k})\right), \tag{3.48}$$

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 (3.46) 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 (3.46). For the bandwidth used in our estimation, how to select optimal one for forecasting is still an open question. We just simply 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 $n = 1000$, 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 chapter), linear logistic regression model 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,

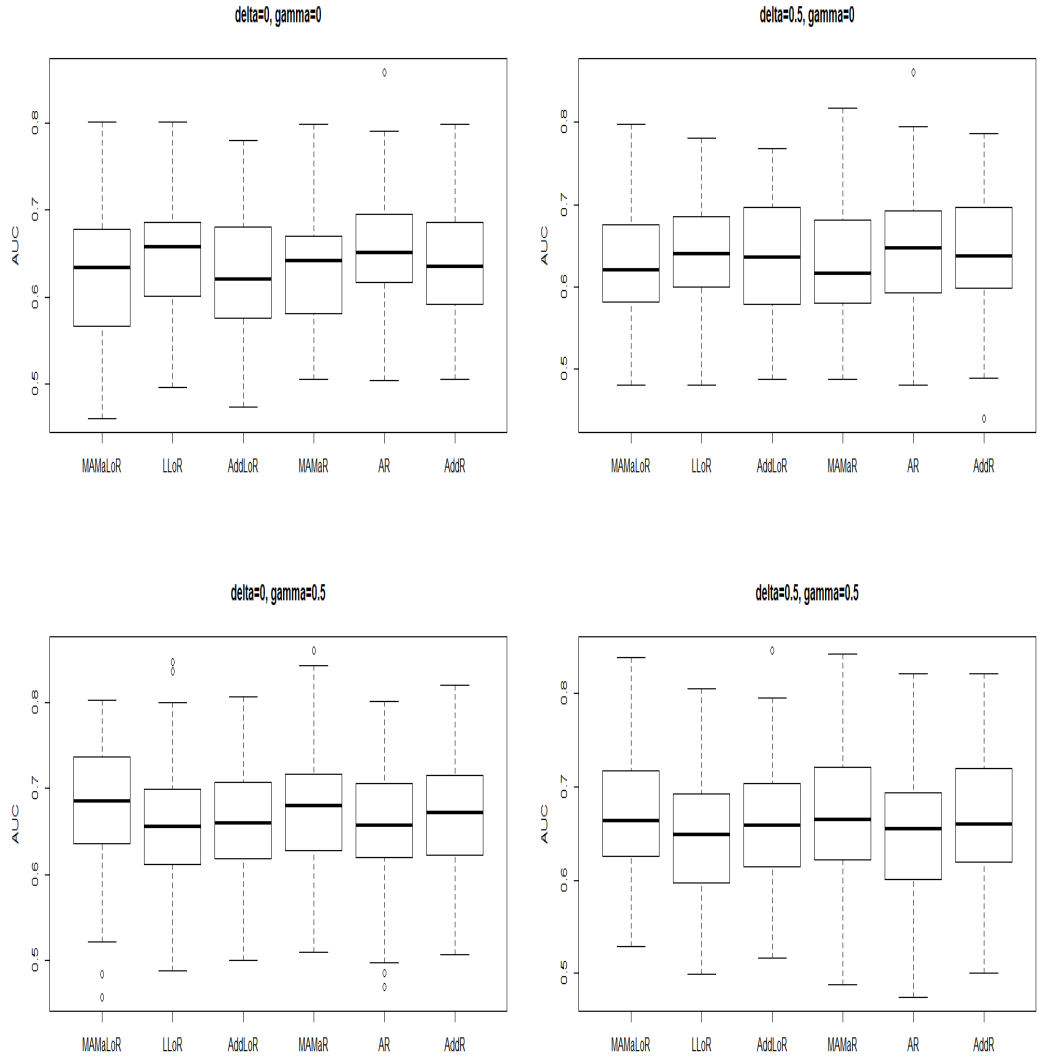

Figure 1: Boxplots of the area under curve (AUC) with 100 repetitions for one-step ahead classification predictions, with $n_p = 50$ observations for testing, of different methods under different true model structures (Top left: linear, Top right: additive, Bottom left & right: nonlinear non-additive)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

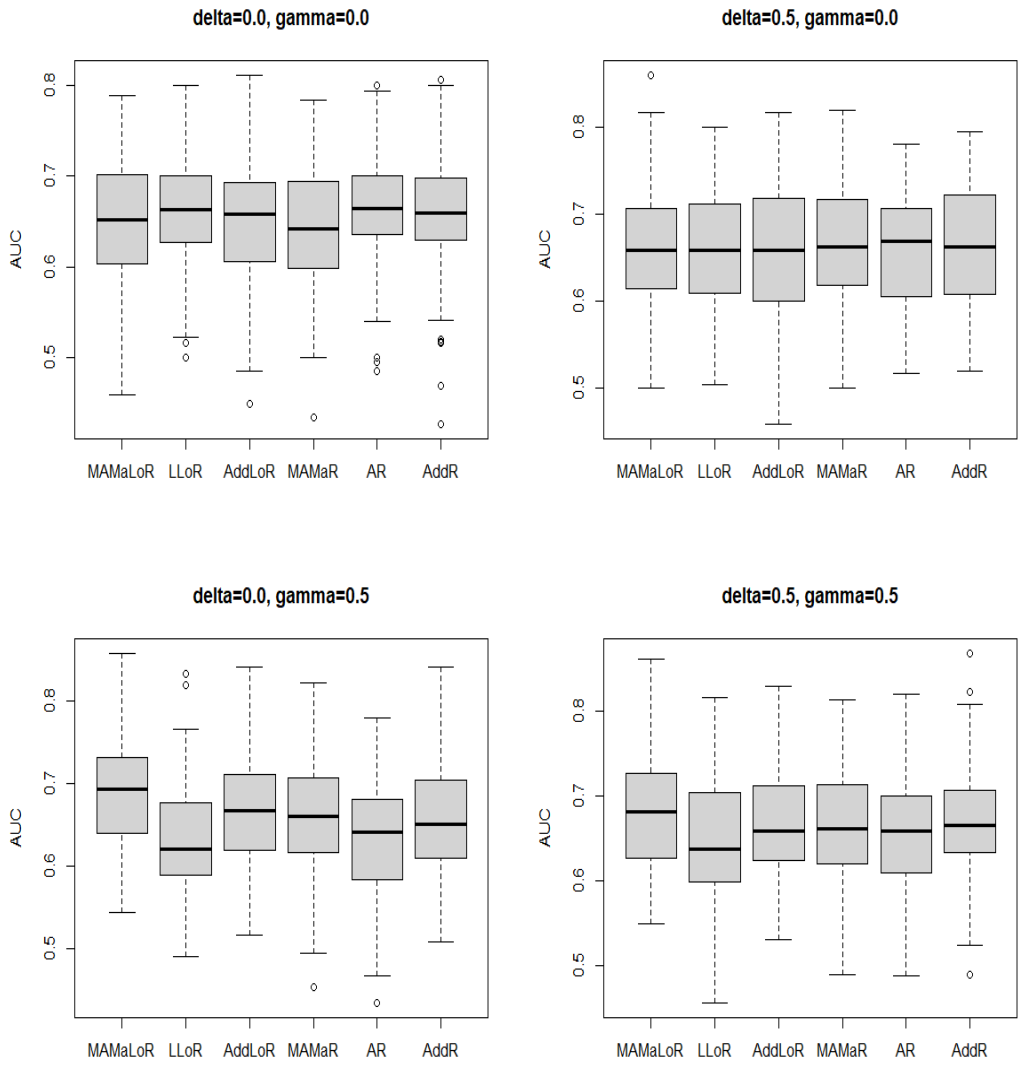

Figure 2: Boxplots of the area under curve (AUC) with 100 repetitions for one-step ahead classification predictions, with $n_p = 50$ observations for testing, of different methods under different true model structures (Top left: linear, Top right: additive, Bottom left & right: nonlinear non-additive)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

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.

(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.

(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

Table 3.2: Parameters specified in Model (3.49)

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		

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 (3.46) to 31. We use the following model along with the parameters summarised in Table 3.2 to generate data for the simulation.

$$\begin{aligned} Y_t &= I(x_t > 0), \\ x_t &= \sum_{k=1}^{31} g_{0k}(x_{t-k}) + \epsilon_t, \\ g_{0k}(x_{t-k}) &= a_k x_{t-k} + \delta \exp(-k x_{t-k})/(1+\exp(-k x_{t-k})) + \gamma \cos(x_{t-k} x_{t-1}), \end{aligned} \tag{3.49}$$

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 (3.49) has an analogue setting to (3.46). Similar to (3.46), 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 settings of non-additive data structure in (3.49), 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 as follows: (i) when the model is not additive ($\gamma \neq 0$), the

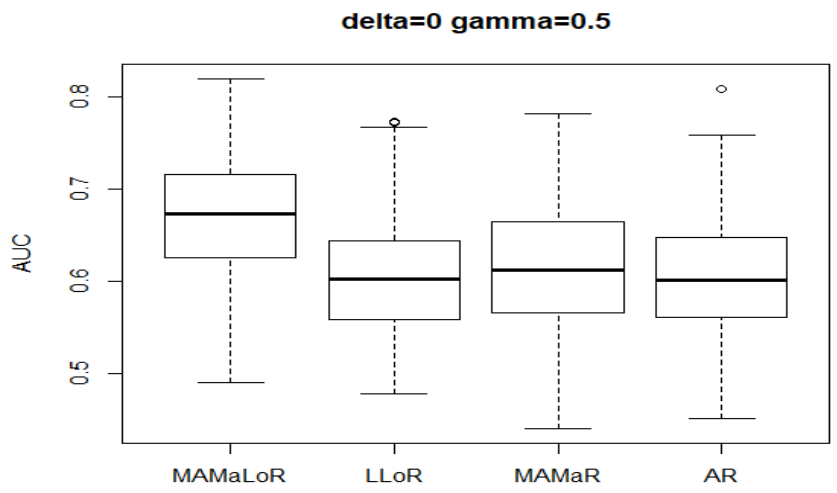

Figure 3: Boxplots of the area under curve (AUC) with 100 repetitions for one-step ahead classification predictions of non-additive data, with $n_p = 50$ observations for testing, for $lag = 31$, based on $n = 1000$ observations for training.

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 are comparable to that for LLoR and AR model, as it only increases in polynomial time when adding more lags (i.e., enlarge the dimensionality 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 variable are added, we should take care to select the relevant variable for prediction, which is beyond the scope of this chapter and left for further study.

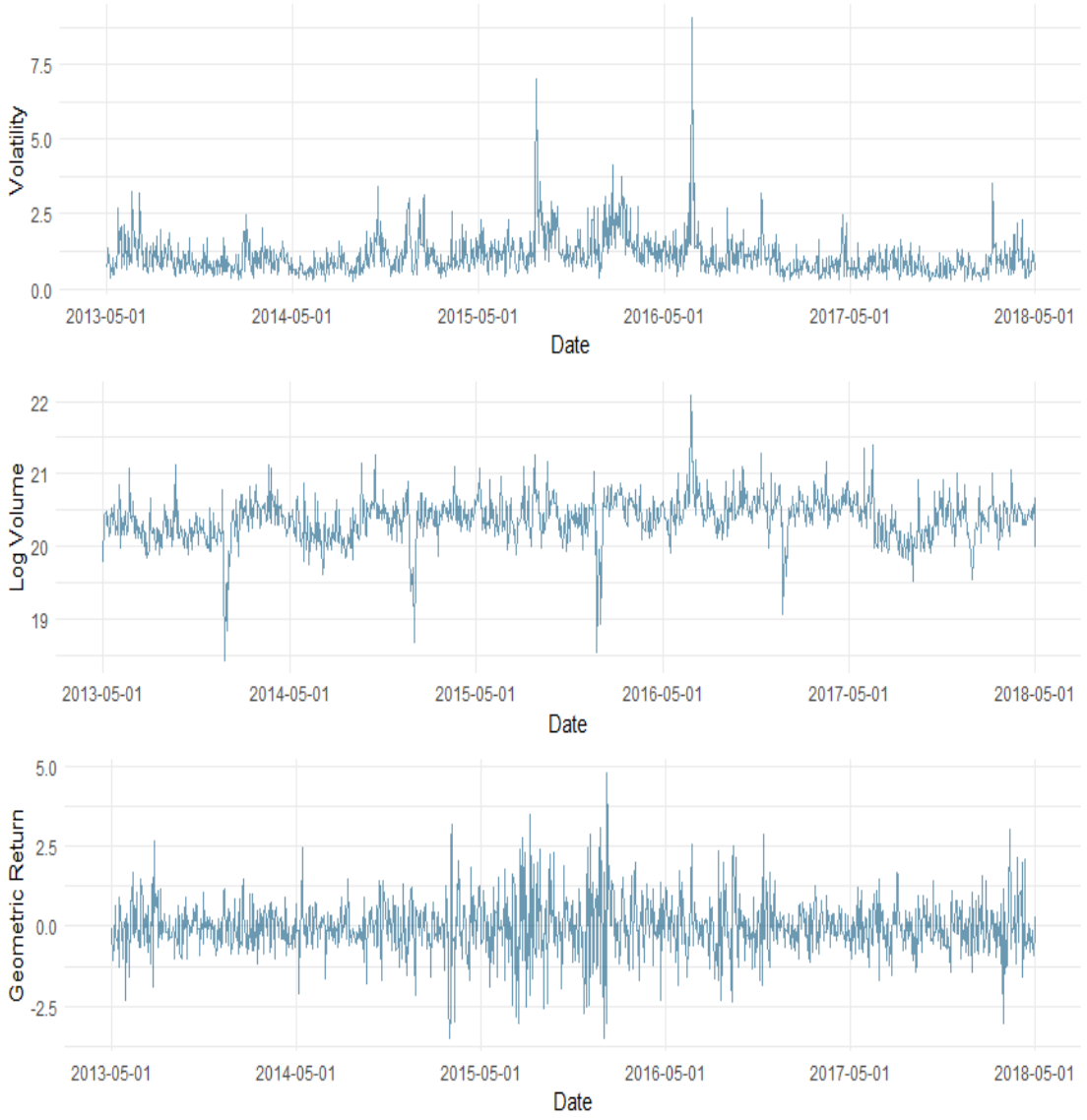

Figure 4: The time series plot of volatility v_t, log-volume V_t and geometric return G_t defined in (3.50).

3.4.2 An application: forecasting market moving direction of FTSE 100 index

In this section, we demonstrate practical advantages of our proposed MA-MaLoR 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 $maxp_t$ 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 going 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

$$Y_t = \begin{cases} 1 & \text{if } cp_t - cp_{t-1} > 0 \\ 0 & \text{else,} \end{cases} \tag{3.50}$$

$$\begin{aligned} v_t &= \log(100\frac{(maxp_t - minp_t)}{\frac{1}{2}(maxp_t + minp_t)}), \\ V_t &= \log(Vlm_t), \\ G_t &= 100log(\frac{cp_t}{cp_{t-1}}). \end{aligned}$$

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 et al. (2015)).

We first estimate the marginal logistic regressions $f_j(\cdot)$'s in (3.3) 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,

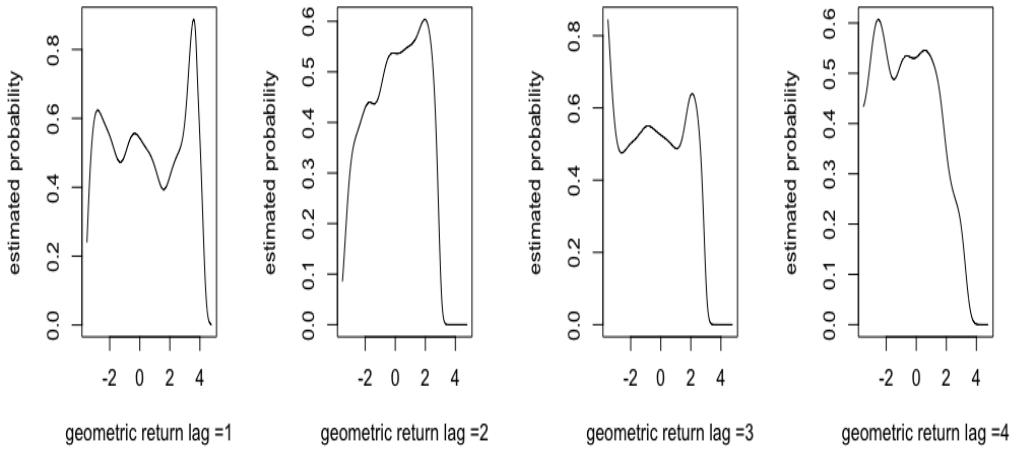

Figure 5: Marginal probability of significant variables in MAMaLoR model

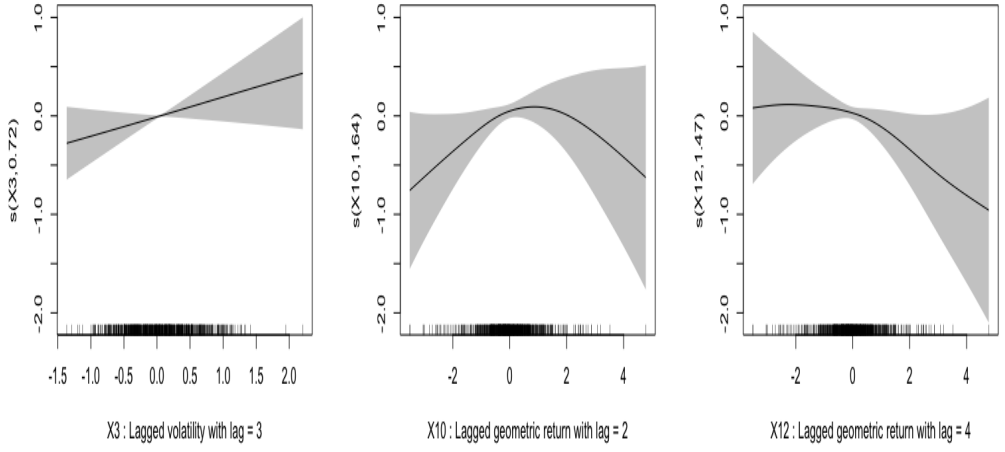

Figure 6: Smooth function for significant variables in GAM model

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

Table 3.3: 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							

curse of dimensionality. We compare the performance of our MAMaLoR model with both GLM and GAM in the forms detailed as follows:

MAMaLoR model:

$$logit(p_t) = \log \frac{p_t}{1-p_t} \approx \alpha_0 + \sum_{j=1}^{4} \alpha_j f_j(v_{t-j}) + \sum_{j=1}^{4} \alpha_{4+j} f_{4+j}(V_{t-j}) + \sum_{j=1}^{4} \alpha_{8+j} f_{8+j}(G_{t-j}), \tag{3.51}$$

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 (3.10) and then α_j's estimated, detailed in Section 3.1;

GLM model:

$$logit(p_t) \approx \alpha_0 + \sum_{j=1}^{4} \alpha_j v_{t-j} + \sum_{j=1}^{4} \alpha_{4+j} V_{t-j} + \sum_{j=1}^{4} \alpha_{8+j} G_{t-j}, \tag{3.52}$$

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

GAM model:

$$logit(p_t) \approx \alpha_0 + \sum_{j=1}^{4} g_j(v_{t-j}) + \sum_{j=1}^{4} g_{4+j}(V_{t-j}) + \sum_{j=1}^{4} g_{8+j}(G_{t-j}), \tag{3.53}$$

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 3.3. 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 (3.53) 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 3.3 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 3.4 with a much smaller AIC value of 968.74 than those in Table 3.3.

We have further compared the AUC values of the forecasting of the market moving direction Y_t based on the significant X_t identified by the above analysis. The group of $X_t = (G_{t-1}, G_{t-3}, G_{t-8}, G_{t-11}, G_{t-13}, G_{t-14}, G_{t-15}, G_{t-16}, G_{t-17}$

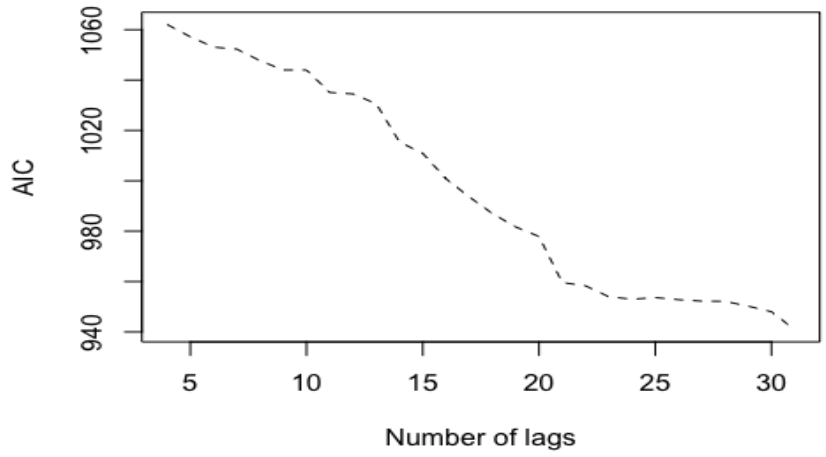

Figure 7: The aic of MAMaLoR model with different number of lagged G_t.

Table 3.4: 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			

, $G_{t-18}, G_{t-21}, G_{t-28}, G_{t-31}$) is identified by the MAMaLoR model given in Table 3.4. The AUC values with the ROC curves for both MAMaLoR model with and without 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

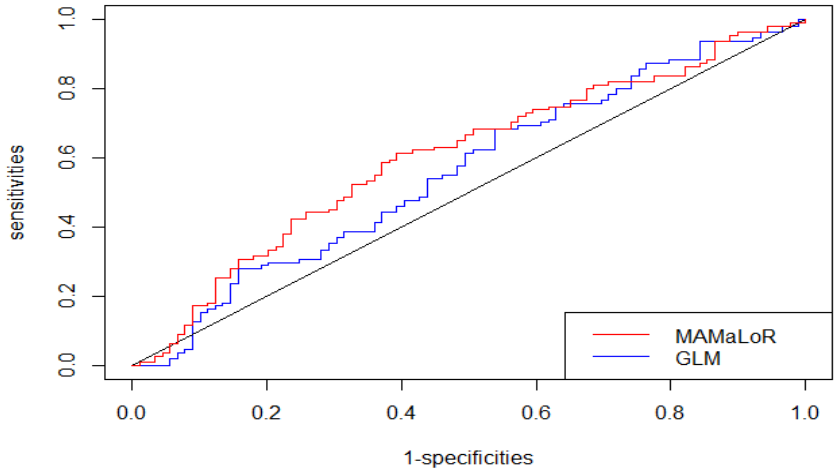

Figure 8: The ROC curves for the MAMaLoR with selected bandwidth (h given in Table 3.5) and the GLM models. Here the corresponding AUC values for MAMaLoR 0.6041 with h selected, and for GLM it is 0.560, respectively.

Table 3.5: Bandwidth selected for the 13 significant variables given in Table 3.4

	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

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 GLM model in time series classification prediction, which is promising. Recognising that the 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 3.4. The selected bandwidth h's are summarised in Table 3.5, 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.

3.5 Conclusion

In this chapter, 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 case.

Chapter 4

Variable Selection in Generalised Model Averaging MArginal Regressions for Discrete-Valued Time Series

To study the discrete-valued time series data of exponential family in high and ultra-high dimensions, we propose the semiparametric model named a 'Generalised Model Averaging MArginal Regression' (GMAMaR) with the variable selection technique called adapted LASSO. The procedure can be viewed as a penalised model averaging method, where each nonparametric and nonlinear marginal regression is estimated in the sense of Kullback-Leibler distance. The asymptotic properties are established under mild conditions for the time series observations that are of β-mixing property. With the computational advantage of low dimensional regression, the GMAMaR model can avoid the 'curse of dimensionality'. The adapted LASSO technique, which can be viewed as a Lagrange penalty, then extracts the important information, and therefore overcomes the problem of overfitting, especially for relatively small size of data. The performance of the procedure are supported by an application to FTSE 100 index data for market direction forecasting compared with traditional Generalised Linear Model (GLM) and popular machine learning technique Random Forest (RF).

4.1 Introduction

With the development of modern technology, high-dimensional discrete-valued time series data has been often faced in various disciplines, e.g., the binary outcome of market performance in financial market and daily increase number of infected patients that follows the poisson or negative-binomial distribution in epidemiology. However the accurate of such estimation tend to deteriorate in high dimension due to the "curse of dimensionality" (Stork et al., 2001). This is due to the computational costs that increase exponentially with the covariate space. To overcome this drawback, Li et al. (2015) have proposed a novel method for estimating unknown form of data and forecasting the unknown future by conditional time series regression, namely the Model Averaging MArginal Regression (MAMAR). Peng and Lu (2021a) later extend the idea to the logistics regression where maximum likelihood method is adopted. In this chapter, we further allow the discrete-value time series to have a distribution in the exponential family that includes the binomial distribution.

However, in many situations of practice, a further challenge for the regression in high-dimensional is to avoid "poor generalisation ability", i.e., overfitting, especially when the data size is relatively small. With noncorrelated variables included, the model will produce additional errors in estimation and it is hard for people to understand the true important variables. The problem is, we don't know what variables to include and what to exclude based on human experience.

The basic idea is to reduce the dimensionality via different tools and get rid of the non-correlated variables. Traditional variable selection methods, such as stepwise selection based on AIC (DeLeeuw, 1992) and BIC (Schwarz et al., 1978) have a number of drawbacks as discussed in Breiman et al. (1996) and Fan and Li (2001). LASSO method (Tibshirani (1996) and Tibshirani (1997)), instead, applies a $\mathfrak{L}_1$-penalty to the conventional logistic regression model, such that the model would shrink in order to force useless variables to be estimated with close-to-zero coefficients, and yields the consistent estimator under mild conditions shown in Zou (2006) and Zhao and Yu (2006). In particular, Zou (2006) has improved the conventional LASSO technique to adaptive LASSO by assigning different weights to different interactions and proved the validity of the "Oracle Property". Thus the collinearity

of parameters does not change the degree of fitting, and the sample size gives no restriction on the number of parameters. As a consequence, when the sample size is rather small, or when it comes to higher dimension, the adapted penalisation method could improve the stability of the estimation. Further state-of-art studies, such as the extension of adaptive LASSO in spatial models, can be referred to Al-Sulami et al. (2019).

To conclude, there are several advantage of using penalisation model:

- When dealing with the interaction of multiple factors, the number of parameters would be very large. However, the penalisation could estimate the coefficients in a better way.
- In the case of which dummy variables are included into the model, the problem of collinearity may arise. Adapted LASSO method, however, overcome this problem and thus can be used to analyse high dimensional data.

The aim of this chapter is thus to apply the adapted LASSO technique to the novel discrete-valued time series model, namely the Generalised Marginal Moving Average Regression (GMAMaR), for data of unknown forms and under the data dependence of a so-called β-mixing condition. The model aims to extract important information of covariates from high and ultra-high data. We will also show in numerical examples that our GMAMaR model with variable selection can work better in forecasting than traditional generalised linear model (GLM)(McCullagh and Nelder, 1989) with logistic regression and also popular machine learning method, namely, the Random Forest (RF)(Liaw and Wiener, 2002).

The remaining of chapter is structured as follows. In Section 2, we first introduce our Generalised Marginal Moving Average Regression (GMAMaR) for the discrete-valued time series that have a distribution of the exponential family. The adapted LASSO penalty is then added to the GMAMaR model as the Lagrange multiplier. The estimation of GMAMaR model is illustrated in Section 3, where the detail of the computational algorithm for the penalised model is also provided. Theoretical results for the asymptotic properties are then given in Section 4. In Section 5, the numerical example of an application of the binomial distribution to forecasting the market price moving direction of FTSE 100 index will be given. Conclusions are summarised in Section 6. All the proofs are relegated to the Appendix.

4.2 Generalised model averaging marginal nonlinear regressions

Consider a β-mixing (stationary) time series process (Y_t, X_t^T) with Y_t the discrete-valued response variable at time t that has a distribution in the exponential family and $X_t = (x_{1t}, ..., x_{dt})^T$ a d-dimensional covariate series representing the available information up to time $t-1$ with possibly also the lagged terms. The dimension d may be rather large in practice.

A well known difficulty, namely the "curse of dimensionality", means that a direct nonparametric or semiparametric estimation may perform very poor in computation when d is large, e.g., $d > 3$. For instance, the computational cost of Generalised Additive Model (GAM)(Hastie and Tibshirani, 1990) is large and it is difficult to converge for small sample size. We therefore suggest the semiparametric procedure, namely Generalised Model Averaging nonlinear Marginal Regressions (GMAMaR), by extending the model averaging idea of Li et al. (2015). The model involves estimating the marginal probability of Y_t given each covariate x_{jt} first, and then combining all the marginal information by model averaging. Due to the cheap computation of one-dimensional nonparametric estimation, the "curse of dimensionality" is thus avoided.

To avoid overfitting, we further apply the adapted LASSO technique (c.f. Zou (2006)) that improves the conventional LASSO method with "oracle property", i.e., asymptotic properties, to our proposed GMAMaR model. By forcing the coefficients of uncorrelated covariates to be zero, the important information is thus extracted. This is introduced at the end of this section.

4.2.1 Semiparametric procedure

The genetic form of density function of exponential family can be given:

$$\mathfrak{m}_Y(Y_t; \boldsymbol{\theta}_t) = exp(Y_t\theta_t - \psi(\theta_t) + \Psi(Y_t, \Theta)), \tag{4.1}$$

where Θ is a known parameter, $\Psi(\cdot)$ and $\psi(\cdot)$ are known functions for a particular distribution family, and θ_t is the canonical parameter depending on the given information in X_t, which can also be expressed by a link function

$\eta(\mu_t)$. Here μ_t is the conditional mean $\mu_t = E(Y_t|X_t)$ that is to be estimated. So the regression problem is to estimate the conditional expectation for the discrete-value time series:

$$\mu_t = E(Y_t|X_t) = \psi'(\theta_t), \tag{4.2}$$

where $\psi'(\cdot)$ stands for the first order derivative of $\psi(\cdot)$. For a specific distribution, $(\psi')^{-1}(\cdot) = \eta(\cdot)$ is a known canonic link function, which is the inverse function of $\psi'(\cdot)$. We then consider the approximation of μ_t with the one-dimensional marginal information as follows:

$$\theta_t = \eta(\mu_t) = (\psi')^{-1}(\mu_t) \approx \alpha_0 + \sum_{j=1}^{d} \alpha_j \eta^{-1}(E(Y_t|x_{jt})) \equiv \theta_t(\boldsymbol{\alpha}), \tag{4.3}$$

where $E(Y_t|x_{jt})$ is the one-dimensional conditional mean to be estimated based on j-th covariate x_{jt}, and $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_d)$ are the unknown coefficients also to be estimated. For the ease of presenting, we defer the detail of estimations later in Section 3. Here $\boldsymbol{\alpha}$ can be seen as the weights assigned to different marginal estimations, and thus this procedure can be viewed as a model averaging, detailed as follows.

The marginal conditional mean based on the j-th component (x_{jt}) can be defined as follows:

$$\mu_{jt} = E(Y_t|x_{jt}),\ j = 1, ..., d. \tag{4.4}$$

The generalised marginal nonparametric regression can be expressed as:

$$\eta(\mu_{jt}) = f_j(x_{jt}), \tag{4.5}$$

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

$$\mu_{jt} = \eta^{-1}(f_j(x_{jt})). \tag{4.6}$$

By combining the one-dimensional marginal regressions, we can re-write (4.3) as follows:

$$\begin{aligned}\theta_t = \eta(\mu_t) &\approx \alpha_0 + \alpha_1\eta(\mu_{1t}) + ... + \alpha_d\eta(\mu_{dt}) \\ &= \alpha_0 + \alpha_1 f_1(x_{1t}) + ... + \alpha_d f_d(x_{dt}) \equiv f_t^{MA} \equiv \theta_t(\boldsymbol{\alpha}),\end{aligned} \tag{4.7}$$

Denote I_{t-1} all the information up to time $t-1$ about time series Y_t. In the MAMAR procedure of Li et al. (2015), they estimate $E(Y_t|I_{t-1})$ by $\alpha_0 + \sum_{j=1}^{d} \alpha_j E(Y_t|x_{jt})$ in terms of $\mathcal{L}_2$ distance, that is $E\{E(Y_t|I_{t-1}) - \alpha_0 - \sum_{j=1}^{d} \alpha_j E(Y_t|x_{jt})\}^2$ is minimised with respect to $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, ..., \alpha_d)$. Here, we estimate the regression (4.7) and (4.3) in terms of the Kullback-Leibler distance (KL-distance), that is a natural distance function from a "true" distribution, to a "target" distribution.

$$KL = 2\sum_{t=1}^{n} E_{y_t}\{\log(\mathfrak{m}(Y_t;\theta_t))\} - \log(\mathfrak{m}(Y_t;\theta_t(\boldsymbol{\alpha}))), \tag{4.8}$$

The KL-distance is minimised with respect to $\boldsymbol{\alpha}$, and we denote this (true) minimiser by $\boldsymbol{\alpha}^{(0)}$. We hence need to estimate the minimiser by maximum likelihood estimation given below:

$$\begin{aligned}\hat{\boldsymbol{\alpha}}^{*(n)} &= \arg\max L(\boldsymbol{\alpha}, \mathbf{f}_0) \\ &= \frac{1}{n}\sum_{t=1}^{n}\left[Y_t(\alpha_0 + \sum_{j=1}^{d}\alpha_j f_j(x_{jt})) - \psi(\alpha_0 + \sum_{j=1}^{d}\alpha_j f_j(x_{jt})) + \Psi(Y_t, \Theta)\right],\end{aligned} \tag{4.9}$$

where $\hat{\boldsymbol{\alpha}}^{*(n)}$ is defined similarly with estimated elements of $\mathbf{f}_0(\cdot) = (f_1(\cdot), \ldots, f_d(\cdot))^T$.

4.2.2 Adapted LASSO

Recall the log-likelihood function (4.9), which is equivalent to

$$L(\boldsymbol{\alpha}, \mathbf{f}_0) = \sum_{t=1}^{n}[\{Y_t(\alpha_0 + \sum_{j=1}^{d}\alpha_j f_j(x_{jt}))\} - \psi(\alpha_0 + \sum_{j=1}^{d}\alpha_j f_j(x_{jt})) + \Psi(Y_t, \Theta)], \tag{4.10}$$

where $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_d)$ is to be optimised in order to find the maximum L. The idea is thus to penalise the coefficients of non-correlated variables to be zero. It is equivalent to multiply the coefficients by a weight factor. To formulate it combined with our original model, we use the *Lagrange Multiplier* method and obtain the estimation $\hat{\boldsymbol{\alpha}}$, viz:

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} -L + \lambda_n \sum_{j=1}^{d} \mathfrak{E}_j |\alpha_j|, \tag{4.11}$$

where the vector $\boldsymbol{\mathfrak{E}} = (\mathfrak{E}_1, \cdots, \mathfrak{E}_d)^T = \frac{1}{|\hat{\boldsymbol{\alpha}}^{*(n)}|^\iota}$, standing for $\mathfrak{E}_j = \frac{1}{|\hat{\alpha}_j^{*(n)}|^\iota}$ component-wisely, for $\iota > 0$. and $\hat{\boldsymbol{\alpha}}^{*(n)} = (\hat{\alpha}_1^{*(n)}, \cdots, \hat{\alpha}_d^{*(n)})^T$ is a $\sqrt{n}$ consistent estimator to the true parameter $\boldsymbol{\alpha}^*$, e.g., an MLE that maximises the log-likelihood L as defined in (4.10) with estimated $f_j(\cdot)$'s. The *Lagrange Multiplier* coefficient λ_n can vary with n.

We then formulate the estimation problem of $\hat{\boldsymbol{\alpha}}$ by minimising the penalised log-likelihood estimation function with estimated $f_j(\cdot)$'s:

$$\begin{aligned}
\hat{\boldsymbol{\alpha}} &= \arg\min \hat{R}(\boldsymbol{\alpha}) \\
&= \sum_{t=1}^{n} [\{-Y_t(\alpha_0 + \sum_{j=1}^{d} \alpha_j \hat{f}_j(x_{jt}))\} + \psi(\alpha_0 + \sum_{j=1}^{d} \alpha_j \hat{f}_j(x_{jt})) - \Psi(Y_t, \Theta)] \\
&+ \lambda_n \sum_{j=1}^{d} \hat{\mathfrak{E}}_j |\alpha_j|,
\end{aligned} \tag{4.12}$$

where $\hat{f}_j(\cdot)$'s are the one-dimensional marginal estimators, which are made by local linear maximum likelihood fitting based on (4.6). Details of the ideas above for estimation are given below.

4.3 Estimation of penalised GMAMaR

In this section, the estimation of the GMAMaR procedure is divided into two stages. In the first stage, we aim to estimate the one-dimensional nonparametric function $f_j(x_{jt})$. This unknown function is allowed to be nonlinear, and estimated based on data given. Then, with estimated $\hat{f}_j(x_{jt})$'s on hand, we can replace $f_j(x_{jt})$'s in (4.7), and treat the semiparametric estimation

problem as a weighted linear regression with links to exponential family. The penalised coefficients, e.g., $\hat{\boldsymbol{\alpha}}$ in (4.12), as well as the Lagrange Multiplier λ_n can be estimated with computational algorithm that requires the estimation of $\hat{f}_j(x_{jt})$ and $\hat{\boldsymbol{\alpha}}^{*(n)}$, which is then introduced at the end of this section.

4.3.1 Estimating $\hat{f}_j(x_{jt})$

As $f_j(x_{jt})$'s are unknown, we need to estimate them first. Here nonparametric method can be used to estimate the marginal probability $\mu_{jt} = E(Y_t|x_{jt})$. In this chapter, we suggest applying maximum likelihood local linear fitting (c.f., Fan et al. (1998a) and Peng and Lu (2021b)) for the estimation of $f_j(\cdot)$ in (4.7) as it is one-dimension, and Y_t given x_{jt} follows exponential family distribution. The conditional local log likelihood is thus given by:

$$\ell_{h,x}(\mu_{jt}; Y_t) = \sum_{t=1}^{n} \log m_{Y_t}(Y_t, \theta_{jt}) k_h(x_{jt} - x_{j0}), \tag{4.13}$$

where $K_h(\cdot) = h^{-1}K(\cdot/h)$ with $K(\cdot)$ is the kernel function on $\mathbb{R}^1$, h is the bandwidth appropriately selected, and $\mu_{jt} = E(Y_t|x_{jt}) = \psi'(\theta_{jt})$ is the expected probability to be estimated.

Note that by taking the Taylor Expansion of $f_j(x_{jt})$ at an arbitrary point x_{j0} given it is differentiable, and knowing x_{jt} that is in the neighbourhood of x_{j0}, we can give its approximation ($f_j(x_{j0})$) as follows:

$$\begin{aligned} f_j(x_{jt}) &\approx f_j(x_{j0}) + f_j'(x_{j0})(x_{jt} - x_{j0}) \\ &\equiv \beta_1 + \beta_2(x_{jt} - x_{j0}), \quad |x_{jt} - x_{j0}| \leq h. \end{aligned} \tag{4.14}$$

The estimation of f_j at x_{j0} as the intercept $\hat{\beta}_1$ in (4.14) is relatively easy as it can be viewed as a weighted linear regression. By letting x_{j0} go through each points in x_{jt}, we thus have the marginal estimation $\hat{f}_j(x_{jt})$.

4.3.2 Estimating $\hat{\boldsymbol{\alpha}}^{*(n)}$

Now we can try to estimate the coefficients $\boldsymbol{\alpha}$ in (4.9) together by replacing the $f_j(x_{jt})$'s with $\hat{f}_j(x_{jt})$'s:

$$\hat{L}(\boldsymbol{\alpha}) = \frac{1}{n}\sum_{t=1}^{n}\left[Y_t(\alpha_0 + \sum_{j=1}^{d}\alpha_j\hat{f}_j(x_{jt})) - \psi(\alpha_0 + \sum_{j=1}^{d}\alpha_j\hat{f}_j(x_{jt})) + \Psi(Y_t,\Theta)\right]. \tag{4.15}$$

To avoid the impacts of the poor estimate of $f_j(\cdot)$'s at extreme values, e.g., where the values of x_{jt} are near the boundaries, we improve the estimation procedure in (4.15) by adding a weight function $w(X_t) = \prod_{j=1}^{d}\mathbf{I}_{(c_{0j}\le x_{jt}\le c_{1j})}$ controlling the edge effects in the estimation with $\mathbf{I}_{(\cdot)}$ being an indicator function with $c_{0j} < c_{1j}$ appropriately chosen:

$$\begin{aligned} L_n(\boldsymbol{\alpha}) = L_n(\boldsymbol{\alpha},\hat{\mathbf{f}}(\cdot)) = \frac{1}{n}\sum_{t=1}^{n}\{Y_t(\alpha_0 + \sum_{j=1}^{d}\alpha_j\hat{f}_j(x_{jt})) \\ - \psi(\alpha_0 + \sum_{j=1}^{d}\alpha_j\hat{f}_j(x_{jt})) + \Psi(Y_t,\Theta)\}w(X_t), \end{aligned} \tag{4.16}$$

with the population (expected) log-likelihood function:

$$\begin{aligned} L(\boldsymbol{\alpha},\mathbf{f}_0(\cdot)) \approx E[Y_t(\alpha_0 + \sum_{j=1}^{d}\alpha_j f_j(x_{jt})) \\ - \psi(\alpha_0 + \sum_{j=1}^{d}\alpha_j f_j(x_{jt})) + \Psi(Y_t,\Theta)]w(X_t), \end{aligned} \tag{4.17}$$

where $\mathbf{f}_0(\cdot) = (f_1(\cdot),\ldots,f_d(\cdot))^T$ and $\hat{\mathbf{f}}(\cdot)$ is defined similarly with estimated elements. 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. These poor estimates of $f_j(x_{jt})$ are thus removed from the estimation of $\boldsymbol{\alpha}$.

From the computational perspective, with $\hat{f}_j(x_{jt})$ known, equation (4.7) can be viewed as a generalised linear regression, which means we can apply relevant technique and algorithm developed. Therefore our GMAMaR procedure is easy to implement in computation.

4.3.3 Estimating λ_n and $\hat{\boldsymbol{\alpha}}$

Recall that the estimation of $\hat{\boldsymbol{\alpha}}$ can be reached by solving the minimization problem of (4.11). Without the penalty term, the objective function of this minimization problem is thus the log-likelihood part of $R(\boldsymbol{\alpha})$, which is equivalently to (4.16):

$$L(\boldsymbol{\alpha}) = \sum_{t=1}^{n}\{-Y_t(\alpha_0+\sum_{j=1}^{d}\alpha_j f_j(x_{jt}))+\psi(\alpha_0+\sum_{j=1}^{d}\alpha_j f_j(x_{jt}))+\Psi(Y_t,\Theta)\}w(X_t). \tag{4.18}$$

We can also write the gradient of the objective function as follows:

$$G(\boldsymbol{\alpha}) = \sum_{t=1}^{n}[Y_t - \psi'(\alpha_0 + \sum_{j=1}^{d}\alpha_j f_j(x_{jt}))]\widetilde{\chi}_t(\mathbf{f}_0)w(X_t), \tag{4.19}$$

where $\widetilde{\chi}_t(\mathbf{f}_0) = (1, f_1(x_{1t}), \ldots, f_d(x_{dt}))$.

Notice that both the log-likelihood function and its gradient function are hard to be solved in a close form. Thus we will use the computer software to tackle the problem more efficiently. Here, a computational procedure is developed to find out the best estimation that maximize the likelihood.

(The algorithm for the GMAMaR model with adapted LASSO)

1. Solve the GMAMaR model to get the initial estimation set: $\hat{\boldsymbol{\alpha}}^{*(n)}$
2. Compute the weight of adaptive LASSO: $\hat{\mathfrak{E}}_j = \frac{1}{|\hat{\boldsymbol{\alpha}}_j^{*(n)}|^{\iota}}$, here $\iota > 0$ can be chosen as 1 for simplicity;
3. Define $\widetilde{f}_j(x_{jt}) = f_j(x_{jt})/\hat{\mathfrak{E}}_j$, for $j = 1, ..., d$;
4. Solve the LASSO model for all λ_n that are considered by tackling the following minimisation problem:

$$\hat{\boldsymbol{\alpha}}^{**}(\lambda_n) = \arg\min_{\boldsymbol{\alpha}} \sum_{t=1}^{n} \{[-Y_t(\alpha_0 + \sum_{j=1}^{d} \alpha_j \widetilde{f}_j(x_{jt}))] + \psi(\alpha_0 + \sum_{j=1}^{d} \alpha_j \widetilde{f}_j(x_{jt})) - \Psi(Y_t, \Theta)\} w(X_t) + \lambda_n \sum_{j=1}^{d} \hat{\mathfrak{E}}_j |\alpha_j|, \tag{4.20}$$

where we use the **lbfgs** package in R (Coppola and Stewart, 2014), which would handle the adaptive LASSO problem by treating it as an optimization problem of the log likelihood function (4.18) plus the $\mathfrak{L}_1$ norm penalisation;

5. Compute the adaptive LASSO estimation: $\hat{\boldsymbol{\alpha}}^* = \hat{\boldsymbol{\alpha}}^{**}/\hat{\mathfrak{E}}_j$, for $j = 1, ..., d$;
6. Define the best estimation $\hat{\boldsymbol{\alpha}}^*(\lambda_n^*)$ and choose the best penalisation coefficient λ_n^* by finding the minimum BIC value:

$$BIC(\lambda_n) = \sum_{t=1}^{n} \{[Y_t(\alpha_0 + \sum_{j=1}^{d} \alpha_j f_j(x_{jt}))] - \psi(\alpha_0 + \sum_{j=1}^{d} \alpha_j f_j(x_{jt})) + \Psi(Y_t, \Theta)\} w(X_t) + klog(\lambda_n)/\lambda_n, \tag{4.21}$$

where $k \neq 0$ is the number of parameters estimated by the model;

7. Output $\hat{\boldsymbol{\alpha}}^*(\lambda_n)$ and λ_n^*.
8. In a more general setting, we can also tune the parameter ι and repeating step $3 - 7$ to find out the best pair $(\hat{\boldsymbol{\alpha}}^*, \lambda_n^*, \iota^*)$.

4.4 Asymptotic properties

In this section, we are going to present the asymptotic properties, namely the uniform consistency and asymptotic normality, of the proposed penalised GMAMaR procedure.

Since the non-correlated estimations are penalised to (near) zero in our estimation, with $\boldsymbol{\alpha}^*$ denoting d-dimensional vector of true estimations, we

denote $\boldsymbol{\alpha}^{*1}$ to be d_0-dimensional vector of non-zero true parameters and $\boldsymbol{\alpha}^{*2}$ to be $(d-d_0)$-dimensional vector of zero true parameters.

$$\boldsymbol{\alpha}^* = \begin{bmatrix} \boldsymbol{\alpha}^{*1} \\ \boldsymbol{\alpha}^{*2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\alpha}^{*1} \\ \mathbf{0} \end{bmatrix}.$$

Consider the true model with sparse representation. Let $\mathcal{A} = \{j : \alpha_j^* \neq 0\} = \{1, ..., d_0\}$, and $d_0 < d$. We can write the Fisher information matrix:

$$\boldsymbol{I} = \begin{bmatrix} \boldsymbol{I}_{11} & \boldsymbol{I}_{12} \\ \boldsymbol{I}_{21} & \boldsymbol{I}_{22} \end{bmatrix}.$$

where $\boldsymbol{I}_{11}$ is a $d_0 \times d_0$ matrix. Then $\boldsymbol{l}_{11}$ is the Fisher information with the true sub-model known. We would like to show that under mild conditions as listed in the following assumption 4.1, the adaptive LASSO estimation of GMAMaR model satisfies the following theorems.

Assumption 4.1. *1. (i) We assume (Y_t, X_t) (Y_t has a distribution in the exponential family) is β-mixing with the mixing coefficient $\beta(t) = O(t^{-b})$ for some $b > max(2(\rho r+1)/(\rho r-2), (r+a)/(1-2/\rho))$ with $a \geq (r\rho-2)r/(2+r\rho-4r)$; (ii) the joint probability density function $g_{X_{t_1},\cdots,X_{t_s}}(x_1, \cdots, x_s)$ is bounded uniformly for any $t_0 < t_1 < \cdots < t_s$ and $0 \leq s \leq 2(r-1)$; (iii) $E|X_t|^{\rho r} < \infty$ for some real number $\rho > 4-2/r$, where $r \geq 1$ is some positive integer.*

2. *(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 $n \to \infty$.*

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

4. *We define $\mathbf{f}_0(\cdot) = (f_1(\cdot), ..., f_d(\cdot))^T$ the vector of the true conditional regression functions, $f_j(\cdot)$'s, which have continuous and bounded second order derivatives.*

5. *We assume under the true parameter $\boldsymbol{\alpha}^*$, $E[\frac{\partial L(\boldsymbol{\alpha}^*;\mathbf{f}_0)}{\partial \alpha_j}] = 0$, where $L(\boldsymbol{\alpha};\mathbf{f})$ is defined in (4.9). The Fisher information matrix at $\boldsymbol{\alpha} = \boldsymbol{\alpha}^*$, $I(\boldsymbol{\alpha}^*) = E[\frac{\partial L(\boldsymbol{\alpha}^*;\mathbf{f}_0)}{\partial \boldsymbol{\alpha}}][\frac{\partial L(\boldsymbol{\alpha}^*;\mathbf{f}_0)}{\partial \boldsymbol{\alpha}}]^T$, is finite and positive definite.*

6. *There is an sufficient large enough open subset* $\mho$ *that contains* $\boldsymbol{\alpha}^*$*(true parameter) ,such that* $\forall \boldsymbol{\alpha} \in \mho$*, there exists a finite function* $\psi_{jks} = E_{\boldsymbol{\alpha}^*}[\mathfrak{S}_{jks}(X)] < \infty$*:*

$$|\frac{\partial L(X,\alpha)}{\partial\alpha_j\partial\alpha_k\partial\alpha_s}| \leq \mathfrak{S}_{jks}(X), \tag{4.22}$$

where this $\mathfrak{S}$ *is the upper bound uniformly with respect to* α*.*

Remark. Note that Assumption 1 gives the weak dependency of time series, which is β-mixing (Fan and Yao, 2003), (Lu et al., 2007). Assumption 2 is standard in time series topics (Fan et al., 2003), (Lu et al., 2007). The edge effect is controlled by Assumption 3, which removes the extreme estimates around the boundaries of X_t, in order to improve the practical performance of the estimation (c.f. Fan et al. (1998b), Fan et al. (2003) and Lu et al. (2007)). Assumption 4 gives the smoothness conditions on the conditional density and regression functions. Assumption 5 and Assumption 6 are often adopted in conventional models to guarantee asymptotic normality of the maximum likelihood estimates (Fan and Li, 2001).

Theorem 4.2. *Let above Assumptions hold. Suppose* $\frac{\lambda_n}{\sqrt{n}} \to 0$ *and* $\lambda_n n^{(\iota-1)/2} \to \infty$*. Then there exist a global minimizer* $\hat{\alpha}$ *of the objective function* $\hat{R}(\boldsymbol{\alpha})$ *defined in (4.12) such that* $\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\| = O_p(\frac{1}{\sqrt{n}})$*, where* $\boldsymbol{\alpha}^*$ *is the parameter under true model.*

Proof. Recall $\mathcal{A} = \{j : \alpha_j^* \neq 0\} = \{1, ..., d_0\}$ (say, without loss of generality), and $d_0 < d$. Note that $\mathfrak{E}_j = \frac{1}{|\hat{\alpha}_j^*|^\iota}$ with $\hat{\alpha}_j^* \equiv \hat{\alpha}_j^{*(n)}$ the pre-estimator root-n consistent to α_j. Let $c_n = \frac{1}{\sqrt{n}}$. Recall $L(\cdot;\cdot)$ is log likelihood function defined in (4.10) and $\hat{R}(\cdot)$ is penalised likelihood function defined in (4.12). Then if the global minimum of $\hat{R}(\cdot)$, $\hat{\boldsymbol{\alpha}}$, satisfies $\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\| = O_p(\frac{1}{\sqrt{n}})$, it equivalently belongs, with probability tending to 1, in the ball $\mathfrak{A} = \{\alpha^* + c_n\boldsymbol{\delta} : \|\boldsymbol{\delta}\| \leq \xi\}$, which is centred around α^* with radius $c_n\boldsymbol{\delta}$. Here $\|.\|$ is the $\mathfrak{L}_2$ Euclidean norm, and ξ is a large constant.

First note by the uniform consistency of the $\hat{f}_j(\cdot)$ to $f_j(\cdot)$ over any compact set (c.f., Theorem 2 of Chapter 2 (Peng and Lu, 2021b)), we can re-write the objective function $\hat{R}(\cdot)$ in (4.12) that satisfies $\hat{R}(\boldsymbol{\alpha}) = (1 + o_P(1))R_n(\boldsymbol{\alpha})$ uniformly w.r.t. $\boldsymbol{\alpha} \in \mathfrak{A}_0 = \{\alpha^* + c_1\boldsymbol{\delta} : \|\boldsymbol{\delta}\| \leq \xi\}$ (c.f., Lu et al. 2007) with

R_n defined below. Now we can examine R_n as follows.

$$\begin{aligned} R_n(\boldsymbol{\alpha}^* + c_n\boldsymbol{\delta}) = &\sum_{t=1}^{n}\{-Y_t(\alpha_0^* + \sum_{j=1}^{d}(\alpha_j^* + c_n\boldsymbol{\delta})f_j(x_{jt}))\}w(X_t) \\ &+ \sum_{t=1}^{n}\{\psi(\alpha_0^* + \sum_{j=1}^{d}(\alpha_j^* + c_n\boldsymbol{\delta})f_j(x_{jt})) + \Psi(Y_t,\Theta)\}w(X_t) \\ &+ \lambda_n\sum_{j=1}^{d}\mathfrak{E}_j|\alpha_j^* + c_n\boldsymbol{\delta}|, \end{aligned} \tag{4.23}$$

and

$$\begin{aligned} R_n(\boldsymbol{\alpha}^*) = &\sum_{t=1}^{n}\{-Y_t(\alpha_0^* + \sum_{j=1}^{d}(\alpha_j^*)f_j(x_{jt}))\}w(X_t) \\ &+ \sum_{t=1}^{n}\{\psi(\alpha_0^* + \sum_{j=1}^{d}(\alpha_j^*)f_j(x_{jt})) + \Psi(Y_t,\Theta)\}w(X_t) + \lambda_n\sum_{j=1}^{d}\mathfrak{E}_j|\alpha_j^*|. \end{aligned} \tag{4.24}$$

Denote $M_n(\boldsymbol{\delta}) = R_n(\boldsymbol{\alpha}^* + c_n\boldsymbol{\delta}) - R_n(\boldsymbol{\alpha}^*)$. By Tylor's expansion, we have:

$$\begin{aligned} M_n(\boldsymbol{\delta}) &= R_n(\boldsymbol{\alpha}^* + c_n\boldsymbol{\delta}) - R_n(\boldsymbol{\alpha}^*) \\ &= -\left[c_n[\frac{\partial L(\boldsymbol{\alpha}^*;\mathbf{f}_0)}{\partial\boldsymbol{\alpha}}]^T\boldsymbol{\delta} + \frac{1}{2}\boldsymbol{\delta}^T\frac{\partial^2 L(\boldsymbol{\alpha}^*;\mathbf{f}_0)}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^T}\boldsymbol{\delta}c_n^2\{1+o_p(1)\}\right] \\ &+ \lambda_n\sum_{j=1}^{d}\mathfrak{E}_j\{|\alpha_j^* + c_n\boldsymbol{\delta}| - |\alpha_j^*|\}, \end{aligned} \tag{4.25}$$

where $L(\boldsymbol{\alpha}^*;\mathbf{f}_0)$ is the true log likelihood function (4.10).

For simplicity, we denote:

$$M_n(\boldsymbol{\delta}) = A_1 + A_2 + A_3, \tag{4.26}$$

with

$$A_1 = -\sum_{t=1}^{n}[(Y_t - \psi'(\mathbf{f}_0,\boldsymbol{\alpha}^*))]c_n\widetilde{\chi}_t(\mathbf{f}_0)^T\boldsymbol{\delta}w(X_t), \tag{4.27}$$

$$A_2 = \sum_{t=1}^{n}\frac{1}{2}[\psi''(\mathbf{f}_0,\boldsymbol{\alpha}^*)\boldsymbol{\delta}^T\widetilde{\chi}_t(\mathbf{f}_0)\widetilde{\chi}_t(\mathbf{f}_0)^T\boldsymbol{\delta}]c_n^2w(X_t)\{1+o_p(1)\}, \tag{4.28}$$

$$A_3 = \lambda_n c_n \sum_{j=1}^{d} \mathfrak{E}_j c_n^{-1}(|\boldsymbol{\alpha}^* + c_n\boldsymbol{\delta}| - |\boldsymbol{\alpha}^*|), \tag{4.29}$$

where $\psi'(\mathbf{f}_0, \boldsymbol{\alpha}^*) = \psi'(\alpha_0^* + \sum_{j=1}^d \alpha_j^* f_j(x_{jt}))$ is the first order derivative of $\psi(\alpha_0^* + \sum_{j=1}^d \alpha_j^* f_j(x_{jt}))$, $\psi''(\mathbf{f}_0, \boldsymbol{\alpha}^*) = \psi''(\alpha_0^* + \sum_{j=1}^d \alpha_j^* f_j(x_{jt}))$ is the second order derivative of $\psi(\alpha_0^* + \sum_{j=1}^d \alpha_j^* f_j(x_{jt}))$ and $\widetilde{\chi}_t(\mathbf{f}_0) = (1, f_1(x_{1t}), \ldots, f_d(x_{dt}))$.

Now we present the asymptotic limit for each term stated above.

Recall the objective function $R_n(\cdot)$ in (4.12) and α^* is the true parameter that minimises the objective function $EL(\boldsymbol{\alpha}; \mathbf{f}_0)$. We then observe its first order derivative A_1, viz:

$$E[(Y_t - \psi'(\mathbf{f}_0, \boldsymbol{\alpha}^*))\widetilde{\chi}_t(\mathbf{f}_0)^T\boldsymbol{\delta}]w(X_t) = 0,$$

and

$$Var[(Y_t - \psi'(\mathbf{f}_0, \boldsymbol{\alpha}^*))\widetilde{\chi}_t(\mathbf{f}_0)^T\boldsymbol{\delta}w(X_t)] = E[\psi''(\mathbf{f}_0, \boldsymbol{\alpha}^*)\boldsymbol{\delta}^T\widetilde{\chi}_t(\mathbf{f}_0)\widetilde{\chi}_t(\mathbf{f}_0)^T\boldsymbol{\delta}w(X_t)]$$

$$= \boldsymbol{V}(\boldsymbol{\alpha}^*).$$

By applying central limit theorem, we have $\frac{1}{\sqrt{n}}\sum_{i=1}^n (Y_t - \psi'(\mathbf{f}_0, \boldsymbol{\alpha}^*))\widetilde{\chi}_t(\mathbf{f}_0)^T w(X_t) = O_p(1)$, and hence $A_1 = A_1/(c_n\sqrt{n}) \xrightarrow{\mathrm{d}} -\boldsymbol{\delta}^T N(0, \boldsymbol{V}(\boldsymbol{\alpha}^*))$. We then have:

$$A_1 = O_p(\sqrt{n}c_n\xi) = O_p(nc_n^2\xi). \tag{4.30}$$

For A_2, similarly we have:

$$E[\psi''(\mathbf{f}_0, \boldsymbol{\alpha}^*)\widetilde{\chi}_t(\mathbf{f}_0)\widetilde{\chi}_t(\mathbf{f}_0)^T]w(X_t) = \boldsymbol{I}(\boldsymbol{\alpha}^*),$$

$$A_2 = \frac{1}{nc_n^2}A_2 \xrightarrow{\mathrm{p}} \frac{1}{2}\boldsymbol{\delta}^T\boldsymbol{I}(\boldsymbol{\alpha}^*)\boldsymbol{\delta},$$

and

$$A_2 = O_p(nc_n^2\xi^2). \tag{4.31}$$

For A_3, if $\alpha_j^* \neq 0$ for $j \in \mathcal{A}$, then $\mathfrak{E}_j = \frac{1}{|\hat{\alpha}_j^*|^\iota} \to \frac{1}{|\alpha_j^*|^\iota}$ and $c_n^{-1}(|\alpha_j^*+c_n\delta_j|-|\alpha_j^*|) \to \delta_j \mathrm{sgn}(\alpha_j^*)$. Thus we have:

$$\lambda_n c_n \sum_{j\in\mathcal{A}} \mathfrak{E}_j c_n^{-1}(|\alpha_j^*+c_n\delta_j|-|\alpha_j^*|) = \lambda_n c_n \sum_{j\in\mathcal{A}} \delta_j \mathrm{sgn}(\alpha_j^*) O_P(1) = O_P(\lambda_n c_n \xi) \xrightarrow{\mathrm{p}} 0$$

for any $\|\delta\| \leq \xi$ because $\lambda_n c_n = O(\lambda_n/\sqrt{n}) \to 0$ by the assumption of this theorem.

If $\alpha_j^* = 0$ for $j \in \mathcal{A}^c$, then $c_n^{-1}(|\alpha_j^* + c_n\delta_j|-|\alpha_j^*|) = |\delta_j|$ and $|c_n^{-1}\hat{\alpha}_j^*| = O_p(1)$ by the root-n consistency of the pre-estimator $\hat{\alpha}_j^*$. Hence

$$\lambda_n c_n \mathfrak{E}_j c_n^{-1}(|\alpha_j^*+c_n\delta_j|-|\alpha_j^*|) = \lambda_n c_n \mathfrak{E}_j |\delta_j| = \lambda_n c_n c_n^{-\iota}(|c_n^{-1}\hat{\alpha}_j^*|)^{-\iota}|\delta_j| = \lambda_n c_n c_n^{-\iota}|\delta_j| O_P(1),$$

which tends to $+\infty$ in probability if $\delta_j \neq 0$ for some $j \in \mathcal{A}^c$, and is equal to zero otherwise, because $\lambda_n c_n c_n^{-\iota} = O(\lambda_n n^{(\iota-1)/2}) \to \infty$ by the assumption of this theorem.

Hence by choosing a sufficiently large ξ, we have:

$$A_3 = \lambda_n c_n \left(\sum_{j\in\mathcal{A}} \mathfrak{E}_j c_n^{-1}(|\alpha_j^* + c_n\delta_j|-|\alpha_j^*|) + \sum_{j\in\mathcal{A}^c} \mathfrak{E}_j c_n^{-1}(|\alpha_j^* + c_n\delta_j|-|\alpha_j^*|) \right), \tag{4.32}$$

which tends to $+\infty$ in probability if $\delta_j \neq 0$ for some $j \in \mathcal{A}^c$, and zero otherwise.

Thus, from (4.30), (4.31) and (4.32), $M_n() = R_n(\boldsymbol{\alpha}^* + c_n\boldsymbol{\delta}) - R_n(\boldsymbol{\alpha}^*) = A_1 + A_2 + A_3$, which tends in distribution to $M(\boldsymbol{\delta}) \equiv -\boldsymbol{\delta}^T N(0, \boldsymbol{V}(\boldsymbol{\alpha}^*)) + \frac{1}{2}\boldsymbol{\delta}^T \boldsymbol{I}(\boldsymbol{\alpha}^*)\boldsymbol{\delta}$ and is of the order $O_p(nc_n^2\xi^2) = O_p(\xi^2)$, for any $\|\delta\| \leq \xi$ with all $\delta_j = 0$ for $j \in \mathcal{A}^c$, and tends to $+\infty$ for any $\|\delta\| \leq \xi$ with some $\delta_j \neq 0$ for $j \in \mathcal{A}^c$.

Note that $\hat{\boldsymbol{\alpha}}$ minimises $\hat{R}(\boldsymbol{\alpha}) = R_n(\boldsymbol{\alpha})(1 + o_P(1))$. If $\hat{\boldsymbol{\alpha}}$ is not within $\{\alpha^* + c_n\delta : \|\delta\| \leq \xi\}$, that is $\hat{\alpha}$ is in $\{\alpha^* + c_n\delta : \|\delta\| \geq \xi\}$, then, owing to convexity of $R_n(\boldsymbol{\alpha}), \hat{\boldsymbol{\alpha}}$ must be on $\{\boldsymbol{\alpha}^* + c_n\boldsymbol{\delta} : \|\delta\| = \xi\}$, with probability tending to one.

When $\|\delta\| = \xi$ holds, by noticing that $M_n(\boldsymbol{\delta}) = A_1 + A_2 + A_3$ and $A_2 \geq 0$ is the largest term (it is bounded by ξ^2, while A_1 is bounded by ξ and ξ is a large constant), we have $\inf_{\|\delta\|=\xi} R_n(\alpha^* + c_n\boldsymbol{\delta}) \geq R_n(\boldsymbol{\alpha}^*)$.

Thus,

$$P(\sqrt{n}|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*| > \xi) = P(\hat{\boldsymbol{\alpha}} \notin \mathfrak{A}) \leq P(\inf_{\boldsymbol{\delta} \in \mathfrak{U}} R_n(\boldsymbol{\alpha}^* + c_n\boldsymbol{\delta}) \leq R_n(\boldsymbol{\alpha}^*)) \to 0,$$

and this completes the proof.

□

Now we would like to introduce the following theorem of the consistency of estimation.

Theorem 4.3. *(Consistency) Let above Assumptions hold. Suppose* $\frac{\lambda_n}{\sqrt{n}} \to 0$*, pick a* $\iota > 0$ *and* $\lambda_n n^{(\iota-1)/2} \to \infty$*. Let* $\boldsymbol{\alpha}^1$ *be* d_0*-dimensional non-zero vector of all* α_j^1 *such that* $j \in A$*. and* $\boldsymbol{\alpha}^2$ *be* $d - d_0$*-dimensional zero vector of all* α_j^2 *such that* $j \in A^c$ *Then the non-correlated variables are estimated to zero with probability tending to one:*

$$P(\hat{\boldsymbol{\alpha}}^2 = \boldsymbol{\alpha}^{*2} = 0) \to 1 \quad as \quad n \to \infty.$$

Proof. Recall that $A = \{j : \alpha_j^* \neq 0\}$. Now we define $\hat{A} = \{j : \hat{\alpha}_j \neq 0\}$. If we have $\forall j \in A, P(j \in \hat{A}) \to 1$, then it suffices to show that $\forall j' \in A^c, P(j' \in \hat{A}^c) \to 0$.

Denote by $\mathfrak{E}^2$ the part of $E = (\mathfrak{E}_1, \cdots, \mathfrak{E}_d)^T$ corresponding to $\mathfrak{E}_j$'s with $j \in A^c$, where A^c stands for the complement of A, and componentwise operations apply where easily seen. We now consider $j' \in \hat{A}$ by taking the derivative of $\hat{R}(\boldsymbol{\alpha})$ with respect to $\boldsymbol{\alpha}^2$ and use the Taylor expansion, viz:

$$\begin{aligned}
\frac{\partial \hat{R}(\hat{\boldsymbol{\alpha}})}{\partial \boldsymbol{\alpha}^2} &= -\frac{\partial L(\hat{\boldsymbol{\alpha}})}{\partial \boldsymbol{\alpha}^2} + \lambda_n \mathfrak{E}^2 \mathrm{sgn}(\hat{\boldsymbol{\alpha}}^2) \\
&= -\frac{\partial L(\boldsymbol{\alpha}^*)}{\partial \boldsymbol{\alpha}^2} - \frac{\partial^2 L(\boldsymbol{\alpha}^*)}{\partial \boldsymbol{\alpha}^2 \partial \boldsymbol{\alpha}^{2T}}(\hat{\boldsymbol{\alpha}}^2 - \boldsymbol{\alpha}^{*2})\{1 + o_p(1)\} + \lambda_n \mathfrak{E}^2 \mathrm{sgn}(\hat{\alpha}^2) \\
&= B_1 + B_2 + B_3, \qquad (4.33)
\end{aligned}$$

with

$$\frac{1}{\sqrt{n}} B_1 = -\sum_{t=1}^{n} [(Y_t - \psi'(\mathbf{f}_{t,A^c}, \hat{\boldsymbol{\alpha}}^2)] \widetilde{\chi}_{t,A^c}(\mathbf{f}_0) w(X_t) \frac{1}{\sqrt{n}} \{1 + o_p(1)\}, \qquad (4.34)$$

where $\mathbf{f}_{t,A^c}$ and $\widetilde{\chi}_{t,A^c}(\mathbf{f}_0)$ stand for the components of $\mathbf{f}_t$ and $\widetilde{\chi}_t(\mathbf{f}_0)$ corresponding to index $j \in A^c$,

$$\frac{1}{\sqrt{n}}B_2 = \frac{1}{n}\sum_{t=1}^{n}[\psi''(\mathbf{f}_{t,A^c}, \hat{\boldsymbol{\alpha}}^2)\widetilde{\chi}_{t,A^c}(\mathbf{f}_0)\widetilde{\chi}_{t,A^c}(\mathbf{f}_0)^T]w(X_t) \times \sqrt{n}(\boldsymbol{\alpha}^{*2} - \hat{\boldsymbol{\alpha}}^2)\{1 + o_p(1)\}, \tag{4.35}$$

$$\frac{1}{\sqrt{n}}B_3 = \frac{1}{\sqrt{n}}\lambda_n \mathfrak{E}^2 \mathrm{sgn}(\hat{\boldsymbol{\alpha}}^2). \tag{4.36}$$

By the law of large number and central limit theorem, $\frac{1}{\sqrt{n}}B_1 \xrightarrow{\mathrm{d}} N(0, \boldsymbol{I}_{22})$ and $\frac{1}{n}\sum_{t=1}^{n}\psi''(\mathbf{f}_{t,A^c}, \hat{\boldsymbol{\alpha}}^2)\widetilde{\chi}_t(\mathbf{f}_{t,A^c})\widetilde{\chi}_t(\mathbf{f}_{t,A^c})^T \xrightarrow{\mathrm{P}} \boldsymbol{I}_{22}$. Hence B_1 and B_2 are of order $O_p(\sqrt{n})$.

B_3 term is of order $n^{\frac{1}{2}}O_p(a_n^*)$, where, under adaptive weights of $\mathfrak{E}$, $a_n^* = \lambda_n \mathfrak{E}^2 \mathrm{sgn}(\hat{\boldsymbol{\alpha}}^2)/\sqrt{n} = \lambda_n n^{(\iota-1)/2}\mathrm{sgn}(\hat{\boldsymbol{\alpha}}^2)$, which tends to zero or ∞ component-wisely, depending on $\hat{\boldsymbol{\alpha}}^2 = 0$ or not component-wisely, by the assumption of this theorem.

Hence $\frac{\partial \hat{R}(\hat{\boldsymbol{\alpha}})}{\partial \boldsymbol{\alpha}^2} = B_1 + B_2 + B_3$, for which $\frac{1}{\sqrt{n}}\frac{\partial \hat{R}(\hat{\boldsymbol{\alpha}})}{\partial \boldsymbol{\alpha}^2}$ should be finite in probability (as $\hat{\boldsymbol{\alpha}}$ is the minimiser of $\hat{R}(\boldsymbol{\alpha})$), is determined by the sign of $\hat{\boldsymbol{\alpha}}^2$. So we have $P(\hat{\boldsymbol{\alpha}}^2 = \boldsymbol{\alpha}^{*2} = 0) \to 1$ as $n \to \infty$.

□

Theorem 4.4. *(Asymptotic Normality) Let above Assumption hold. Suppose* $\frac{\lambda_n}{\sqrt{n}} \to 0$ *and* $\lambda_n n^{(\iota-1)/2} \to \infty$. *Then* $\sqrt{n}(\hat{\boldsymbol{\alpha}}^1 - \boldsymbol{\alpha}^{*1}) \xrightarrow{p} N(0, \boldsymbol{I}_{11}^{-1})$.

Proof. Recall that $A = \{j : \alpha_j^* \neq 0\}$. There exists a global minimiser of the objective function $\hat{R}(\boldsymbol{\alpha})$, which is as same as the minimiser of $\hat{R}(\boldsymbol{\alpha}^1)$, viz:

$$\frac{\partial \hat{R}(\boldsymbol{\alpha}^1)}{\partial \boldsymbol{\alpha}^1}\Big|_{\{\boldsymbol{\alpha}^1 = \hat{\boldsymbol{\alpha}}^1\}} = \mathbf{0}.$$

By taking Taylor expansion, we have:

$$\frac{\partial \hat{R}(\hat{\boldsymbol{\alpha}})}{\partial \boldsymbol{\alpha}^1} = C_1 + C_2 + C_3, \tag{4.37}$$

with

$$\begin{aligned}\frac{1}{\sqrt{n}}C_1 &= -\sum_{t=1}^{n}[(Y_t - \psi'(\mathbf{f}_{t,A}, \hat{\boldsymbol{\alpha}}^1)]\widetilde{\chi}_{t,A}(\mathbf{f}_0)w(X_t)\frac{1}{\sqrt{n}}\{1+o_p(1)\} \\ &= -\sum_{t=1}^{n}[(Y_t - \psi'(\mathbf{f}_{t,A}, \boldsymbol{\alpha}^{*1})]\widetilde{\chi}_{t,A}(\mathbf{f}_0)w(X_t)\frac{1}{\sqrt{n}}\{1+o_p(1)\}, \qquad (4.38)\end{aligned}$$

where $\mathbf{f}_{t,A}$ and $\widetilde{\chi}_{t,A}(\mathbf{f}_0)$ stand for the components of $\mathbf{f}_t$ and $\widetilde{\chi}_t(\mathbf{f}_0)$ corresponding to index $j \in A$,

$$\begin{aligned}\frac{1}{\sqrt{n}}C_2 &= \frac{1}{n}\sum_{t=1}^{n}[\psi''(\mathbf{f}_{t,A}, \hat{\boldsymbol{\alpha}}^1)\widetilde{\chi}_t(\mathbf{f}_0)\widetilde{\chi}_t(\mathbf{f}_0)^T]w(X_t) \\ &\quad \times \sqrt{n}(\boldsymbol{\alpha}^{*1} - \hat{\boldsymbol{\alpha}}^1)\{1+o_p(1)\} \\ &= \frac{1}{n}\sum_{t=1}^{n}[\psi''(\mathbf{f}_{t,A}, \boldsymbol{\alpha}^{*1})\widetilde{\chi}_t(\mathbf{f}_0)\widetilde{\chi}_t(\mathbf{f}_0)^T]w(X_t) \\ &\quad \times \sqrt{n}(\boldsymbol{\alpha}^{*1} - \hat{\boldsymbol{\alpha}}^1)\{1+o_p(1)\}, \qquad (4.39)\end{aligned}$$

$$\frac{1}{\sqrt{n}}C_3 = \frac{1}{\sqrt{n}}\lambda_n \boldsymbol{\mathfrak{E}}^1 \mathrm{sgn}(\hat{\boldsymbol{\alpha}}^1). \qquad (4.40)$$

For C_1,

$$E[(Y_t - \psi'(\mathbf{f}_{t,A}, \boldsymbol{\alpha}^{*1}))\widetilde{\chi}_{t,A}(\mathbf{f}_0)w(X_t)] = 0,$$

and

$$\begin{aligned}&Var[(Y_t - \psi'(\mathbf{f}_{t,A}, \boldsymbol{\alpha}^{*1}))\widetilde{\chi}_{t,A}(\mathbf{f}_0)w(X_t)] \\ &= E[\psi''(\mathbf{f}_{t,A}, \boldsymbol{\alpha}^{*1})\widetilde{\chi}_{t,A}(\mathbf{f}_0)\widetilde{\chi}_{t,A}(\mathbf{f}_0)^T w(X_t)] =: \boldsymbol{I}_{11}(\boldsymbol{\alpha}^{*1}).\end{aligned}$$

For C_2, similarly we have:

$$\frac{1}{n}\sum_{t=1}^{n}[\psi''(\mathbf{f}_{t,A}, \boldsymbol{\alpha}^{*1})\widetilde{\chi}_{t,A}(\mathbf{f}_0)\widetilde{\chi}_{t,A}(\mathbf{f}_0)^T w(X_t)] \xrightarrow{\mathrm{p}} \boldsymbol{I}_{11}(\boldsymbol{\alpha}^{*1}).$$

For C_3, when $n \to \infty$, $\lambda_n \boldsymbol{\mathfrak{E}}^1 \mathrm{sgn}(\hat{\boldsymbol{\alpha}}^1) = \lambda_n \boldsymbol{\mathfrak{E}}^1 \mathrm{sgn}(\boldsymbol{\alpha}^{*1})\{1+o_p(1)\}$. Also we know that $\frac{1}{\sqrt{n}}C_3 = O_P(\lambda_n/\sqrt{n}) \to 0$ by the assumption of this theorem.

Therefore:

$$
\begin{aligned}
0 = & -\frac{1}{\sqrt{n}} \sum_{t=1}^{n} [(Y_t - \psi'(\mathbf{f}_{tj}, \hat{\boldsymbol{\alpha}}^1)] \widetilde{\chi}_t(\mathbf{f}_{tj}) w(X_t) \\
& + \frac{1}{n} \sum_{t=1}^{n} [\psi''(\mathbf{f}_{t,A}, \hat{\boldsymbol{\alpha}}^1) \widetilde{\chi}_{t,A}(\mathbf{f}_0) \widetilde{\chi}_{t,A}(\mathbf{f}_0)^T w(X_t)] \sqrt{n} (\boldsymbol{\alpha}^{*1} - \hat{\boldsymbol{\alpha}}^1) \{1 + o_p(1)\} \\
& + o_p(1) \qquad (4.41)
\end{aligned}
$$

Because $\hat{\alpha}^1$ is a consistent estimation as shown in previous theorems, by central limit theorem, we have:

$$
-\frac{1}{\sqrt{n}} \sum_{t=1}^{n} [(Y_t - \psi'(\mathbf{f}_{t,A}, \hat{\boldsymbol{\alpha}}^1)] \widetilde{\chi}_t(\mathbf{f}_{t,A}) w(X_t) \xrightarrow{\mathrm{d}} N(0, \boldsymbol{I}_{11}) \qquad (4.42)
$$

and

$$
\frac{1}{n} \sum_{t=1}^{n} [\psi''(\mathbf{f}_{tj^c}, \hat{\boldsymbol{\alpha}}^1) \widetilde{\chi}_{t,A}(\mathbf{f}_0) \widetilde{\chi}_{t,A}(\mathbf{f}_0)^T w(X_t)] \xrightarrow{\mathrm{p}} \boldsymbol{I}_{11}. \qquad (4.43)
$$

Thus,

$$
\sqrt{n}(\hat{\boldsymbol{\alpha}}^1 - \boldsymbol{\alpha}^{*1}) \xrightarrow{\mathrm{d}} N(0, \boldsymbol{I}_{11}^{-1} \boldsymbol{I}_{11} \boldsymbol{I}_{11}^{-1}) = N(0, \boldsymbol{I}_{11}^{-1}). \qquad (4.44)
$$

The asymptotic normality part is proven.

□

Theorems 4.2-4.4 together give the asymptotic normality and consistency (which is called *Oracle Property*) for the proposed Adaptive LASSO semi-parametric regression model.

4.5 An application to FTSE 100 index

In this section, we give an application to the FTSE 100 index data, assuming the market price direction follows a binomial distribution, to show the strength of our penalised GMAMaR model, as an extension to Chapter 3(Peng and Lu (2021a)). The data include the open price op_t, close price cp_t, the maximise price of the day $maxp_t$ and the minimum price of the day $minp_t$, the trading volume Vlm_t from $01-May-2013$ to $01-May-2018$, of 1263 observations. We are concerned with whether the market price go up ($Y_t = 1$) or not ($Y_t = 0$) with the relationship between volatility, volume and geometric return, depicted in figure (1) and defined, respectively, by

$$
\begin{aligned}
Y_t &= \begin{cases} 1 & \text{if } cp_t - cp_{t-1} > 0; \\ 0 & \text{else.} \end{cases}, \\
v_t &= log(100\frac{(maxp_t - minp_t)}{\frac{1}{2}(maxp_t + minp_t)}), \\
V_t &= log(Vlm_t), \\
G_t &= 100log(\frac{cp_t}{cp_{t-1}}).
\end{aligned} \tag{4.45}
$$

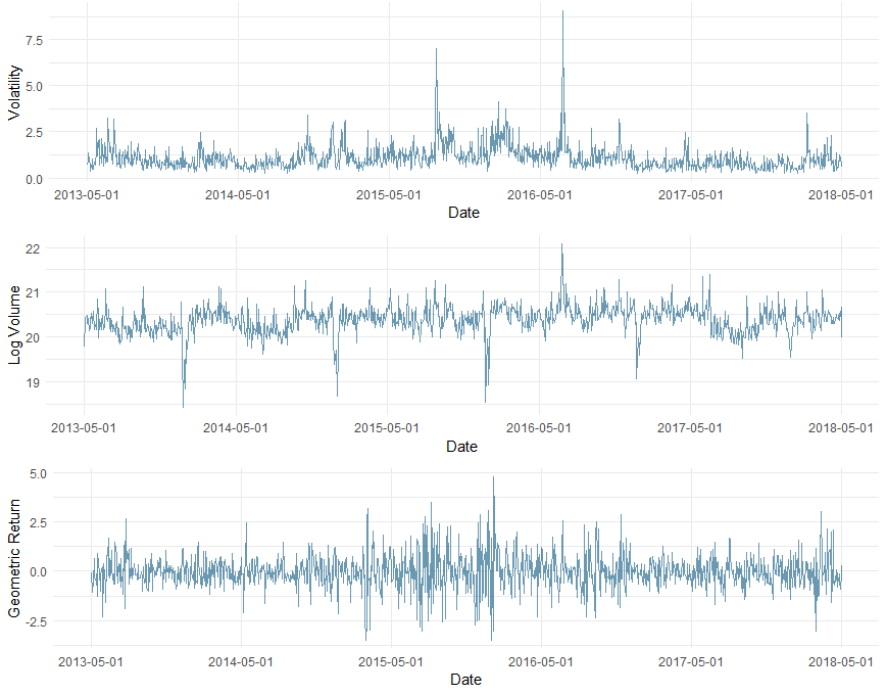

Figure 1: The time series plot of volatility V_t, volume v_t and geometric return G_t

We are interested in the ability to select only the important independent variables when the dimensionality d is large. We thus examine the performances of the variable selection using the one-step-ahead prediction of the market price Y_t with the information of long lags l (from max lag $l = 30$ to $l = 50$) of all volatility, volume and geometric return to check if they are helpful in improving the explanation or prediction of market price. That is, each lagged term will be treated as a single covariate to be fed to the model and we are using $X_t = (v_{t-1}, v_{t-2}, v_{t-3}, \ldots, v_{t-l}, V_{t-1}, V_{t-2}, V_{t-3}, \ldots, V_{t-l}, G_{t-1}, G_{t-2}, G_{t-3}, \ldots, G_{t-l})$ to predict Y_t. The training sample we used is from the 1st observation to the 1100th observation. Our evaluation sample of the prediction is

the following 50 observations (1101 to 1150) right after the training sample. Since Y_t is binary, we are computing the receiver operating characteristic curve (ROC) and area under the curve (AUC) to compare the performance (see Ballings et al. (2015)). In particular, the value of AUC is calculated based on the Receiver Operating Characteristic curve (ROC), which depicts the true positive rate against the false positive rate. In this sense, the larger AUC, the better the model.

When the max lag $l = 30$, we have in total 90 covariate variables in the model with most of them not statistically significant. After applying the penalty, only three covariates are selected that can contribute to explaining the market evolve direction, see Table 4.1. We notice that the volatility term v_3 is significant at the confidence level of 95%, while the geometric return terms G_{14} and G_{21} are significant at confidence level of 99%. This is to say, the past information of geometric return is more powerful in explaining the market evolution and we can expect such impact to have a (roughly) weekly cycle. However, since term G_7 is missing here, the market cycle cannot be confirmed yet. One possible conjecture could be that the fluctuation in the previous week has covered the normal cycle pattern.

Penalised GMAMaR model (lag=30)			
	Estimate	Std.Error	Pr(—z—)
Intercept	-0.04868	0.07157	0.49639
v_3	1.07576	0.49916	0.03115 *
G_{14}	0.47238	0.15948	0.00306 **
G_{21}	0.37952	0.14072	0.00700 **
AIC		1465.6	
Signif codes: *** 0.001 ** 0.01 * 0.05 . 0.1 1			

Table 4.1: GMAMaR model with max lag l = 30

Now we extend the max lag l to 50 to feed the model more past information. Table 4.2 summaries that there are 7 statistically significant covariates. It is interesting that now v_3 term is omitted, when compared to Table 4.1. Instead, the model identifies the Volume term V_{23}. In the sense of practice, a larger Volume does imply a more active market, and thus leads to larger volatility and market evolution. However, we may suspect that this term will ease if we can include more correlated variables, as such relationship is indirect. As to the geometric return, we can see the weekly pattern more clearly. Though the model suggests the term G_8 instead of G_7 and gives

Penalised GMAMaR model (lag = 50)			
	Estimate	Std.Error	Pr(—z—)
Intercept	-0.4960	0.1205	3.88E-05 ***
V_{23}	1.6516	0.7613	0.030052 *
G_8	0.6695	0.2862	0.019306 *
G_{14}	0.4893	0.1731	0.004701 **
G_{21}	0.3964	0.1464	0.006774 **
G_{34}	0.5688	0.2362	0.016028 *
G_{35}	0.9514	0.4231	0.024522 *
G_{48}	0.7719	0.2167	0.000368 ***
AIC		1416.7	
Signif codes: *** 0.001 ** 0.01 * 0.05 . 0.1 1			

Table 4.2: GMAMaR model with max lag l = 50

two continuous term G_{34} and G_{35}, we notice that it could be a result of the turbulence in the market evolution observed from data, and thus they are only significant at 95% confidence level.

We also notice that the AIC value has decreased from 1465.6 to 1416.7 with more past information available, which encourages us to include more correlated variables into the model. Though it is out of the scope of this chapter to identify all the correlated variables explaining the stock market, this points out the potential value of our model as it can significantly reduce the dimensionality of large data by removing the non-correlated terms. This is especially important in practice since we are often unable to identify which covariates are useful by human experience.

Another perspective to discuss is the prediction ability of the proposed model. Here we compute the ROC plot for the GMAMaR model without (Group 1) and with (Group 2) variable sections. To further distinguish the power of our model, we introduce the basic generalised linear regression model (GLM)(McCullagh and Nelder, 1989) and the random forests method (RF)(Liaw and Wiener, 2002) to compare with. Note that both the GLM and RF have the i.i.d assumption for the data set.

Figures 2 and 3 suggest that the prediction ability for our GMAMaR model does improve with variable selection (Group 2). We notice that, with more past information being included, the final prediction ability has also been enhanced after applying the penalty.

On the other hand, the overfitting problem is observed for GLM model. At first glance, the prediction ability of GLM is surprisingly good, especially

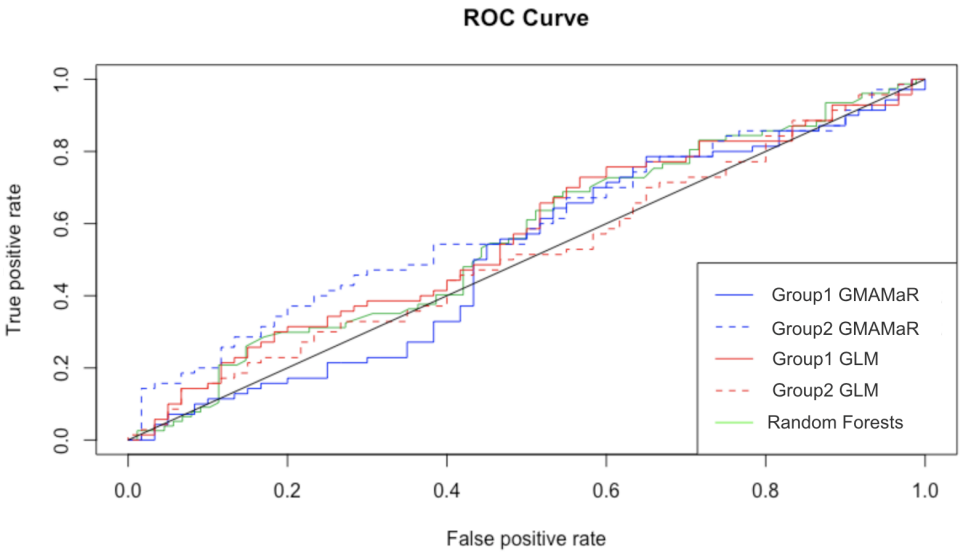

Figure 2: The ROC curve for Group 1 and Group 2 that is without and with variable selections, respectively, of max lag $l = 30$

when more variables are included without selection (Group 1). However, with more non-correlated variables included in the model, the modelling error term shall increase, and thus leads to worse estimation and prediction. We therefore confirm that GLM model cannot capture the true relationships between correlated variables and Y_t, as its performances fail to improve after variable selection.

As to RF method, based on its theory, there is no need to add lag terms for it. Hence the prediction power (AUC) is consistent in both case, which is 0.5704. This is to say, the RF method cannot capture the true relationship as it is unable to identify the time dependent trends. However, since it can be seen as the model average of decision tree method, it doesn't suffer from overfitting problem. We notice that, for lag $l = 50$, as summarised in Table 4.3, our proposed GMAMaR model have the best prediction power as indicated by a higher AUC value with variable selection. Thus for FTSE 100 index data, the result suggests that it is more likely to have a dependent data structure as assumed in this chapter.

In fact, for the prediction of a binary variable, any pure guessing shall has a success rate near 50% according to the law of large number. Thus any prediction model with AUC being around 0.5 can hardly be distinguished from pure guessing. Back to our case, however, there are two important factors to be noticed. On one hand, it is hard to have a good performance

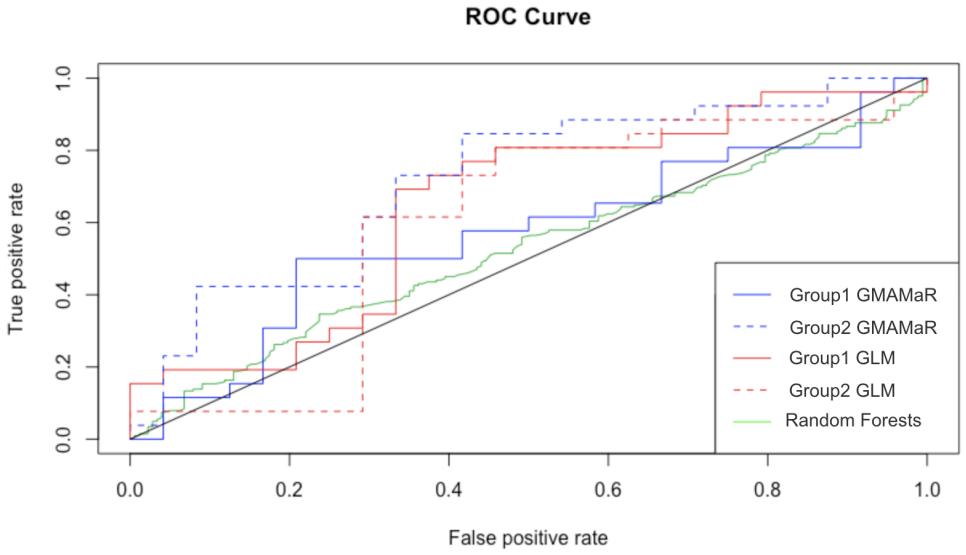

Figure 3: The ROC curve for Group 1 and Group 2 that is without and with variable selections, respectively, of max lag $l = 50$

Table 4.3: AUC comparison with and without variable selection

	Lag = 30		Lag = 50	
Model	Group 1	Group 2	Group 1	Group 2
GMAMaR	0.5100	0.5624	0.5673	0.718
GLM	0.5888	0.5174	0.6458	0.5978
Random Forests		0.5704		

Group 1: Without Adaptive LASSO;
Group 2: With Adaptive LASSO.

due to the limit of test sample size to reduce the variance. On the other hand, there is no guarantees or guidance for us to choose the possible correlated variables to be included into the model. Hence we never know if the model has performed at its best. In fact, it is realised that human experience has large deficiency in this domain, and many studies start to include more variables, with some of them being non-financial factors, such as weather and social aspects (e.g., mood). Thus we would like to conclude here that our proposed GMAMaR model with variable selection technique is a good fit into this domain to help people understanding the true relationships between variables in a possibly high and ultra-high dimension with dependent data structure.

4.6 Conclusion

In this chapter, we propose a novel semiparametric procedure GMAMaR with variable selection technique, namely the adapted LASSO, such that it is enabled to deal with large scale discrete-valued time series data of unknown forms, particularly to select the correlated variables in high or ultra-high dimensional space. The computation of the penalised GMAMaR method is cheap to avoid to well-known "curse of dimensionality", and thus easy to implement in practice. Theoretical results, e.g., the uniform consistency and asymptotic normality of the penalised procedure are established as well. A numerical example based on FTSE 100 index data is presented to further validate the power of the model with comparison to traditional GLM method and popular machine learning method Random Forest. We believe the contributions of this chapter close the gap of variable selection in high and ultra high dimension for discrete-valued time series modelling and provide the industry with a more robust method to deal with real data, in particular where variables are not completed understood. This model can also be further extended to spatio-temporal domain such that not only the relations of time dependency but also location dependency can be included.

The study of high and ultra-high dimensional data is still active. We hope to demonstrate that the proposed penalised GMAMaR procedure can further contribute to this field in further research.

Chapter 5

Modelling the COVID-19 Data in the UK: A Spatio-Temporal Analysis of Count-Valued Data

The COVID-19 pandemic has impacted the way people live worldwide, including the UK. In this chapter we propose a spatio-temporal model for count data with both nonlinear time trends and autoregressive and spatial neighbouring effects as well as other mobility and news index data considered for estimation and prediction of the COVID-19 data in the UK. Our proposed model is shown to perform more effectively with the aid of variable selection technique. We can thus extract useful information providing more insights empirically into the key factors contributing to the daily confirmed cases at different locations. We find that the success of interventions varied depending on locations, as a location may subject to its population, medical resource and its role in the national or international transportation network. Our findings also show that the neighbouring effects are significant, and hence limiting public transportation nearby is effective to control the spread of pandemic by reducing contacts. Moreover, we empirically find that the media effects are significant, which may well promote self-protection awareness in controlling the spread of pandemic. It is further shown that all these effects are varied with locations. We can expect these findings and techniques would be useful in guiding and supporting the policy making and

allocation of the resources that are location based for pandemic controls, and in making contribution to the related COVID epidemiology studies.

5.1 Introduction

The COVID-19 pandemic has impacted the way people live worldwide, including the UK. The development of modern techniques has made it possible for spatial time series data to be collected along time and over locations, which becomes prevalent in today's research. For example, Ferguson et al. (2020) used the data collected in mainland China to predict the infected number of COVID-19 for the UK, which turns out to be probably one of the famous works on COVID-19 to the public. Similarly, the Institute of Health Metrics (IHME) at the University of Washington has published regular predictions in the United States for specific locations. However, their figures reported are criticised by later studies (c.f., Avery et al. (2020)), which fail to incorporate the spatio-temporal effect of infectious rate. Indeed, traditional epidemiology models, e.g., SIR (susceptible/infected/recovered) model, do not consider the neighbouring effect when facing spatial data, see e.g., Unwin et al. (2020). While many spatio-temporal methods have been developed in the literature (c.f., Cressie and Wikle (2015)), they are yet rarely, or only preliminarily, applied to the study of pandemic, though it is believed in epidemiology studies that the infectious rate, often known as R_0, is a time-varying and location dependent value. Aisyah et al. (2020), for instance, described the COVID-19 data in Indonesia in the spatial-temporal format, which supports the argument of Avery et al. (2020). Another attempt of discovering the spatial and temporal effects of COVID-19 data in Africa can be found in Gayawan et al. (2020), where the spatio-temporal effects are assumed to be random effects and estimated using Bayesian inference techniques with prior assumptions of distributions. Giuliani et al. (2020), on the contrary, adopted the time-series mixed effects generalised linear model to understand and predict the spatio-temporal effects of the infections of COVID-19 in Italy. However, we notice that they only consider the past information of lag 1 and spatial effects of neighbouring provinces that share a border with the reference location. Unlike spatio-temporal modelling, time series analysis for the COVID-19 may have been examined more widely in the literature, see e.g., Kim et al. (2020) and Zhu et al. (2021). These studies

assume the Poisson distribution for the daily death number, which is commonly adopted for count data. Alternatively, other distributions are applied for epidemiology models (see, e.g., Ferguson et al. (2020)), and it is also often assumed that the true infectious number follows a Gamma distribution while the observed infectious number follows a Negative-Binomial distribution.

Differently from the references mentioned above, in this chapter, we aim to present more effective statistical techniques for the modelling and analysis of spatio-temporal daily confirmed number data of the COVID-19 at 367 local authorities in the UK (see details of the data illustration in Section 2 below). We will suggest our spatio-temporal models for COVID-19 data based on the following points. First of all, following Kim et al. (2020) and Zhu et al. (2021), we propose to apply in this chapter the Poisson distribution to model the count data of daily confirmed number of COVID-19, as it is reported that for an uncontrolled pandemic the distributions of the time series of cases are roughly symmetric and bell-shaped (c.f., Farr (1840)). This is different from the spatio-temporal modelling with continuous-valued responses in the current literature (c.f., Lu et al. (2009), Al-Sulami et al. (2017)). Secondly, for spatio-temporal data, assumption of some kind of stationarity is often needed for statistical inference, see e.g., Lu et al. (2009). However, this may often be violated in practice. The spread and development of COVID-19 cases are known to have had two waves to the date (as of the data considered). The time series of daily confirmed number of new cases show clearly two peaks. This, in other words, implies that the time series of COVID-19 data is not stationary in time. Therefore, we suggest allowing the time trend to be modelled in a nonlinear function instead of being a constant, in order to permit the non-stationary series with time trend to be modelled into a time series model framework at each location. Moreover, we follow Al-Sulami et al. (2017) and Lu et al. (2009), allowing the spatial data are non-stationary across spacial locations on irregular sampling grids. Thirdly, we will suggest extending the idea of spatial neighbouring lag interactions for the continuous-valued responses as considered in Al-Sulami et al. (2017) and Lu et al. (2009) to modelling of spatial neighbouring lag interactions with discrete-valued daily confirmed number of new cases for the COVID-19 data.

The purpose of this chapter is therefore to apply our proposed spatio-temporal

model for modelling and analysis of the COVID-19 data collected at irregularly spaced locations of 367 local authorities in the UK. The feature of the model considers the relationship between response and covariates as well as both the temporal lag and spatial neighbouring effects. It is noted that such consideration is advocated in literature, e.g., Avery et al. (2020). We further allow the time trend at each location to be nonlinear instead of constant to ease the non-stationary nature of COVID-19 data in modelling. For instance, we can consider possible relationships of mobility and news index on daily confirmed cases for 367 local authorities in the UK, which will be examined in this chapter. Optimal temporal lags as well as spatial neighbouring lags are selected for each location. Note that the dimension of the regression in our model at one location can be as high as $(21 * 367 + 10) = 7717$ maximum on the basis of its 21-day (3 weeks) temporal lags over 367 locations, plus 10 covariates on mobility and news index[1]. Therefore, the selection of the significant spatial and temporal lags and the related covariates is important for our understanding of the dynamic behaviours of the pandemic over time and across space. As a comparison to our proposed model, we will also present the results for the models without consideration of spatial neighbouring effects and the models without consideration of time trends. We will see that the variable selection technique applied in the analysis will help to extract the information from data and therefore provide further insights for decision makers and researchers to understand how such pandemic dynamically develops in both time and spatial dimensions and how to more effectively control the development of such pandemic.

In particular, the empirical findings show that: (1) The daily confirmed number has strong time trends across different time periods of interventions; (2) A neighbouring effect is also identified that areas of key role in a transportation network often suffer more serious infection of COVID-19, and thus the action of lockdown is indeed an effective measure in the combat of pandemic; (3) The media effects are significant, which may well promote self-protection awareness in controlling the spread of pandemic; (4) However, we demonstrate that all these effects identified are varied with locations, so different local authorities may need to implement varied policies regarding

[1]For the ease of understanding, we would like to defer the discussion of these covariates later to Section 2.

control measures such as lockdown, not only because of the varied population size and medical resources, etc., but also depending on their roles in a transportation network. Therefore, limiting public transportation and using media advertisement of COVID-19 information to promote the public's awareness of self-protection (e.g., wearing masks) should be of the first priority in the sense of preventing the spread of this pandemic. We can expect these findings and the suggested methods will be useful in guiding and supporting the policy making and allocation of the resources that are location based for pandemic controls, and in making contribution to the related COVID epidemiology studies.

The remainder of this chapter is structured as follows. In section 2, we present the COVID-19 data used in this chapter with a background introduction. We present the spatio-temporal model with methods in Section 3 including the techniques adopted for model estimation and variable selections. Empirical findings are illustrated in Section 4, where details of numerical results and more insights are presented. Finally, conclusions are summarised in Section 5.

5.2 The COVID-19 data

5.2.1 Background

As confirmed on 31st January 2020, the first case of COVID-19 was identified in the UK. In fact, the world has experienced a serious global pandemic since then. The number of confirmed cases in UK was among the highest of the world. Globally, the total number grew to 102,283,784, including 2,219,236 deaths, on 31st January 2021. As to the UK, there were 3,818,423 confirmed cases of COVID-19 as well as 105,571 deaths at that time. We will focus on analysis of a UK COVID-19 data set that we have collected in a spatio-temporal manner in this chapter.

The impact of COVID-19 had deeply affected the way people live. To deal with such a global pandemic, travels and transportations were under a stricter supervision to prevent its spread. As a consequence, it has greatly damaged the global economy and businesses. For instance, the crude oil price of West Texas Intermediate crude dropped to negative for the first

time ever on 20th April 2020. In the UK, the stock market index, namely FTSE 100 Index, reached 4993.89 on 23rd March 2020 due to the first wave of COVID-19. Its performance had not yet been back to the level of 7000 ever since, at the date of 10th January 2021.

In the combat against COVID-19, all efforts, including both governmental policies and individual behaviours, were made, aiming to reduce the value of infectious rate R_0, in order to save lives. Avery et al. (2020) had stated that two actions would be expected facing such a serious contagious disease: (1) Government would enact policies, such as lockdown, to slow or stop the spread; (2) People will need to change their behaviour to avoid getting infected. Both actions would lead to reducing R_0 with COVID-19 spread, and thus save thousands of lives. In order to prevent the further spread of COVID-19, the UK government had announced three national lockdowns. The first lockdown started on 23rd March 2020 and lasted into July 2020. The second one came into force on 5th November 2020, and ended on 2nd December 2020. The latest lockdown was implemented on 4th January 2021 and it was expected to last until March. During the lockdown, people were restricted to stay at home, e.g., work from home, unless for essential needs. Most of the social events, as well as sports, are banned.

To explore more clearly the impact of these reactions on the spread of COVID-19, there have been some previous studies, e.g., Unwin et al. (2020), who have proposed to use the mobility data obtained from mobile devices to reflect the implementation of lockdown, during which people are restricted to live, study and work from home. Consequently, the mobility rate would decrease dramatically under lockdown. Note that such number would also differ, if we compare different areas, e.g., at residential areas and commercial districts, or via different transportation means, e.g., by driving and walking. It is thus worthwhile to investigate if the impacts of lockdown are identical for each location or if lockdown is only necessary for some of them. Moreover, people travelling with different transportation means may be subjected to heterogeneous levels of risks with locations.

On the other hand, it is relatively hard to collect data linked directly to people's behaviours like social distancing and wearing masks. The studies on the USA cases, e.g., Jamieson and Albarracin (2020) and Barrios and Hochberg (2020), have paid attention to the public media and detected the correlation between people's awareness/knowledge of safe behaviours and

the related resources available on media. In this chapter, we have therefore consulted the UK News Index for this type of data, roughly representing how many media discussions on COVID-19 are published every day in UK.

5.2.2 Data

In this chapter, we consider a roughly 10 months data set of COVID-19 daily increase number in the UK, published by the UKGovernment (2021), covering 352 days with the time period from 15th-Feb-2020 to 31st-Jan-2021 (as of the date when this work was started) and 367 locations of the UK lower tier local authorities in total. Here Northern Ireland was excluded in this study as it is separated from the other parts of the UK by the sea with less transportation because of the pandemic control, so we only consider the study of the areas of the Great Britain in this chapter. See Figure 1 for the locations in which we have plotted the accumulated confirmed cases of the COVID-19 in our date set. In this figure, the accumulated number of cases, from 62 case in Orkney Island (the smallest number) to 87641 cases in Birmingham (the largest number), are grouped with the five values in quantiles at 0% (62 cases), 25% (4569 cases), 50% (6974 cases), 75% (12762 cases) and 100% (87641 cases) respectively. We can see that darker areas suffer from the COVID-19 more seriously, while lighter areas are barely impacted (except some unavailable data-NA part).

Following Unwin et al. (2020), we consulted mobility data obtained both from Google (2021) and Apple (2021) to represent the effect of lock down. For instance, the mobility would drop fast once lockdown is implemented, and increase otherwise.

In particular, Google data are calculated based on a benchmark setting such that we can see the change of mobility in those places:

- Grocery & pharmacy: Mobility data collected from grocery markets and pharmacies.
- Parks: Mobility data collected from public parks.
- Transit stations: Mobility data collected from public transport hubs.
- Retail & recreation: Mobility data collected from public place of entertainment.

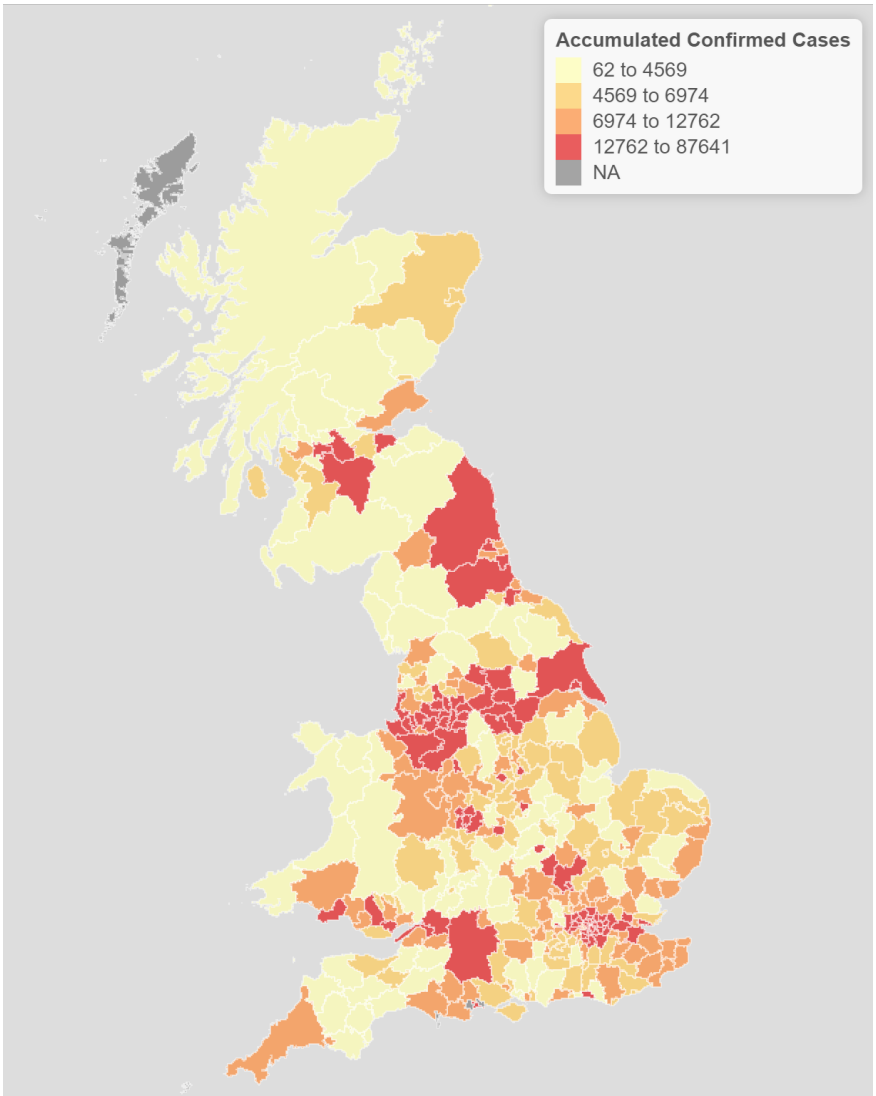

Figure 1: Accumulated Confirmed Case (partitioned in quantiles at 0% (62 cases), 25% (4569 cases), 50% (6974 cases), 75% (12762 cases) and 100% (87641 cases)) of Great Britain up to 31st January 2021 (The darker colour & larger percentage indicate the more serious accumulated number of infected patients in that area).

- Residential: Mobility data collected from residence.
- Workplaces: Mobility data collected from workplace.

Apple data, on the other hand, only classify the mobility into public transportation, walking and driving to represent the daily travels. We notice that all these data are collected based on the mobile devices using either Google or Apple services and there should be little overlaps.

In addition, we also consider the UK Daily News Index available from an online website EconomicPolicyUncertainty (2021), which is also known as

newspaper-based Economic Policy Uncertainty (EPU) Index. The EPU refers to the count of articles that at least match one of the three term sets as given by its name. It may serve as an indication on how COVID-19 has been connected to our daily lives in many different aspects.

5.3 Methodology: A spatio-temporal model for Covid-19 data

5.3.1 Model assumption and structure

Let $Y_t(s_k)$ and $X_t(s_k)$ denote two spatio-temporal processes with observations at discrete time points, $t = 1, ..., n$, that covers the $n = 352$ days from 15th February 2020 to 31st January 2021, where $Y_t(s_k)$ is the concerned daily confirmed number of COVID-19 cases at a given location s_k among $m = 367$ local authorities in Great Britain, and $X_t(s_k)$ is the covariate vector that contains $D = 10$ covariate variables including Google and Apple mobility data and News index, as introduced in Section 2. The spatial unit of a local authority, s_k, is denoted as $s_k := (u_k, v_k) \in \mathbf{R}^2$, where u_k are v_k are x (longitude) and y (latitude) coordinates of representing locations of local authorities with index $k = 1, ..., m$. At a given spatial location s_k, we can consider a time series model with our spatio-temporal structure as follows.

Denote by I_{t-1} for the information up to time $t-1$ about time series $Y_j(s_k)$ for $j \leq (t-1)$ with all considered locations of $k = 1, 2, \cdots, m$. The regression is to model the following conditional expectation:

$$\lambda_t(s_k) = E[Y_t(s_k)|I_{t-1}]. \tag{5.1}$$

As suggested in Kim et al. (2020) and Zhu et al. (2021), we further assume the daily number of new cases $Y_t(s_k)$ follows a conditional Poisson distribution:

$$Y_t(s_k)|I_{t-1} \sim Poisson(\lambda_t(s_k)). \tag{5.2}$$

By noticing that the $Y_t(s_k)$ is both time and spatio dependent, we need to include the temporal-lag autoregressive term $Y_{t-p}(s_k)$, for $p = 1, ..., P$, and the spatial neighbour time lagged term $Y^s_{t-q}(s_k) = \sum_{i=1}^{m} w_{ki} Y_{t-q}(s_i)$,

for $q = 1, ..., Q$. We will specify spatial weight w_{ki} as well as the spatial neighbouring effect $Y^s_{t-q}(s_k)$ with details in the next subsection. Then we can write our family of location-dependent spatial-temporal model, extending Al-Sulami et al. (2017) to a count data process $Y_t(s_k)$, as follows:

$$\log(\boldsymbol{\lambda}_t(s_k)) = f(\tilde{t}, s_k) + \sum_{p=1}^{P} \alpha_p(s_k) Y_{t-p}(s_k) + \sum_{q=1}^{Q} \beta_q(s_k) Y^s_{t-q}(s_k) + \sum_{d=1}^{D} \boldsymbol{\gamma}_d(s_k) X_{td}(s_k). \tag{5.3}$$

Here $f(\cdot, s_k)$ denotes a nonlinear time trend function at location s_k with $\tilde{t} = t/n$ for $t = 1, 2, \ldots, n$, and is of own interest in view of the time trending with non-stationary count process of the daily increase number, $Y_t(s_k)$, of COVID-19 cases at location s_k (c.f., as illustrated in Figure 4 below). This point is different from the usual idea of differencing operation in time that is popularly applied to make time series data stationary for continuous-valued data (c.f., Al-Sulami et al. (2017)). Note that it is more difficult to apply the differencing to the count-valued time series data $Y_t(s_k)$ as it makes it harder to model (5.1) through (5.2) with changes of the distribution for the new data. Therefore introducing this trending function is important to model the daily increase number series of COVID-19 cases in this chapter. Estimation of $f(\tilde{t}, s_k)$ will be further discussed in Subsection 5.3.2.1. The coefficients, $\boldsymbol{\alpha}(s_k) = (\alpha_1(s_k), ..., \alpha_P(s_k))'$, $\boldsymbol{\beta}(s_k) = (\beta_1(s_k), ..., \beta_Q(s_k))'$ and $\boldsymbol{\gamma}(s_k) = (\gamma_1(s_k), ..., \gamma_D(s_k))'$ in model (5.3) are the parameter vectors to be estimated, which allow to be location s_k dependent, where the $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the spillover weights assigned to dynamic lag effects of the response $Y_t(s_k)$ at the location s_k itself and its spatial weighted neighbours, respectively, and $\boldsymbol{\gamma}$ represents the spillover effects of the covariates on mobility and news indexes, with A' standing for a transpose of a vector or matrix A.

For notational simplicity, we let $Z_t(s_k) = [Y_{t-1}(s_k), ..., Y_{t-P}(s_k), Y^s_{t-1}(s_k), ..., Y^s_{t-Q}(s_k), X_{t1}(s_k), ...X_{tD}(s_k)]'$ and $\boldsymbol{\theta}(s_k) = [(\boldsymbol{\alpha}(s_k))', (\boldsymbol{\beta}(s_k))', (\boldsymbol{\gamma}(s_k))']'$. The formula (5.3) can be written as:

$$\log(\boldsymbol{\lambda}_t(s_k)) = f(\tilde{t}, s_k) + (\boldsymbol{\theta}(s_k))' Z_t(s_k). \tag{5.4}$$

5.3.2 Estimation

For model (5.4), before discussion of estimation of $\boldsymbol{\theta}(s_k)$ via maximum log-likelihood method, we first present some ideas on how to construct the time trend function $f(\cdot, s_k)$ by splines and the spatial neighbour $Y^s_{t-q}(s_k)$ by weight matrix as precursors.

5.3.2.1 Nonlinear time trend function

For the time trend term $f(\tilde{t}, s_k)$, the idea is to capture the nonlinear trending of time series at each location with $\tilde{t}$ increasing for the whole time period considered. A popular idea for nonlinear function fitting is by a spline method that uses the linear combination of basis functions to present the nonlinear structure. We can write the smooth trending function at a chosen location s_k as follows:

$$f(\tilde{t}, s_k) = \sum_{r=0}^{R+\delta} \mu_r(s_k)\eta_r(\tilde{t}), \tag{5.5}$$

where δ is the degree of basis polynomial, R is the total number of inner knots used in construction of spline, and $\eta_r(\tilde{t})$'s are the set of spline basis functions with coefficients $\mu_r(s_k)$ at location s_k; c.f., Perperoglou et al. (2019).

It is widely reported that cubic splines ($\delta = 3$) are popular as a perfect curve smoothing; see e.g., Hastie et al. (2009). In this chapter, we thus follow this standard and generate a smoothed nonlinear curve using the B-spline basis functions. This can be implemented by R package **Splines** with function **bs** (c.f., Venables et al. (2009)).

We then fit the smoothed time trend $f(\tilde{t}, s_k)$ in cubic spline into the regression model (5.3) replacing the constant intercept, and we will obtain the estimated coefficients $\mu_l(s_k)$ in the estimation subsection below. In this chapter, we select the degrees of freedom to df $= 6$ according to the AIC values of the fitted model (5.3): for example, AIC$= 2802.5, 2803.7, 2622, 2580$ and 2625 respectively for df $= 3, 4, 5, 6$ and 7 at City of London. That is, when specifying the degrees of freedom to df $= 6$, the number of inner knot points is $R = \text{df} - \delta = 3$. In addition, the intercept, i.e., the term η_0, is omitted in the computations of this chapter (otherwise it may lead to multicollinearity in regression).

5.3.2.2 Spatial neighbouring effect

For a given location s_k, the corresponding spatial neighbouring effect can be characterised by:

$$Y^s_{t-q}(s_k) = \sum_{i=1}^{m} w_{ki} Y_{t-q}(s_i), \tag{5.6}$$

where $w_{ki} \geq 0$ is the spatial weight characterising the neighbouring effect from location s_i to s_k, standardised to satisfy $w_{ii} = 0$ and $\sum_{i=1}^{m} w_{ki} = 1$. This term is often pre-specified a priori in econometrics; see Anselin (2013)

In this chapter, to determine the spatial weight w_{ki}, we use the inverse distance between locations, i.e., $w_{ki} = 1/d_{ki}$, where d_{ki} is measured in distance on Euclidean space between locations s_k and s_i (c.f. Wilhelmsson (2002)). It is adopted in Al-Sulami et al. (2017). This weight corresponds to the assumption that a virus host would travel more frequently between locations nearby with larger impacts. We then standardise it, such that $\sum_{i=1}^{m} w_{ki} = 1$. The weighted spatial effect Y^s_t is obtained henceforth. In the model, this spatial effect is pre-calculated before estimation and then treated as the covariate variables. We aim to investigate the corresponding neighbouring effect on response Y_t at each location.

5.3.2.3 Estimating the unknown parameters

Given the Poisson model (5.3) above, we want to find the estimators, $\hat{\boldsymbol{\theta}}(s_k) = [(\hat{\boldsymbol{\alpha}}(s_k))', (\hat{\boldsymbol{\beta}}(s_k))', (\hat{\boldsymbol{\gamma}}(s_k))']' = [\hat{\alpha}_1(s_k), ..., \hat{\alpha}_P(s_k), \hat{\beta}_1(s_k), ..., \hat{\beta}_Q(s_k), \hat{\gamma}_1(s_k), ...\hat{\gamma}_D(s_k)]'$, via maximum log-likelihood method.

Recall that $Z_t(s_k) = [Y_{t-1}(s_k), ..., Y_{t-P}(s_k), Y^s_{t-1}(s_k), ..., Y^s_{t-Q}(s_k), X_{t1}((s_k), ...X_{tD}((s_k)]$ and I_{t-1} is denoted for all the available information known up to time $t-1$. Thus, based on the time series observations, $\{(Z_t(s_k), Y_t(s_k)), t = 1, 2, \cdots, n\}$, at location s_k, we can write the conditional likelihood function following from (5.2), given the initial available information I_0, for estimation of $\boldsymbol{\theta}(s_k)$ as follows:

$$\begin{aligned} Lik_S(\boldsymbol{\theta}(s_k)) &= \prod_{t=1}^{n} p_S(Y_t(s_k)|I_{t-1}(s_k); \boldsymbol{\theta}(s_k)) \\ &= \prod_{t=1}^{n} \frac{e^{Y_t(s_k)[f(\tilde{t},s_k)+(\boldsymbol{\theta}(s_k))' Z_t(s_k)]} e^{-e^{f(\tilde{t},s_k)+(\boldsymbol{\theta}(s_k))' Z_t(s_k)}}}{Y_t(s_k)!}, \end{aligned} \tag{5.7}$$

where for notational simplicity, we suppose I_0 stands for the information $\{(Z_t(s_k), Y_t(s_k)),$
$t \leq 0\}$, $p_S(\cdot|\cdot)$ for conditional Poisson probability function, and $Y_t(s_k)|I_{t-1}(s_k)$, $t = 1, 2, \cdots, n$, are (conditionally) independent.

Taking nature log of the equation (5.7), the log likelihood function is:

$$L_S(\boldsymbol{\theta}(s_k)) = \sum_{t=1}^{n} \{Y_t(s_k)[f(\tilde{t}, s_k) + (\boldsymbol{\theta}(s_k))' Z_t(s_k)] - \exp(f(\tilde{t}, s_k) + (\boldsymbol{\theta}(s_k))' Z_t(s_k)) - \log(Y_t(s_k)!)\}, \tag{5.8}$$

With the population (expected)log-likelihood function:

$$L(\boldsymbol{\theta}(s_k)) = \sum_{t=1}^{n} E[\{Y_t(s_k)[f(\tilde{t}, s_k) + (\boldsymbol{\theta}(s_k))' Z_t(s_k)] - \exp(f(\tilde{t}, s_k) + (\boldsymbol{\theta}(s_k))' Z_t(s_k)) - \log(Y_t(s_k)!)\}], \tag{5.9}$$

we are then seeking to estimate the parameter vector $\hat{\boldsymbol{\theta}}(s_k) = (\hat{\boldsymbol{\alpha}}(s_k), \hat{\boldsymbol{\beta}}(s_k), \hat{\boldsymbol{\gamma}}(s))$ by maximising the log likelihood function (5.8) with $f(\tilde{t}, s_k)$ specified in (5.5). Note that $\hat{\boldsymbol{\theta}}(s_k) = \arg\max L_S(\boldsymbol{\theta}(s_k))$ giving the estimator $\hat{\boldsymbol{\theta}}(s_k)$ from sample data, and $\boldsymbol{\theta}_0(s_k) = \arg\max L(\boldsymbol{\theta}(s_k))$ giving the true parameter vector $\boldsymbol{\theta}_0(s_k) = (\alpha_0(s_k), \beta_0(s_k), ..., \gamma_0(s_k))^T$.

Note for each s_k the log-likelihood function (5.8) are continuous and twice differentiable. We now find the first order derivatives of the log-likelihood function (5.8) with respect to $\boldsymbol{\theta}(s_k)$, leading to the following type of estimation equations:

$$L_S^{(1)}(\boldsymbol{\theta}(s_k)) = \frac{\partial L_S(\boldsymbol{\theta}(s_k))}{\partial \theta(s_k)} = \sum_{t=1}^{n} [Y_t(s_k) - \exp(f(\tilde{t}, s_k) + (\boldsymbol{\theta}(s_k))' Z_t(s_k))] Z_t(s_k) = 0. \tag{5.10}$$

The empirical Fisher information matrix, the inverse of which is an estimator of the asymptotic variance of the estimator $\hat{\boldsymbol{\theta}}(s_k)$, is obtained by deriving the negative second-order partial derivatives, and is positive definite:

$$L_S^{(2)}(\boldsymbol{\theta}(s_k)) = -\frac{\partial^2 L_S(\boldsymbol{\theta}(s_k))}{\partial\theta(s_k)\partial(\theta(s_k))'}$$
$$= \sum_{t=1}^{n} Z_t(s_k)[\exp(f(\tilde{t}, s_k) + (\boldsymbol{\theta}(s_k))' Z_t(s_k))](Z_t(s_k))'. \quad (5.11)$$

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

Intuitively, if $L_S^{(1)}(\boldsymbol{\theta}(s_k))$ is asymptotically close to $E[L_S^{(1)}(\boldsymbol{\theta}(s_k))]$ uniformly with respect to $\boldsymbol{\theta}(s_k)$ over a compact set with $\boldsymbol{\theta}_0(s_k)$ being its interior point, then $\hat{\boldsymbol{\theta}}(s_k)$ should be close to the solution of $E[L_S^{(1)}(\boldsymbol{\theta}(s_k))] = 0$, and is a consistent estimator of $\boldsymbol{\theta}_0(s_k)$.

5.3.3 Variable selection

When estimating the model above, another practical issue is on time lag orders which need be identified. Further, there may be only a few of the variables that are actually helpful for explaining the daily increase number Y_t. We hence need to optimally select the best time lag orders, and to identify the important variables that can extract the useful information.

5.3.3.1 Selection of time lag orders

To find out the optimal settings of the time lag orders P and Q in Model (5.3), one can use the Akaike Information Criterion (AIC) value for model selection. In particular, considering small sample size of data available in this chapter, we adopt the AICc method as suggested by Hurvich and Tsai (1993), where a penalty term of sample size and number of parameters are considered. When the sample size goes to infinity, such penalty term would converge to zero. However, if the sample size is small, overfitting with too many parameters can therefore be avoided:

$$\text{AICc}(P, Q) = \text{AIC}(P, Q) + \frac{2(P + Q + D)^2 + 2(P + Q + D)}{n - (P + Q + D) - 1}, \quad (5.12)$$

where

$$\text{AIC}(P,Q) = 2(P+Q+D) - 2[\sum_{t=1}^{n}[Y_t(s_k)[f(\tilde{t},s_k) + (\boldsymbol{\theta}(s_k))'Z_t(s_k)] \\ - \exp(f(\tilde{t},s_k) + (\boldsymbol{\theta}(s_k))'Z_t(s_k)) - \log(Y_t(s_k)!\,)]].$$

The longest incubation period for the patient to show symptoms and to be confirmed by NHS is known to be less than three weeks. Thus, for each location, combined with the knowledge of the COVID-19, we optimally decide the settings of (P,Q) for $P,Q \leq 21$. It is based on the intuition that it takes maximum 14 days for the patient to show symptoms and then it may take a few more days for him/her to be treated and confirmed, see e.g., Lauer et al. (2020). We thus use 21 days as maximum time lags here. Further lags can be easily incorporated into the model if needed.

5.3.3.2 Extracting feature variables

It is noticed that overfitting is often found when having a large number of independent variables. To overcome it, variable selection techniques, such as LASSO and other ridge penalties (c.f., Friedman et al. (2010)), can be used here. Recalling the log-likelihood $L_S(\boldsymbol{\theta}(s_k))$ in (5.8), we can give the penalised log-likelihood, that is the negative log-likelihood plus a penalty as follows:

$$\tilde{L}_S(\boldsymbol{\theta}(s_k)) = -L_S(\boldsymbol{\theta}(s_k)) + \lambda P_\pi(\boldsymbol{\theta}(s_k)), \tag{5.13}$$

where $P_\pi(\boldsymbol{\theta}) = \frac{1-\pi}{2}\|\boldsymbol{\theta}\|_2^2 + \pi\|\boldsymbol{\theta}\|_1 = \sum_{j=1}^{M}[\frac{1-\pi}{2}\boldsymbol{\theta}_j^2 + \pi\boldsymbol{\theta}_j]$ with $M = P+Q+D$, and λ is the tuning parameter to be adjusted. We are seeking the coefficient vector $\hat{\boldsymbol{\theta}}(s_k)$ for each location that minimises the penalised log-likelihood function:

$$\hat{\boldsymbol{\theta}}(s_k) = \arg\min_{\boldsymbol{\theta}} - \sum_{t=1}^{n}[Y_t(s_k)[f(\tilde{t},s_k) + (\boldsymbol{\theta}(s_k))'Z_t(s_k)] \\ - exp(f(\tilde{t},s_k) + (\boldsymbol{\theta}(s_k))'Z_t(s_k)) - \log(Y_t(s_k)!\,)] + \lambda P_\pi(\boldsymbol{\theta}).$$

We consider the popular LASSO estimation with $\pi = 1$, i.e., we only use the l_1 norm ($P_\pi(\boldsymbol{\theta}) = \pi\|\boldsymbol{\theta}\|_1$), which automatically estimate the coefficients

of unimportant covariates being zero. In order to find out the best λ as well as penalised coefficients $\hat{\boldsymbol{\theta}}(s_k)$, we are actually facing a model selection problem here for different λ. The candidates models with different values of λ are normally evaluated by various information criteria, such as AIC as discussed above. For this purpose, we use AIC to find out the best λ and the associated estimated coefficients $\hat{\boldsymbol{\theta}}(s_k)$.

Therefore, after variable selection, only the non-zero parameters remain. The coefficients of other parameters would be forced to equal to zero. As a consequence, we then extract the information from the data that shows what are the important factors to consider. It is noted that such technique can be implemented by R package ***glmnet*** of Friedman et al. (2010) (although the i.i.d data situation is considered in their paper).

5.4 Empirical findings

In this section, we aim to provide a rigorously empirical analysis to understand spatio-temporal dynamic behaviour of the UK COVID-19 daily cases with the mentioned issues above being properly handled, the analysis of which is hence more comprehensive than those in the literature. For example, existing spatio-temporal studies of COVID-19, e.g., Gayawan et al. (2020) and Giuliani et al. (2020) for the Italy and African data, respectively, do not include the extraneous factors such as population mobility and only consider the spatio effects as a component of the autoregressive model in a simpler manner. As to other conventional epidemic studies, e.g., Unwin et al. (2020), the spatio effects are ignored, even though spatial data is used. Also, it is not well addressed that the data is actually non-stationary in raw, which needs to be carefully dealt with.

We believe it is of practical interest to consider micro variables to reveal the importance of different intervention actions or behaviours. In the perspective of pandemic control, the findings of our analysis will provide deeper understanding of the dynamic spread of COVID-19 and the effects of interventions adopted in Great Britain that are local authority level based. The numerical results for 8 local authorities, namely Birmingham, Cardiff, Edinburgh, Glasgow, (City of) London, Leeds, Liverpool and Manchester, are

particularly presented. These selected cities cover the most areas in the UK except for Northern Ireland.

5.4.1 Initial model selection

We first consider an initial model selection for taking nonlinear time trends and spatial neighbouring into account or not. For stating convenience, we denote our proposed Model (5.3) as ST, standing for spatial model with the consideration of nonlinear trends. Similarly, Model NS refers to the model without the terms of spatial effects, i.e., Y_t^s, in Model (5.3). Model NT would assume no temporal trend and thus the term $f(\tilde{t}, s_k)$ is ignored in Model (5.3). Finally, Model NSNT omits both the terms Y_t^s and $f(\tilde{t}, s_k)$ in Model (5.3), which is commonly adopted in the relevant literature of COVID-19 analysis (c.f., Wu et al. (2020)). The temporal lag order P and spatial neighbouring lag order Q are optimally selected as seen in Section 5.3.3.1. In particular, to measure the performance of a model in estimation and prediction, we define the mean absolute error (MAE) and absolute error (AE) as follows:

$$MAE = \frac{\sum_{i=1}^{n} |Y_i - \hat{Y}_i|}{n}, \tag{5.14}$$

$$AE_i = |Y_i - \hat{Y}_i|, \tag{5.15}$$

where $\hat{Y}_i$ is the estimated or predicted daily number of confirmed cases given in (5.1), obtained by the corresponding estimated models. In this sense, both AE and MAE give an indicator of how good the model performs in the context of estimation and prediction.

We now generate the box-plot given in Figure 2 and 3 for the AE obtained from the four models mentioned for the whole period of 324 days. We can summarise that the model considering spatial effects as well as nonlinear time trends would perform consistently better for all locations when compared with the other models mentioned above. In particular, as understandable, the impact of nonlinear time trends seem to be larger than the impact of spatial effects, characterising the changing dynamic nature of daily confirmed cases. For all locations, adopting Model NSNT would lead to the largest

error for a poor estimation. This is also further confirmed in Table 5.1, which summarises the MAE of the 324 estimates compared to true values.

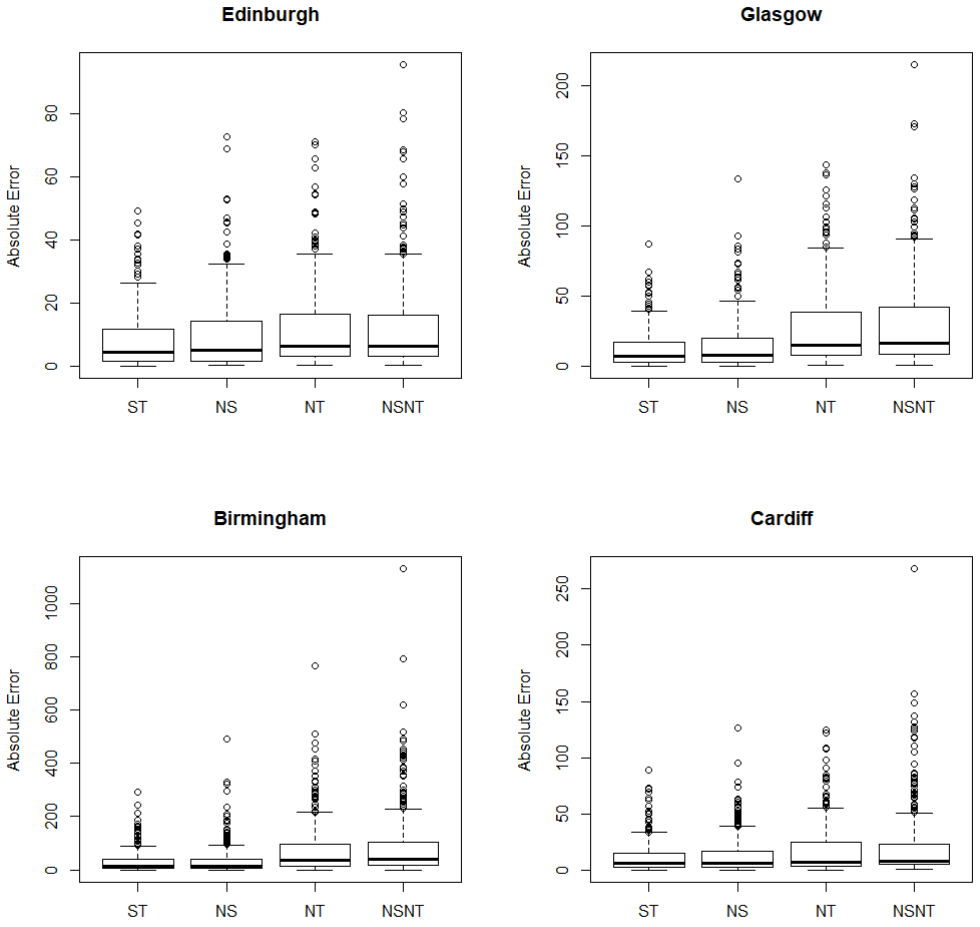

Figure 2: Boxplot of Absolute Error of estimations for Edinburgh, Glasgow, Birmingham and Cardiff with four different models, namely ST, NS, NT and NSNT.

Table 5.1: Mean Absolute Error (MAE) of estimations for 8 selected locations with four different models, namely ST, NS, NT and NSNT (The orders for temporal lag effects, P, and for spatial neighbouring lag effects, Q, are optimally selected by AIC).

Model/Location	Birmingham P=21,Q=21	Cardiff P=21,Q=21	Edinburgh P=21,Q=18	Glasgow P=21,Q=21	(City of) London P=20,Q=21	Leeds P=21,Q=17	Liverpool P=21,Q=21	Manchester P=20,Q=21
ST	30.1780	12.6959	7.7395	12.2162	7.6367	18.7671	16.1971	13.0598
NS	34.8147	13.0159	9.51804	14.2291	11.2405	23.7226	23.6703	20.3178
NT	76.6852	17.6132	11.9138	26.5844	12.1871	37.0070	25.4384	24.5482
NSNT	89.1189	21.1876	12.4761	29.9867	18.2338	42.1779	29.9730	30.2575

Now we turn our attention to focus on Model ST. We notice that not all variables are statistically significant according to the estimated model. In order to extract the important information, we apply the variable selection

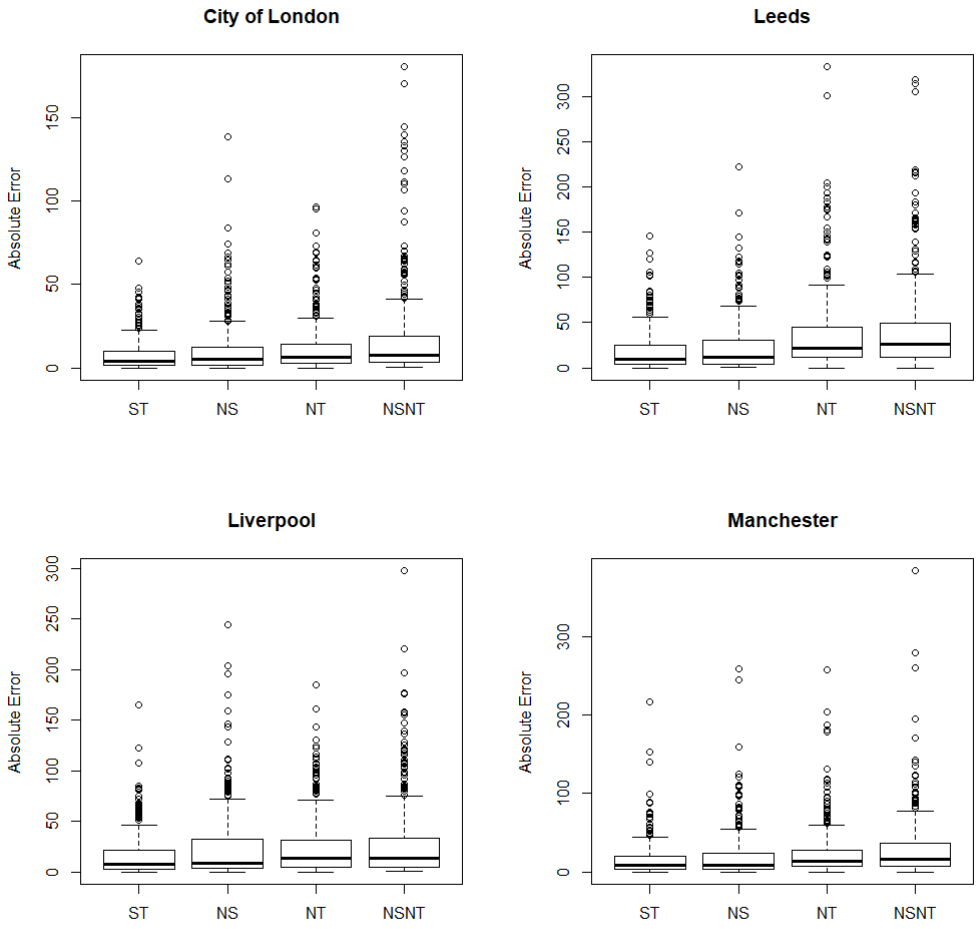

Figure 3: Boxplot of Absolute Error of estimations for (City of) London, Leeds, Liverpool and Manchester of estimations with four different models, namely ST, NS, NT and NSNT.

technique LASSO here (recall Section 5.3.3.2). By solving the Model ST with LASSO (5.13), coefficients of important variables are extracted. The fitted values of 324 days for each location are then depicted in Figure 5 to 8 (We will illustrate the predictions in the last subsection). To make comparison, the fitted value of Model ST (5.3) is also provided. It is clear that the estimations after variable selection are more accurate compared to those without variable selection.

5.4.2 Time trend and lockdown effect

Recall that UK had implemented three lockdowns up to January 2021, as mentioned in the Introduction section. It is of interest to investigate what

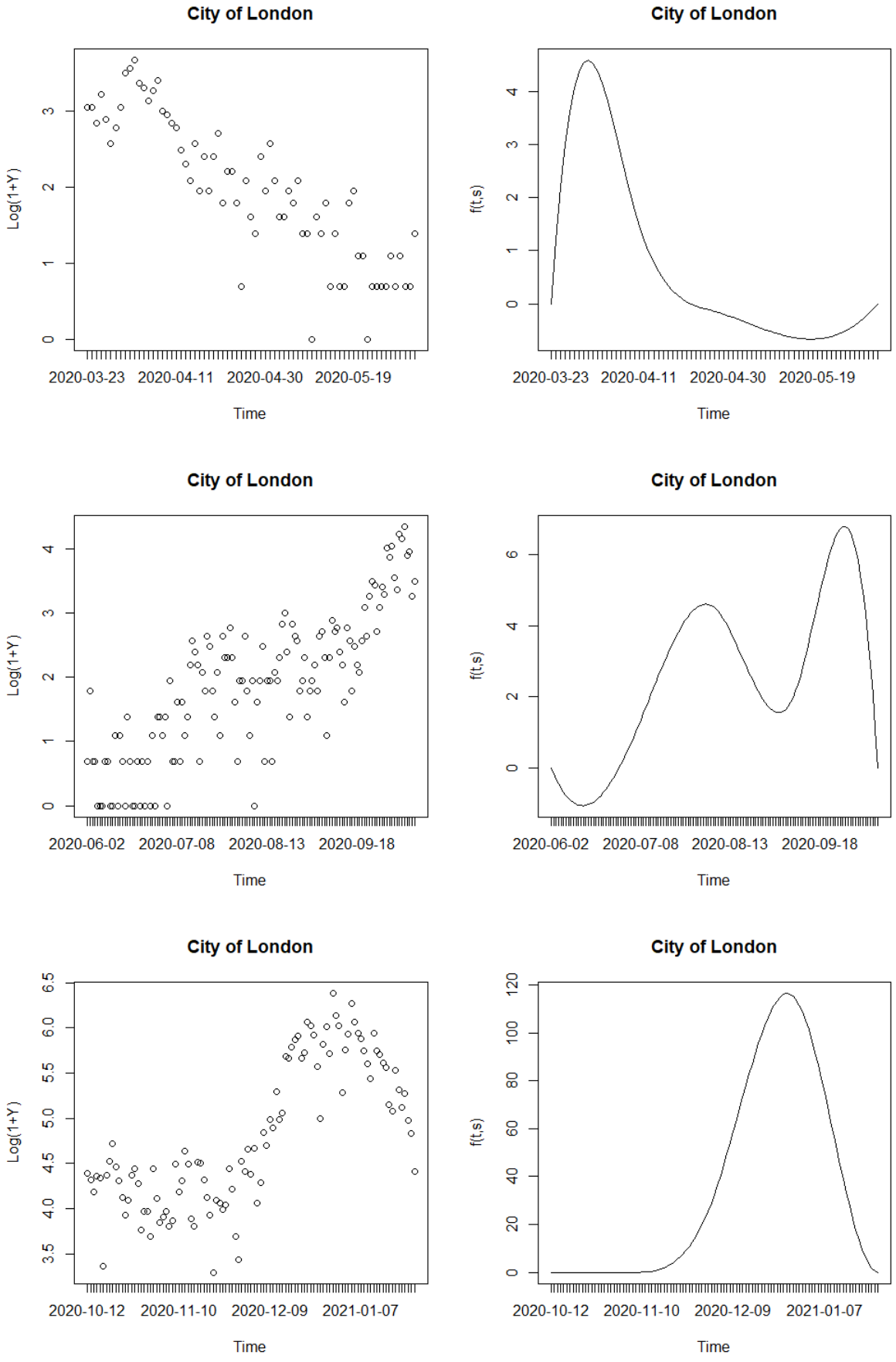

Figure 4: Effects correspond to the first lockdown, the lifting period and the second and third lockdown.The fitted time trend $f(\tilde{t}, s_k)$ are given on right-hand side, which performs very closely to the true pattern given on the left-hand side.

are the impacts of such lockdowns.

Taking City of London for example, we sort the data into three different time periods, which correspond to the first lockdown, the lifting period and the second and third lockdown. We then fit the data to Model ST with LASSO (5.13). This is because we are curious about two things: (1) can we observe different patterns of the time series for different periods of time (that are impacted by the lockdown); (2) and if so, can we capture such time trends with our proposed spline functions. For ease of comparison, we take the log of daily confirmed cases Y plus 1 to avoid the situations where $Y = 0$. The patterns for the three time periods considered are depicted in Figure 4 on the left-hand side, where the estimated time trends from Model ST with LASSO (5.13) are given on the right-hand side. Here we estimate the time trends by summing up the basis (curves) functions given in (5.5) with the coefficients $\mu_l(s_k)$ obtained from Model ST with LASSO (5.13).

According to Figure 4, the nonlinear structure of data is noticeable. The fitted time trend $f(\tilde{t}, s_k)$, which can be viewed as a nonlinear function of time t, performs closely to the true pattern. This, therefore, indicates that the trend term has been captured well, and thus the model remaining (i.e., when deducting the time trend term) can be considered stationary.

In particular, the patterns of trends are very different for all time periods. The daily increase number of the first lockdown period (the first row in Figure 4) quickly reaches its peak and then decrease slowly everyday, which is consistent to the expectations according to epidemiology studies. When the lockdown has been lifted (the second row in Figure 4), we notice a trough between two peaks, which happens around Summer, at which the first lockdown has been (partly) lifted while the second lockdown has not yet been fully implemented. For the last period, the festival of family reunion brings the sharp peak at the end of December.

It now seems that the lockdowns do have an impact on the spread of COVID-19, and therefore result in these different patterns of daily increase number. Moreover, we notice that, the implementation of lockdown is often delayed compared to the outbreak of COVID-19, and policies in different locations, as well as different time periods, would be subjected. It is, as to a macro variable itself, not good enough to provide us more information in details.

Therefore, in this chapter, we use the micro variables introduced in Section 5.2 for further analysis.

5.4.3 Feature variable selection

By the analysis of feature variables with the help of our Model ST with LASSO (5.13), the coefficients of important variables are summarised in Table 5.2 for the time period of 324 days at selected locations. The first six terms $S(t_n)$ are the coefficients assigned to the basis functions of time trends. It is noted that the time trends at different locations varied.

We can now summarise the results as follows:

- **Autoregressive Effects**: The time series of the daily new cases (or death) of an uncontrolled epidemic are roughly symmetric and bell-shaped (Farr, 1840), as further adopted in Ferguson et al. (2020). This gives an intuition that the temporal effect may have a cluster effect that the number observed for tomorrow is likely to be high, if it is high today, but not likely if it was high a month ago. As expected, we do observe this property by finding the negative coefficients for medium and large time lags as indicated from Table 5.2. For Cardiff, such negative relationships are also identified in small time lags, however, with the coefficients being close to zero.

  Indeed, different cities have different significant time lags as seen from Table 5.2. This may be a result of the incubation period and diagnosis as it is now known that it takes on average 5 days (range 1-11 days and the maximum is 14 days) for the patient to show symptoms and then it may take some time for him or her to be treated and confirmed by NHS. It may therefore also be affected by the availability of local healthcare resources as well as the local level of COVID-19 outbreak. For example, in Edinburgh, the daily increase number depends on its past values of 1-2, 5-12, 14-15 and 17 days ago, while in City of London, which is a small district of the Greater London, it is sensitive to a different and smaller range of time lags, including 1-2, 3, 6-7 and 21 days ago.

- **Spatial Effects**: It is expected that both the transportation volumes and origins of these cities are different. Though intuitively one would expect a similar result (e.g., cluster effect) as observed for temporal

Table 5.2: Model ST with variable selection for selected local authorities

	Birmingham	Cardiff	Edinburgh	Glasgow	London	Leeds	Liverpool	Manchester
$S_1(tn)$	-82.04113	-24.02168	-10.26135	-55.42546				
$S_2(tn)$	-33.20170	-7.89398	-12.20415				2.21372	
$S_3(tn)$		-31.40977	-21.36209	-3.63540		-0.46822	-12.42109	
$S_4(tn)$		27.94630	36.18937	22.10380		80.92337	103.75950	-3.60122
$S_5(tn)$	222.80925	99.51638	22.52940	36.25856	76.06458	0.45148	10.18499	74.23492
$S_6(tn)$	2.35191		12.43009	36.26249		27.45319	73.83587	
Y_{t-1}	0.25647	0.57550	0.36679	0.61484	0.23629	0.57354	0.39210	0.19412
Y_{t-2}	0.10967	0.08178	0.18025	0.06174	0.09187	0.14452	0.00276	0.13729
Y_{t-3}		0.07724		0.00841				0.00014
Y_{t-4}	0.02528	-0.00942		0.10928	0.07447	0.12908		0.08377
Y_{t-5}			0.05359					
Y_{t-6}	0.18616	0.12310	0.11318		0.00571	0.01359	0.12396	
Y_{t-7}	0.03578	0.15177	0.30046	0.15940	0.19434	0.16837		0.18703
Y_{t-8}		-0.00292	-0.00983			-0.18532	-0.00073	0.01486
Y_{t-9}		-0.10498	-0.02723	-0.04720		-0.00374		
Y_{t-10}		0.02029	-0.00852	-0.01925		-0.07322		
Y_{t-11}		-0.02543	-0.12234					
Y_{t-12}			-0.00899				0.09052	
Y_{t-13}		0.05893				0.01521	0.03120	0.00007
Y_{t-14}		-0.06501	0.03537	0.01262		0.15802		
Y_{t-15}		-0.05861	-0.00996			-0.11750	-0.02168	
Y_{t-16}		-0.02732					-0.02137	
Y_{t-17}			0.00064					
Y_{t-18}								
Y_{t-19}				-0.00319				
Y_{t-20}		-0.11038		-0.03170				
Y_{t-21}	-0.00589				-0.05104			-0.07083
News Index			0.00461	0.01501			0.00305	-0.00001
Driving		-0.04490		-0.89792		-0.28084	-0.17296	
Transit	0.47460	0.31224	0.02869	1.04618	0.02502	0.62448	0.65237	0.30250
Walking	-0.35153	-0.06937		0.09292		-0.17927	-0.33310	-0.15872
Retail&Receation		0.07281		0.00181	0.16486			0.00037
Grocery&Pharmacy	-2.16797	-1.21571	-0.52264	-1.16849	-0.68130	-0.65893	-0.72166	-0.84669
Parks	0.02147	0.10101	0.07058	0.23340		0.09713	0.09830	0.01510
Transit Stations				-0.03511				0.14548
Workplace		0.05348	0.00895		0.07890	0.09913	0.07025	0.15079
Residential			-0.07994	-0.28307			-0.02893	
Y^S_{t-1}	1.88134		0.16000	0.32715	0.35654		1.26303	0.40870
Y^S_{t-2}		0.00505						
Y^S_{t-3}				0.31705			0.05630	
Y^S_{t-4}		-0.09938	-0.13002					
Y^S_{t-5}		-0.00861	-0.01192		0.04338		-0.53373	0.03793
Y^S_{t-6}				0.68604	0.18276	0.01043	0.20813	0.20967
Y^S_{t-7}	1.73824	0.41208	0.15347	0.20106		0.87037	1.88149	
Y^S_{t-8}			-0.21439	-0.92061		-0.00121	-0.95371	
Y^S_{t-9}		-0.01280	-0.10891				-0.20787	
Y^S_{t-10}	-0.72425	-0.42295	-0.08480	-0.61870	-0.23529	-0.24593	-0.13900	-0.31148
Y^S_{t-11}		-0.06478		-0.13365		-0.00005	-0.12930	
Y^S_{t-12}				0.01743				
Y^S_{t-13}						0.00018	0.01002	0.00019
Y^S_{t-14}	0.08959	-0.00003					-0.50969	
Y^S_{t-15}				0.00011				
Y^S_{t-16}			-0.05583			-0.38683		
Y^S_{t-17}	-0.42047						0.00609	
Y^S_{t-18}	-0.35707				-0.01812	-0.04056		-0.00015
Y^S_{t-19}		0.02058					-0.29319	
Y^S_{t-20}	-0.65820		0.00186				0.00276	-0.00009
Y^S_{t-21}			-0.00703				-0.00241	

effects, we now observe from Table 5.2 many negative coefficients for all ranges of time lags. For example, there are 4-5 day negative lag effects for cities of Cardiff, Edinburgh and Liverpool, while longer day negative lag effects are observed for other cities. This may, perhaps, be partially credited to the increasing of self-protection awareness boosted by the bloom of epidemic in nearby areas. However, for transportation hubs, e.g., Birmingham and Liverpool, the main spatial effects of COVID-19 are positive and stronger than other areas, with lag-1 coefficients as

large as 1.88134 and 1.26303. Here we would like to call for further research to investigate such spatial neighbouring effects with links to social studies, transportation research and epidemiology.

- **Google data - Mobility in Areas**: Our finding generally agrees with the common sense that virus would spread with close contacts especially in the areas where people are crowed with no fresh air. For example, from Table 5.2, we find the daily increase number positively depends on mobility in retail areas for Cardiff, Glasgow, (City of) London and Manchester. However, no further evidences are observed for other cities. Similarly, such positive effects are observed in work place. As to the residential areas, staying at home is shown to help reducing the increase of daily confirmed cases in cities, such as Edinburgh, Glasgow and Liverpool. It is expected that for different locations, the living style of residents would also lead to some kinds of variety, which may call for future study of such social links.

  It is not very surprising to see from Table 5.2 that mobility in groceries and pharmacies at all locations would actually decrease the number of infections, as people are required to shop there with medical equipments, e.g., masks, for protection. The access to pharmacies also provides people the chance to get such medical equipments, and thus help preventing the spread of COVID-19.

  For the mobility data of 'Parks', the coefficients are positive except City of London as shown in Table 5.2. It looks surprising. This has now been explained by updated medical research that the virus could still be alive and therefore spread in the air (c.f., The Lancet Respiratory Medicine (2020)).

  Surprisingly, for mobility at transit stations, the coefficient is mildly negative for Glasgow. On one hand, the reduce of mobility in transit stations may lead to the increase of mobility in other areas. On the other hand, it could be actually a result of lockdown that people use public transportation more often when the spread of COVID-19 is under control and avoid it when the situation is bad.

- **Apple data - Mobility by transport**: As expected, from Table 5.2, we do observe significantly positive relationships for 'Transit' from Apple Data. Especially, in the areas like Glasgow, such relationship is

identified strongly positively correlated with the coefficient for 'Transit' as large as nearly 1.05. It shows consistent impacts on the daily increase number that it is still very risky for people to take public transportation in general. We thus would like to suggest limiting the mobility of public transportation as the first priority.

The effect of the 'walking' from Apple data, mostly believed by people to be safe as it is an open space with fresh air, is shown in Table 5.2 to be negatively correlated to the number of daily new cases in most cities except Edinburgh, Glasgow and City of London. Note that it is positive for 'walking' in City of Glasgow.

Driving , according to the results, seems to be safe and can even contribute to reduce the spread of COVID-19. This can be understood as it would reduce the chance of people using public transportations, and avoid contacts.

- **UK News Index**: The Index value is positively correlated to the spread of COVID-19 for locations like Edinburgh, Glasgow and Liverpool, mainly in the northern part of England and in Scotland. These cities with large populations have witnessed the serious local outbreak of COVID-19 and local lockdown, which should in general be reflected in the News index. However, we observe a negative coefficient with News index for Manchester, which is close to 0. To better study the impact of media promotion of self-protection awareness in the combat of epidemic of COVID-19, more detailed data other than UK News Index is needed. Research in this direction may provide further information and deeper understanding that guide the government to better react to these emerging events.

5.4.4 Forecasting comparison

In order to further examine the advantage of the feature variable selection technique, we apply both Model ST (5.3) and Model ST with variable selection (5.13) to predict the number of daily new cases for the last available week in the data, i.e., from 7th March 2020 to 24th January 2021. This is done by a one-step ahead prediction for 7 days in total, i.e., from 25th January 2021 to 31st January 2021.

The prediction with the variables other than time trend is rather straightforward (as they are accessible in data). However, for the time trend term, we need to first forecast its value before calculating the predicted number of confirmed cases. It is argued that the spline method only uses one side of past time points when forecasting, e.g., at $t = n + 1$, so the extrapolation of the future may not work well. However, for a relatively short term period, say next day, it is fair to assume that the current time trend would last. Recalling Section 5.3.2.1, we use cubic spline to first obtain the time trend at time n, and then feed its value to the model for prediction at time $(n+1)$. We can thus make prediction for time $(n + 1)$ from our models.

We calculate the MAE of predictions based on model 5.13 for each location and present it in Table 5.3 (see also Figure 5 to 8). It is noted that some of the values predicted by the Model ST without variable selection would actually lay outside of the plot, i.e., exceeding the maximum of y-axis in the figures. It is clear that Model ST with variable selection has much smaller MAE values of the predictions, and thus a stronger prediction power, for all cities indicated in Table 5.3. For Birmingham, it has the largest number of accumulated confirmed cases, and the trends there are rather volatile. Clearly, Model ST without variable selection fails to capture the true pattern well and is impacted by the outliers instead, as seen in Figure 5. It is also confirmed by Figures 5 to 8 that Model ST without variable selection often overestimates the daily confirmed cases in prediction.

Finally we make a comment before ending this subsection. We turn our attention back to Table 5.2 for explaining the predictions given in Table 5.3. The obtained coefficients for Model ST with variable selection in Table 5.2 indicate that the time trend term actually plays an important role in explaining the daily number of new cases. For some locations, e.g., Cardiff, City of London, Leeds and etc., not every basis function of time trends is selected. In fact, as one can observe from Figures 5 to 8, the time trend patterns of daily new cases are quite different at different locations. This agrees with the epidemiology consensus that the investigations of COVID-19 need to be done on a location to location basis.

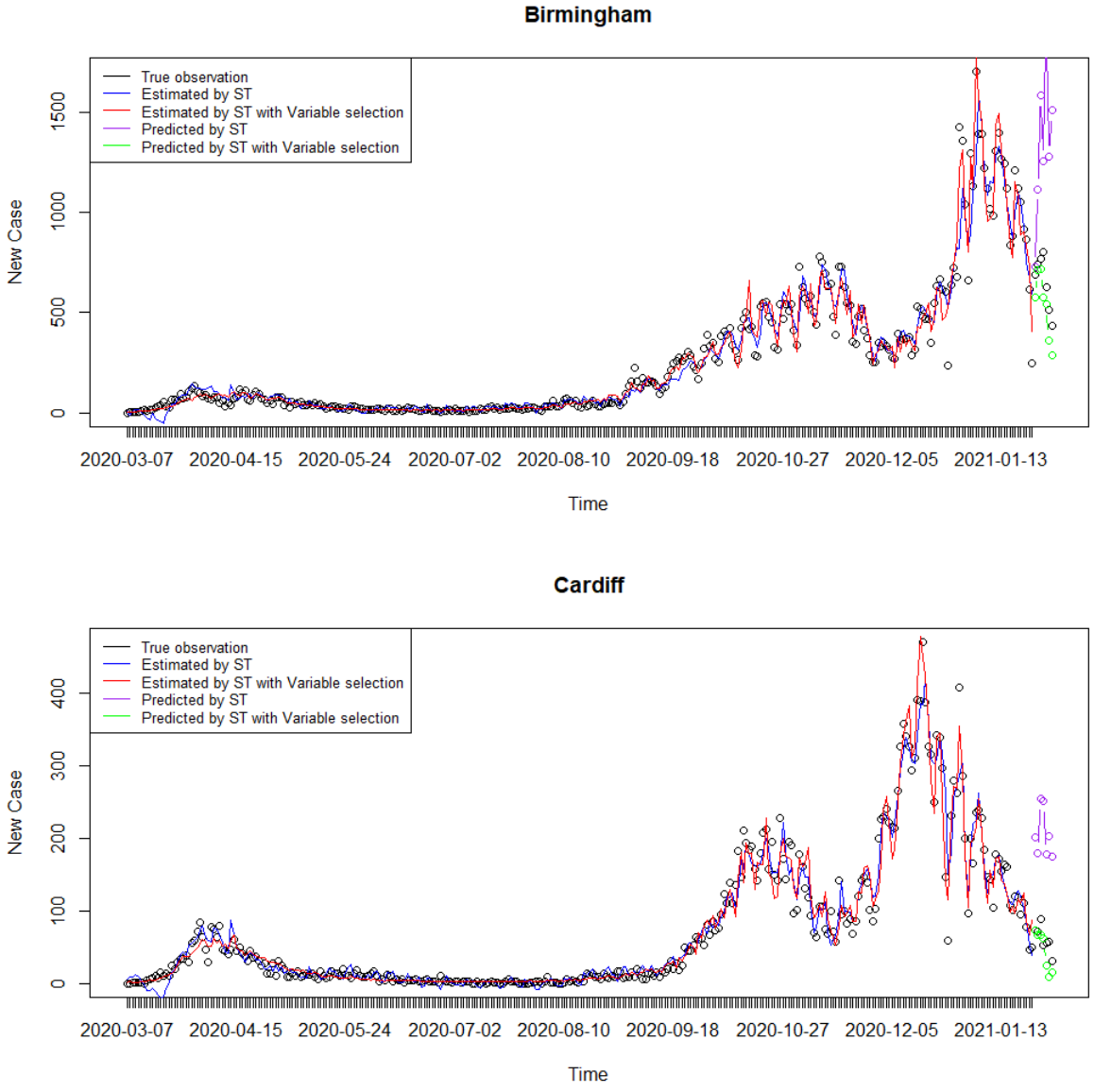

Figure 5: Estimation and Prediction of Model ST for Birmingham and Cardiff

Table 5.3: Mean Absolute Errors of predictions of 7 days by Model ST with/without variable selection.

	Birmingham	Cardiff	Edinburgh	Glasgow	City of London	Leeds	Liverpool	Manchester
ST without variable selection	682.5164	144.4767	223.6387	3534.29	220.9273	349.0468	713.9196	189.745
ST with variable selection	114.3103	18.3848	5.9312	25.3739	12.4171	19.9994	26.0801	17.2645

5.4.5 Implications

From the above analyses, we believe the proposed model is acceptable in accuracy and robustness for analysing the UK COVID-19 data. The time trends of COVID-19 are not negligible across different time periods, though without interventions, the development of pandemic is roughly bell shaped. Also, the empirical findings reveal that the development of COVID-19 in each city depends on its role in the national (or international) transportation network. As a consequence, lockdown is the effective action to limit the neighbouring effects identified. However each local authority should implement individual policies based on its own status. For example, it may be beneficial to keep the access to Grocery & Pharmacy for the public. Indeed, the key of lockdown is not to limit the activity of people to certain areas, but to limit the contacts and promote self-awareness of prevention. Our findings therefore support the epidemic consensus that the limitation of public transportation is the most important action from the perspective of pandemic control.

5.5 Conclusion

In this chapter, we have presented a spatio-temporal model with the consideration of nonlinear trends, temporal lag and spatial neighbouring effects of the daily cases as well as the mobility effects for the empirical modelling of the COVID-19 data in the UK. The key findings include:

(1) The daily confirmed number has strong time trends across different time periods of interventions. It also has interesting strong temporal lag effects, with strongly positive lag effect lasting for about one week in the past but with a self-recovering of negative lag effect after more than one week up to 3 weeks (in particular 3 weeks lag effect identified for Birmingham, City of London and Manchester).

(2) A neighbouring effect is also identified that the areas of key role in a transportation network often suffer more serious infection of COVID-19, and thus the action of lockdown is indeed an effective measure in the combat of pandemic.

(3) The media effects are significant, which may well promote self-protection awareness in controlling the spread of pandemic. The mobility data classified by either areas or means of travel can provide further guidance on how to implement the multi-level lockdown for different locations.

(4) We demonstrate that all these effects identified are however varied with locations, so different local authorities may need to implement varied policies regarding control measures such as lockdown, not only because of the varied population size and medical resources, etc., but also depending on their roles in a transportation network.

These findings imply that limiting public transportation and using media advertisement of COVID-19 information to promote the publics awareness of self-protection (e.g., wearing masks) should be of the first priority in the sense of preventing the spread of this pandemic, in particular in view of the one-week strongly positive lag effect. We need to pay more attention to public transportation rather than walking or driving. The increase of mobility to pharmacies actually reflects the awareness of individual protection, and thus can help limit the spread of COVID-19. We can expect these findings and the suggested methods will be useful in guiding and supporting the policy making and allocation of the resources that are location based for pandemic controls.

We comment that our modelling is based on the conditional Poisson distribution of the daily new cases given the past available information. It means that unconditional distribution of the daily new cases is a mixture of Poisson, or a non-Poisson, distribution, so our model distribution is in general reasonable. However, study of spatio-temporal dynamic behaviour of the daily new cases under other model distributions such as negative binomial (c.f., Giuliani et al. (2020)) could also be further investigated, which is left for future consideration.

The investigation into COVID-19 is still an active area for many disciplines. We believe both the methods and the findings in this chapter thus can further contribute to the related studies.

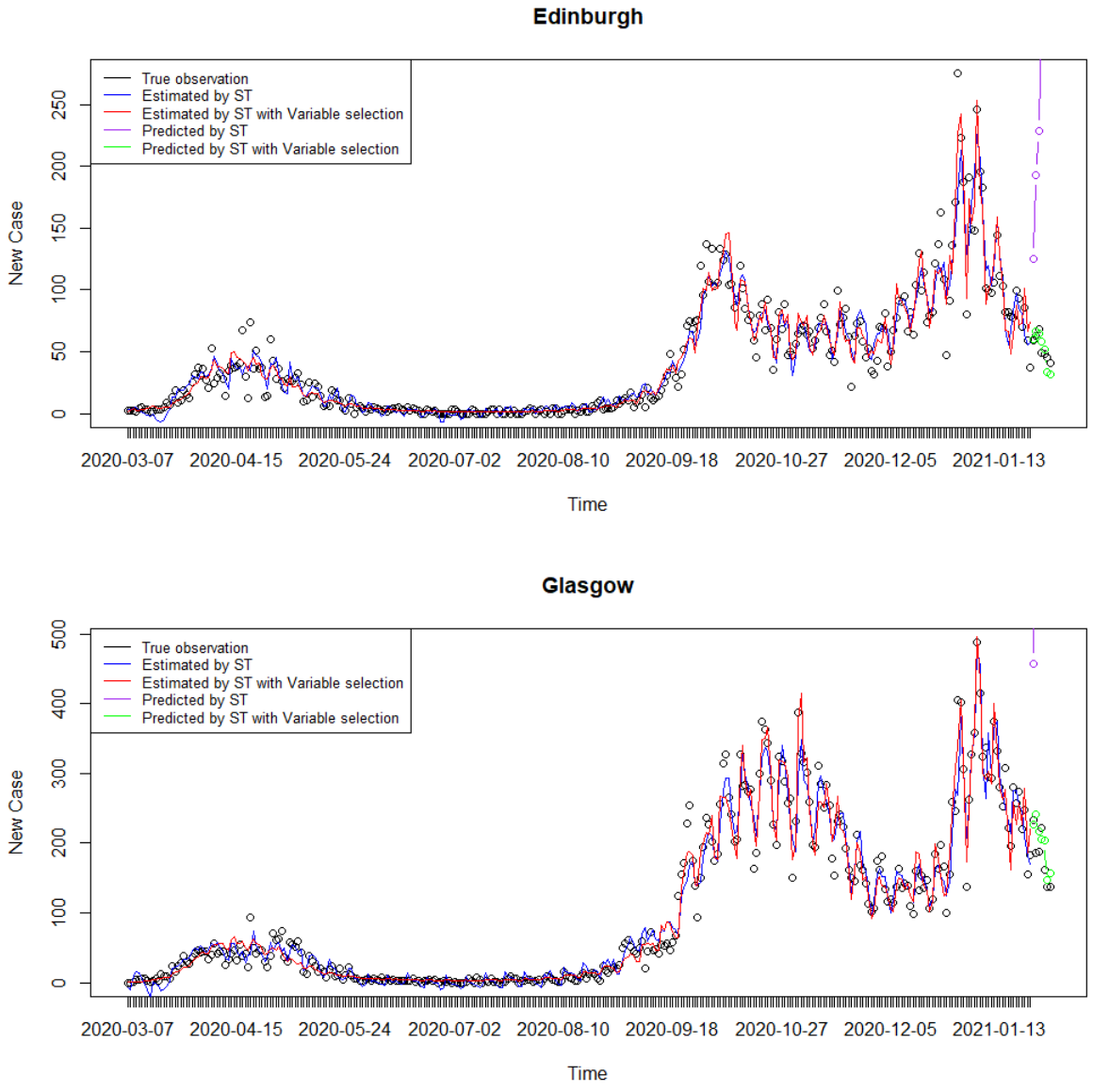

Figure 6: Estimation and Prediction of Model ST for Edinburgh and Glasgow

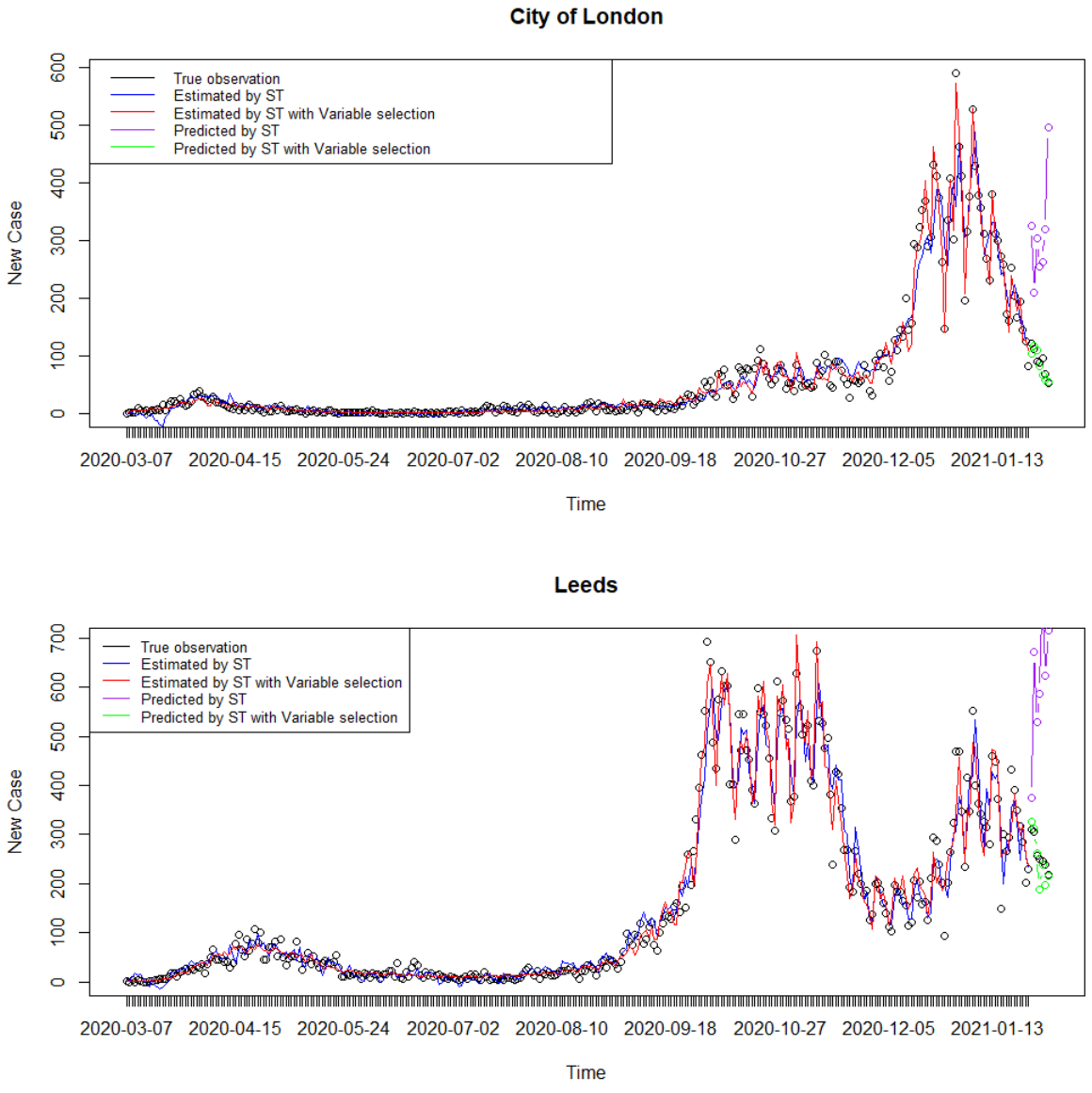

Figure 7: Estimation and Prediction of Model ST for (city of) London and Leeds

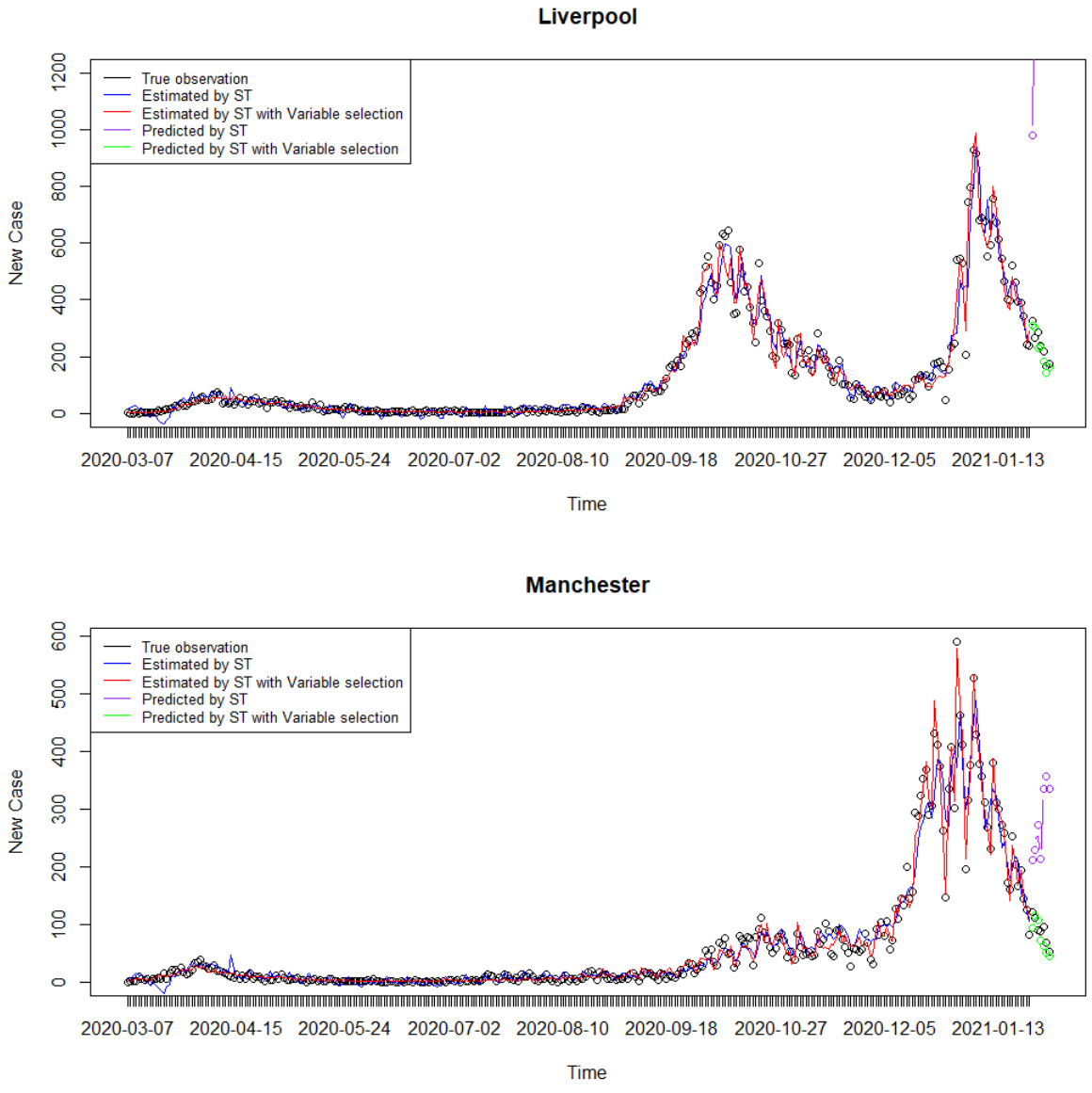

Figure 8: Estimation and Prediction of Model ST for Liverpool and Manchester

Chapter 6

Conclusion, Challenges and Future Work

6.1 Conclusion

In this research, we have proposed several novel estimation procedures for the discrete-valued time series data. The theoretical works of asymptotic properties contribute to the literature of high and ultra-high dimensions time series data of discrete-valued response and maximum likelihood estimations. The numerical examples of empirical applications have shown the great robustness and prediction ability of the proposed procedures for dealing real world problems. We thus demonstrate the main contributions as follows:

- A class of discrete-valued time series models have been developed with flexibility to incorporate also exogenous variables that are easy to interpret and compute.
- To cope with the unknown nature of data, we consider nonlinear structure instead of commonly adopted linear assumptions. This makes sure that the models are "data-driven", which is important for situations where human experiences are absent.
- Conventional techniques, such as local linear regression (see Chapter 2) and adapted LASSO (see Chapter 4), are adapted from i.i.d assumption to β-mixing conditions. It further allows researchers in the filed of time series to adopt these classic methods in related studies.

- "Curse of dimensionality", which is often faced in large scale data problems, has been avoided by the proposed semi-parametric model GMAMaR. Therefore, it provides the solid foundation for further applications of our proposed models to real world problems, e.g., in the field of epidemiology, finance and etc.

6.2 Challenges and future works

Though the models developed in this thesis are proved to be elegant and efficient theoretically, in practice, there are still some challenges to be solved:

(a) We have examined the impact of spatial effects in a simple model given in Chapter 5, which shows great potentials. A challenge is the spatial effect caused by the economic globalisation that countries are coming together as one large economy to make international trading easier. This is to say, the possible dependency of space has to be considered if we want to estimate or to predict the complex responses such as default rate and market evolution probability.

(b) As partly examined in this thesis, the developed models are robust and powerful, when dealing with discrete time series data in high and ultra-high dimension. Therefore, another possibility is to apply these models to other practical applications such as credit scoring.

In the next stage of research, we would like to extend our novel estimations procedures for discrete-valued β-mixing time series data to a broader area that could possibly solve the challenges mentioned above.

6.2.1 Spatial-temporal modelling

Empirical studies in econometrics and statistics often find spatio-correlations in time series data, due to the development of globalisation. For instance, they are in wide range of disciplines, such as environmental economics and industrial organisation economics. We aim to extend our proposed GMAMaR method to the domain of spatio-temporal models. For example, the logistic spatio-temporal model will have a form as follows:

$$logit(p(s,t)) = c(s_k) + b(s,t)x(s,t) + \mathcal{W}(s,t), \tag{6.1}$$

where $p(s,t) = P(Y(s,t) = 1|I_{t-1})$, s is the index of location and t is the index of time, $b(s,t)$ is the coefficient, $\mathcal{W}$ is a spatial-temporal process and $c(s_k)$ is the intercept term.

Unlike t, which is one dimension, the location can have different directions and it depends on how we measure it. There are two different methods found in literature. The first one is to assume that s can go through every location. This is the case where continuous x-axis and y-axis coordinates are assigned with s, e.g., in Al-Sulami et al. (2017). The other way is to assume the spatial sites are of the form of countable lattices, e.g., in Besag (1974) and Lee (2004). This can be considered as a simplified case, where the directions are often restricted to, e.g., north, south, west and east.

The linear relationship of spatial(-temporal) models has been widely studied in literature, see Cox and Isham (1988), Yang et al. (2005) and Cressie and Wikle (2015). In recent years, we notice an increasing interest of the investigation of the nonlinear structure Xu and Lee (2015). For instance, Lu and Chen (2004), Hallin et al. (2004) and Gao et al. (2006) study some statistical properties of a spatial regression under mixing conditions. Jenish (2012) establishes asymptotic properties of local linear estimators under spatial near-epoch dependence (NED). It is recognised that allowing nonlinear relationships can improve the performance of models as it better captures the true relationship of the data. Among the spatial-temporal models, the study of discrete-valued dependent variable is not as extensive as the continuous-valued cases. (c.f., Zhu et al. (2008), Wang et al. (2013) and Cressie and Wikle (2015).)

In the context of discrete-valued time series modelling procedure, the extension to spatio-temporal models would require careful checks of the assumptions and novel designs of estimation methods. Here we consider a situation given as follows.

Let $Y_t(s_k)$ and $X_t(s_k)$ denote two spatio-temporal processes at discrete time point $t = 1, ..., n$, where $Y_t(s_k)$ is binary taking on value of 0 or 1, and $X_t(s_k)$ contains d dimensional covariate variables, which may involve the spatial-temporal lag and variables of different time series data. A spatial unit is defined as $s_k := (u_k, v_k) \in \mathbf{R}^2$, where u_k, v_k are x and y coordinates

at discrete time point $t = 1, .., n$ and location index $k = 1, ..., r$. At a given spatial location s_k, we consider a semi-parametric nonlinear regression time series model as follows.

Denote $I_{t-1}(s_k)$ for the information up to time $t-1$ about time series $Y_t(s_k)$. The regression is to estimate the following conditional probability:

$$p_t(s_k) = P(Y_t(s_k) = 1|I_{t-1}(s_k)). \tag{6.2}$$

Because of the curse of dimensionality, we would like to first look at the marginal effects of each covariate. Here we define the marginal probability of the jth covariate $(x_{jt}(s_k))$ as follow:

$$p_{jt}(s_k) = P(Y_t(s_k) = 1|x_{jt}(s_k)), j = 1, ..., d; k = 1.., r. \tag{6.3}$$

Let F be the logistic cumulative distribution function(c.d.f), i.e., $F(u) = \frac{e^u}{1+e^u}$. Then the marginal non-parametric logistic regression is $logit(p_{jt}(s_k)) = f_j(x_{jt}(s_k), s_k)$, and therefore, we have:

$$p_{jt}(s_k) = F(f_j(x_{jt}(s_k), s_k)), \tag{6.4}$$

where $f_j(.)$ is the j-th marginal smooth function.

Our second step is then to combine the marginal logistic regressions by using the idea of model averaging as follows:

$$\begin{aligned} logit(p_t(s_k)) = c(s_k) + \Big[\alpha_1(s_k), \cdots, \alpha_d(s_k)\Big] \begin{bmatrix} logit(p_{1t}(s_k)) \\ \vdots \\ logit(p_{dt}(s_k)) \end{bmatrix} \\ + \gamma_1(s_k)\boldsymbol{\omega}(s_k) \begin{bmatrix} Y_{t-1}(s_1) \\ \vdots \\ Y_{t-1}(s_r) \end{bmatrix} + ... + \gamma_q(s_k)\boldsymbol{\omega}(s_k) \begin{bmatrix} Y_{t-q}(s_1) \\ \vdots \\ Y_{t-q}(s_r) \end{bmatrix} \end{aligned} \tag{6.5}$$

where $\boldsymbol{\alpha}(s_k) = (\alpha_1(s_k), ..., \alpha_d(s_k))$ and $\boldsymbol{\gamma}(s_k) = (\gamma_1(s_k), ..., \gamma_q(s_k))$ are the vectors of coefficients to be estimated. This can be seen as the model averaging. The $\boldsymbol{\alpha}$ can be seen as the temporal weights assigned to temporal models and $\boldsymbol{\gamma}$ can be seen as the spatial weights of spatial models.

$\boldsymbol{\omega}(s_k) = (\omega_1(s_k), \ldots, \omega_r(s_k))$ is the corresponding spatio weight vector given location s_k, which is often assumed to be a priori is popular in econometrics (c.f., Al-Sulami et al. (2017)). $c(s_k)$ is the constant term that will be estimated simultaneously, which reflects the long term average of $Y_t(s_k)$.

This is an example of logistic model with the asymptotic properties of proposed estimation method left for future work. We hope the development of such model for discrete-valued time-series of exponential family can potentially contribute both in theoretical and practical way by providing the uniform consistency properties and showing accurate estimation and prediction results for solving real world problems. For instance, an application to COVID-19 data, as what we have studied in Chapter 5, can be expected with the new model designed.

6.2.2 Credit scoring

As to the existing time series models, an application to credit scoring appears to be of great potential. Credit scoring is an important method widely used in todays business where a lender (or such a decision maker) would decide whether or how to offer credits to the borrower (or a consumer specified). One of the most critical processes in a credit decision system is to evaluate the credit. To assess such decision, it is necessary to collect, analyses and classify different credit attributes. This process to evaluate the credit, aiming to minimise the expected loss of the loan defaults, is normally referred to as credit scoring. It is also said to be the modelling of assessing creditworthiness by Hand and Jacka (1998).

The decision maker normally would have two questions: first, how to deal with new applicants and then, how to cope with existing customers. This would include questions such as "shall we accept a new application?" or "how to decide their credit limits (could be increase or decrease)". No matter what kinds of models are used to answer the listed question, it is crucial to have a big data of previous applicants with associated detailed information and significant credit history. Credit scoring models all use the given sample to investigate the connection between the attributes of the consumer and the subsequent performance in history. Each attribute would be given a score and the total score summed up describes the risk of the particular consumer and whether it is too bad to accept.

Table 6.1: Statistical Models for Credit Scoring

Methods	Main technique	Summary
Discriminant analysis	Decision theory	Use the conditional probability to minimize the expected loss of default
Discriminant analysis	Classify two groups	Classify cases into groups, by drawing the perpendicular cut-off line
Discriminant analysis	Linear regression	Find the best linear combination of variables to minimize the mean square error
Logistic regression	Maximum likelihood estimation	Use the MLE method to find the best coefficients of the log transferred regression
Probit and tobit analysis	Nonlinear regression	Use individual approach to minimize the total sum of error
Classification tree	Recursive partitioning algorithm	Uses decision tree to maximise differences between sub-sets
Nearest-neighbour	Nonparametric	Choose metric to find the nearest neighbours

Table 6.2: Non-statistical Models for Credit Scoring

Methods	Summary
Linear programming	Minimize the sum of absolute errors (MSAE) or maximum error (MME)
Integer programming	Minimize the number of misclassification or the total cost associated
Neural network	Minimize the average mean of all trained samples
Genetic algorithm	Search the set of possible solutions and find out the local or global optimum

An increasing number of literatures in the last few years indicates that the research and application of credit scoring is developing rapidly. The most up to data review paper is proposed by Lessmann et al. (2015) that aims to give a benchmark to the classification algorithms using credit data as an update of their previous work of Baesens et al. (2003). Crook et al. (2007) earlier had introduced the history and its following development of credit scoring techniques with a comprehensive review. All of their work are based on the ground of previous research such as a review of different classification models given by Rosenberg and Gleit (1994), the result comparison review done by Hand and Henley (1997) and the additional introduction of behaviour scoring techniques proposed by Thomas (2000) as well as a further discussion of Thomas et al. (2002) on this topic. Abdou and Pointon (2011) then provided a review 214 related publications as a continued work of Thomas et al. (2002).

According to these reviews, the credit scoring models could be divided into statistical scoring methods, non-statistical scoring methods and the repayment behaviour methods. The representatives of them are linear/ logistic regression, mathematical programming and survival analysis respectively, see also Table 6.1 and 6.2.

Except for standard methods mentioned above, research of credit scoring has been carried on with many newly developed techniques. These new models are aimed to answer the question such as "how to score with limited sample set", "can the conventional classifiers be combined", "what if there exists more than one scorecard", "what if one estimates the inter-media variables before making a binary decision" and "how about predict when would default happen rather than whether it will happen or not". All the mentioned

questions being investigated lead the proposed techniques to include more outcome variables and allow them to be continuous rather than traditional answer as yes or no.

Such trend shows an increasing interest in the study of credit scoring in a more complex environment in a timely manner, which could hardly be explained or included using the existing statistical and non-statistical models such as logistic regression. Since credit scoring also reflects the lender's prediction of the future economic scenario, one may consider to include the economic environment and other important variables related. These financial data are often needed to be analysed using time-series analysis both for predicting the future and study the underlying laws of economic. Hence, there is a potential to combine and investigate the credit scoring methods and time series analysis methods. As a consequence, it would ideal to apply our proposed discrete-valued time series models (e.g., MAMALOR) into the field of credit scoring.

References

Abdou, H. A. and Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: a review of the literature. *Intelligent systems in accounting, finance and management*, 18(2-3):59–88.

Aisyah, D. N., Mayadewi, C. A., Diva, H., Kozlakidis, Z., and Adisasmito, W. (2020). A spatial-temporal description of the sars-cov-2 infections in indonesia during the first six months of outbreak. *PloS one*, 15(12):e0243703.

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

Al-Sulami, D., Jiang, Z., Lu, Z., and Zhu, J. (2019). On a semiparametric data-driven nonlinear model with penalized spatio-temporal lag interactions. *Journal of Time Series Analysis*, 40(3):327–342.

Anselin, L. (2013). *Spatial econometrics: methods and models*, volume 4. Springer Science & Business Media.

Apple (2021). Mobility Trends Reports. https://covid19.apple.com/mobility.

Avery, C., Bossert, W., Clark, A., Ellison, G., and Ellison, S. F. (2020). An economist's guide to epidemiology models of infectious disease. *Journal of Economic Perspectives*, 34(4):79–104.

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., and Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the operational research society*, 54(6):627–635.

Ballings, M., Van den Poel, D., Hespeels, N., and Gryp, R. (2015). Evaluating multiple classifiers for stock price direction prediction. *Expert Systems with Applications*, 42(20):7046–7056.

Barrios, J. M. and Hochberg, Y. (2020). Risk perception through the lens of politics in the time of the covid-19 pandemic. Technical report, National Bureau of Economic Research.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, 31(3):307–327.

Bosq, D. (2012). *Nonparametric statistics for stochastic processes: estimation and prediction*, volume 110. Springer Science & Business Media.

Box, G. E., Jenkins, G. M., and Reinsel, G. (1970). Time series analysis: forecasting and control holden-day san francisco. *BoxTime Series Analysis: Forecasting and Control Holden Day1970*.

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

Bradley, R. C. (2005). Basic properties of strong mixing conditions. a survey and some open questions. *arXiv preprint math/0511078*.

Breiman, L. et al. (1996). Heuristics of instability and stabilization in model selection. *The annals of statistics*, 24(6):2350–2383.

Brown, R. G. (2004). *Smoothing, forecasting and prediction of discrete time series*. Courier Corporation.

Buckley, D. and Bulger, D. (2012). Trends and weekly and seasonal cycles in the rate of errors in the clinical management of hospitalized patients. *Chronobiology international*, 29(7):947–954.

Carroll, R. J., Ruppert, D., and Welsh, A. H. (1997). *Nonparametric estimation via local estimating equations, with applications to nutrition calibration*. Humboldt-Universität zu Berlin, Wirtschaftswissenschaftliche Fakultät.

Chen, J., Li, D., Linton, O., and Lu, Z. (2016). Semiparametric dynamic portfolio choice with multiple conditioning variables. *Journal of Econometrics*, 194(2):309–318.

Chen, J., Li, D., Linton, O., and Lu, Z. (2018). Semiparametric ultra-high dimensional model averaging of nonlinear dynamic time series. *Journal of the American Statistical Association*, 113(522):919–932.

Coppola, A. and Stewart, Brandon M. Okazaki, N. (2014). *lbfgs: Efficient L-BFGS and OWL-QN Optimization in R.* R package version 1.2.1.

Cox, D. R. and Isham, V. (1988). A simple spatial-temporal model of rainfall. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, 415(1849):317–328.

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

Cressie, N. and Wikle, C. K. (2015). *Statistics for spatio-temporal data*. John Wiley & Sons.

Crook, J. N., Edelman, D. B., and Thomas, L. C. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research*, 183(3):1447–1465.

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

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

Davis, R. A. and Wu, R. (2009). A negative binomial model for time series of counts. *Biometrika*, 96(3):735–749.

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

DeLeeuw, J. (1992). Introduction to akaike (1973) information theory and an extension of the maximum likelihood principle. In *Breakthroughs in statistics*, pages 599–609. Springer.

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

EconomicPolicyUncertainty (2021). UK Daily News Index. http://www.policyuncertainty.com/uk˙daily.html.

Fan, J., Farmen, M., and Gijbels, I. (1998a). Local maximum likelihood estimation and inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(3):591–608.

Fan, J. and Gijbels, I. (1995). Adaptive order polynomial fitting: bandwidth robustification and bias reduction. *Journal of Computational and Graphical Statistics*, 4(3):213–227.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, volume 66. CRC Press.

Fan, J., Härdle, W., and Mammen, E. (1998b). Direct estimation of low-dimensional components in additive models. *The Annals of Statistics*, 26(3):943–971.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

Fan, J. and Yao, Q. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika*, 85(3):645–660.

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

Fan, J., Yao, Q., and Cai, Z. (2003). Adaptive varying-coefficient linear models. *Journal of the Royal Statistical Society: series B (statistical methodology)*, 65(1):57–80.

Farr, W. (1840). Progress of epidemics. *Second report of the Registrar General of England and Wales*, pages 16–20.

Ferguson, N., Laydon, D., Nedjati Gilani, G., Imai, N., Ainslie, K., Baguelin, M., Bhatia, S., Boonyasiri, A., Cucunuba Perez, Z., Cuomo-Dannenburg, G., et al. (2020). Report 9: Impact of non-pharmaceutical interventions (npis) to reduce covid19 mortality and healthcare demand.

Ferland, R., Latour, A., and Oraichi, D. (2006). Integer-valued garch process. *Journal of Time Series Analysis*, 27(6):923–942.

Fokianos, K., Rahbek, A., and Tjøstheim, D. (2009). Poisson autoregression. *Journal of the American Statistical Association*, 104:488:1430–1439.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1.

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

Gao, J., Lu, Z., and Tjøstheim, D. (2006). Estimation in semiparametric spatial regression. *The Annals of Statistics*, 34(3):1395–1435.

Gayawan, E., Awe, O. O., Oseni, B. M., Uzochukwu, I. C., Adekunle, A., Samuel, G., Eisen, D. P., and Adegboye, O. A. (2020). The spatio-temporal epidemic dynamics of covid-19 outbreak in africa. *Epidemiology & Infection*, 148.

Giuliani, D., Dickson, M. M., Espa, G., and Santi, F. (2020). Modelling and predicting the spatio-temporal spread of covid-19 in italy. *BMC infectious diseases*, 20(1):1–10.

Google (2021). Covid-19 Community Mobility Report. https://www.google.com/covid19/mobility.

Hallin, M., Lu, Z., and Tran, L. T. (2004). Local linear spatial regression. *The Annals of Statistics*, 32(6):2469–2500.

Hamilton, J. D. (2020). *Time series analysis*. Princeton university press.

Hand, D. J. and Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 160(3):523–541.

Hand, D. J. and Jacka, S. D. (1998). *Statistics in finance*. John Wiley & Sons.

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

Hardle, W., Hall, P., Ichimura, H., et al. (1993). Optimal smoothing in single-index models. *The annals of Statistics*, 21(1):157–178.

Hastie, T. and Tibshirani, R. (1987). Generalized additive models: some applications. *Journal of the American Statistical Association*, 82(398):371–386.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models*, volume 43. CRC press.

Hurvich, C. M. and Tsai, C.-L. (1993). A corrected akaike information criterion for vector autoregressive model selection. *Journal of time series analysis*, 14(3):271–279.

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

Jamieson, K. H. and Albarracin, D. (2020). The relation between media consumption and misinformation at the outset of the sars-cov-2 pandemic in the us. *The Harvard Kennedy School Misinformation Review*.

Jenish, N. (2012). Nonparametric spatial regression under near-epoch dependence. *Journal of Econometrics*, 167(1):224–239.

Jones, M., Davies, S., and Park, B. (1994). Versions of kernel-type regression estimators. *Journal of the American Statistical Association*, 89(427):825–832.

Kim, T., Lieberman, B., Luta, G., and Pena, E. (2020). Prediction regions for poisson and over-dispersed poisson regression models with applications to forecasting number of deaths during the covid-19 pandemic. *arXiv preprint arXiv:2007.02105*.

Kristensen, D. (2009). Uniform convergence rates of kernel estimators with heterogeneous dependent data. *Econometric Theory*, pages 1433–1445.

Lahiri, K. and Yang, L. (2016). A non-linear forecast combination procedure for binary outcomes. *Studies in Nonlinear Dynamics & Econometrics*, 20(4):421–440.

Lauer, S. A., Grantz, K. H., Bi, Q., Jones, F. K., Zheng, Q., Meredith, H. R., Azman, A. S., Reich, N. G., and Lessler, J. (2020). The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: estimation and application. *Annals of internal medicine*, 172(9):577–582.

Lee, L.-F. (2004). Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. *Econometrica*, 72(6):1899–1925.

Lessmann, S., Baesens, B., Seow, H.-V., and Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1):124–136.

Li, D., Linton, O., and Lu, Z. (2015). A flexible semiparametric forecasting model for time series. *Journal of Econometrics*, 187(1):345–357.

Li, D., Lu, Z., and Linton, O. (2012). Local linear fitting under near epoch dependence: uniform consistency with convergence rates. *Econometric Theory*, pages 935–958.

Li, Q. and Racine, J. S. (2007). *Nonparametric econometrics: theory and practice*. Princeton University Press.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2(3):18–22.

Liebscher, E. (1996). Strong convergence of sums of α-mixing random variables with applications to density estimation. *Stochastic Processes and Their Applications*, 65(1):69–80.

Liesenfeld, R., Nolte, I., and Pohlmeier, W. (2006). Modelling financial transaction price movements: a dynamic integer count data model. *Empirical Economics*, 30(4):795–825.

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

Lu, Z. (2001). Asymptotic normality of kernel density estimators under dependence. *Annals of the Institute of Statistical Mathematics*, 53(3):447–468.

Lu, Z. and Chen, X. (2004). Spatial kernel regression estimation: weak consistency. *Statistics & probability letters*, 68(2):125–136.

Lu, Z. and Linton, O. (2007). Local linear fitting under near epoch dependence. *Econometric Theory*, pages 37–70.

Lu, Z., Steinskog, D. J., Tjøstheim, D., and Yao, Q. (2009). Adaptively varying-coefficient spatiotemporal models. *Journal of the Royal Statistical Society: series B (statistical methodology)*, 71(4):859–880.

Lu, Z., Tjøstheim, D., and Yao, Q. (2007). Adaptive varying-coefficient linear models for stochastic processes: asymptotic theory. *Statistica Sinica*, 17(1):177–198.

Masry, E. (1996). Multivariate local polynomial regression for time series: uniform strong consistency and rates. *Journal of Time Series Analysis*, 17(6):571–599.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, volume 37. CRC Press.

Mcdonald, D., Shalizi, C., and Schervish, M. (2011). Estimating beta-mixing coefficients. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 516–524. JMLR Workshop and Conference Proceedings.

McKenzie, E. (1985). Some simple models for discrete variate time series. *JAWRA Journal of the American Water Resources Association*, 21(4):645–650.

Nielsen, S. F. (2005). Local linear estimating equations: Uniform consistency and rate of convergence. *Nonparametric Statistics*, 17(4):493–511.

Peng, R. and Lu, Z. (2021a). Semiparametric averaging of nonlinear marginal logistic regressions and forecasting for time series classification. Submitted for publication.

Peng, R. and Lu, Z. (2021b). Uniform consistency for local maximum likelihood estimation of time series non-parametric regression by local linear estimating equations. Submitted for publication.

Perperoglou, A., Sauerbrei, W., Abrahamowicz, M., and Schmid, M. (2019). A review of spline function procedures in r. *BMC medical research methodology*, 19(1):1–16.

Rosenberg, E. and Gleit, A. (1994). Quantitative methods in credit management: a survey. *Operations research*, 42(4):589–613.

Ryabko, D. and 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(1):2837–2856.

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

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Seifert, B. and Gasser, T. (1996). Finite-sample variance of local polynomials: analysis and solutions. *Journal of the American Statistical Association*, 91(433):267–275.

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

Stork, D. G., Duda, R. O., Hart, P. E., and Stork, D. (2001). Pattern classification. *A Wiley-Interscience Publication*.

Taylor, S. J. (2008). *Modelling financial time series*. world scientific.

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

The Lancet Respiratory Medicine (2020). Covid-19 transmission - up in the air. *The Lancet Respiratory Medicine*, 8(12):1159.

Thomas, L. C. (2000). A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International journal of forecasting*, 16(2):149–172.

Thomas, L. C., Edelman, D. B., and Crook, J. N. (2002). *Credit scoring and its applications*. SIAM.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tibshirani, R. (1997). The lasso method for variable selection in the cox model. *Statistics in medicine*, 16(4):385–395.

Tibshirani, R. and Hastie, T. (1987). Local likelihood estimation. *Journal of the American Statistical Association*, 82(398):559–567.

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

Turner, R., Hayen, A., Dunsmuir, W., and Finch, C. F. (2011). Air temperature and the incidence of fall-related hip fracture hospitalisations in older people. *Osteoporosis international*, 22(4):1183–1189.

UKGovernment (2021). The official UK Government website for data and insights on Coronavirus. https://coronavirus.data.gov.uk.

Unwin, H. J. T., Mishra, S., Bradley, V. C., Gandy, A., Mellan, T. A., Coupland, H., Ish-Horowicz, J., Vollmer, M. A., Whittaker, C., Filippi, S. L., et al. (2020). State-level tracking of covid-19 in the united states. *Nature communications*, 11(1):1–9.

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

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

Venables, W. N., Smith, D. M., Team, R. D. C., et al. (2009). An introduction to r.

Waller, L. A., Carlin, B. P., Xia, H., and Gelfand, A. E. (1997). Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical association*, 92(438):607–617.

Wang, H., Iglesias, E. M., and Wooldridge, J. M. (2013). Partial maximum likelihood estimation of spatial probit models. *Journal of Econometrics*, 172(1):77–89.

Wilhelmsson, M. (2002). Spatial models in real estate economics. *Housing, theory and society*, 19(2):92–101.

Wong, K. C., Li, Z., Tewari, A., et al. (2020). Lasso guarantees for *beta*-mixing heavy-tailed time series. *Annals of Statistics*, 48(2):1124–1142.

Wu, X., Nethery, R. C., Sabath, M., Braun, D., and Dominici, F. (2020). Air pollution and covid-19 mortality in the united states: Strengths and limitations of an ecological regression analysis. *Science advances*, 6(45):eabd4049.

Xia, Y. and Li, W. (1999). On single-index coefficient regression models. *Journal of the American Statistical Association*, 94(448):1275–1285.

Xu, X. and Lee, L.-f. (2015). Maximum likelihood estimation of a spatial autoregressive tobit model. *Journal of Econometrics*, 188(1):264–280.

Yang, C., Chandler, R., Isham, V., and Wheater, H. (2005). Spatial-temporal rainfall simulation using generalized linear models. *Water Resources Research*, 41(11).

Zhang, X., Yu, D., Zou, G., and Liang, H. (2016). Optimal model averaging estimation for generalized linear models and generalized linear mixed-effects models. *Journal of the American Statistical Association*, 111(516):1775–1790.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine learning research*, 7(Nov):2541–2563.

Zhu, G., Zhu, Y., Wang, Z., Meng, W., Wang, X., Feng, J., Li, J., Xiao, Y., Shi, F., and Wang, S. (2021). The association between ambient temperature and mortality of the coronavirus disease 2019 (covid-19) in wuhan, china: a time-series analysis. *BMC Public Health*, 21(1):1–10.

Zhu, J., Zheng, Y., Carroll, A. L., and Aukema, B. H. (2008). Autologistic regression analysis of spatial-temporal binary data via monte carlo maximum likelihood. *Journal of agricultural, biological, and environmental statistics*, 13(1):84–98.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.