Statistics and Its Interface Volume 0 (2020) 1–4

Uniform consistency for local fitting of time series non-parametric regression allowing for discrete-valued response

Rong Peng, Zudi Lu*

Local linear kernel fitting is a popular nonparametric 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 paper, 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. The rate of convergence is also provided under mild conditions. Performances of the proposed method are demonstrated by a Monte-Carlo simulation study and an application to COVID-19 data. There is a huge potential for the developed theory contributing to further development of discrete-valued response semiparametric time series models.

AMS 2000 subject classifications: Primary 62M10, 62M20; secondary 62G05.
Keywords and phrases: Uniform consistency, Discrete-valued time series, Exponential family, Local linear fitting, β-mixing, Non-parametric.

1. INTRODUCTION

The research of local linear regression is of wide interest in statistical and econometric nonlinear and nonparametric modelling (c.f., Fan and Gijbels [10], Fan and Yao [11], Li and Racine [22], Lu and Linton [26]). 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, nonparametric models, and especially semiparametric models that combine nonparametric and parametric methods, are particularly of interest to deal with such situation of nonlinear time series analysis; see e.g., Gao [15] and Terasvirta et al. [36].

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) is another example that plays an important role in areas of disease diagnosing, default rate checking, and so on. Within the discrete-valued time series models, parametric linear or nonlinear autoregression technique is very popular. The reader is referred to Davis, Dunsmuir and Wang [5], Fokianos, Rahbek and Tjøstheim [14] and Davis et al. [6] for a comprehensive review on the related developments.

Differently from those parametric models which suffer from model misspecification, in this paper we propose analysing time series regression in a nonparametric 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 [37], where they have applied it to the generalised linear models and proportional hazards models for independent data. Later Fan, Farmen and Gijbels [9] have discussed the good properties of it in local polynomial fitting. Related research also includes Carroll, Ruppert and Welsh [3], 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. [40], are established in literature as a way to extending results from i.i.d cases to dependent structure (c.f., Bradley [2], Lu [25] and Lu and Linton [26]). In particular, β-mixing, which is often discussed in machine learning [30], 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). The β-mixing property also implies the α-mixing condition as it is stronger and with a faster decay rate. For a more

arXiv: 0000.0000
*Corresponding author.

detailed discussion of β-mixing conditions, the reader is referred to Doukhan, Massart and Rio [7] [Section2.4].

Our focus in this paper is thus to establish the asymptotic properties of the local linear maximum likelihood (LLML) fitting for time series nonparametric regression allowing for discrete-valued response under β-mixing condition. As is well known, the uniform consistency results of such nonparametric kernel-based estimators are widely useful in further developments such as semiparametric modelling (c.f., Nielsen [32], Hansen [17] and Kristensen [19]). Investigations into the method, although extensively made for continuous-valued time series (c.f., Liebscher [23], Masry [28], Bosq [1], Fan and Yao [11], Hansen [17] and Kristensen [19], Li, Lu and Linton [21], and the references therein), are still rare for the time series that are discrete-valued. In this paper, 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. The rate of convergence is also provided under additional mild conditions. Differently from the local least squares based estimation with available analytical solution in the literature (c.f., Li, Lu and Linton [21]), 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, Tjøstheim and Yao [27]) in this paper. Performances of the proposed method are demonstrated by a Monte-Carlo simulation study and an application to COVID-19 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 paper is structured as follows. We will introduce the local linear estimating model in Section 2, followed by the establishment of its uniform consistency with rate of convergence discussed in Section 3. In Section 4, the numerical examples including a Monte-Carlo simulation and an application to COVID-19 data will be demonstrated before the conclusion in Section 5.

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 $\{Z_s, s \geq t+n\}$.

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:

$$m_Y(y;\theta_t) = a(y)\exp(y\theta_t - \phi(\theta_t)), \tag{1}$$

where $a(.)$ 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:

$$\eta(\mu_t) = \theta_t = f(X_t), \tag{2}$$

with $f(\cdot)$ the unknown function that we need to estimate. Therefore this problem of nonparametric estimation is essentially semi-parametric in the sense that nonparametric 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

$$\ell_{h,x}(\boldsymbol{\mu}; Y) = \sum_{t=1}^{n} \log m_{Y_t}(Y_t, \boldsymbol{\theta}_t) K_h(X_t - x), \tag{3}$$

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 $n \to \infty$.

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 (3) being zero,

$$\frac{1}{n}\sum_{t=1}^{n} \omega(Y_t, \theta_t) K_h(X_t - x) = 0, \tag{4}$$

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

$$E[\omega(Y_t, \theta_t)|X_t] = 0. \tag{5}$$

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

$$f(X_t) \approx f(x) + f'(x)(X_t - x) + ... + \frac{f^{(p)}(x)}{p!}(X_t - x)^p$$

$$\equiv \mathbf{x}_t^T\boldsymbol{\beta}, \quad |X_t - x| \le h, \tag{6}$$

where $\mathbf{x}_t = (1, ((X_t - x)/h), ..., ((X_t - x)/h)^p)^T$, with the superscript T denoting a transpose, and $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1 h..., \beta_p h^p)^T$ 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, especially when the dimension d of X_t is greater than 1. In this regard, a local linear fitting is usually preferred, i.e., $p = 1$ (c.f., Fan, Farmen and Gijbels [9]). We are considering a general dimension d for X_t below.

Thus under the first order partial derivative,

$$\begin{aligned} f(X_t) &\approx f(x) + f'(x)^T(X_t - x) \\ &\equiv \beta_0 + h\boldsymbol{\beta}_1^T(X_t - x)/h, \quad if\ |(X_t - x)| \le h, \end{aligned} \tag{7}$$

where $\boldsymbol{\beta}_1 = f'(x)$ is the derivative of $\beta_0 = f(x)$ w.r.t. x, and $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1^T)^T \in \mathbb{R}^{1+d}$ is a vector of local coefficients at x, with f a generic function (not necessarily being the true function) in (2).

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

$$\Omega_n(\boldsymbol{\beta}, x, h) = \begin{pmatrix} \Omega_n^{(1)}(\boldsymbol{\beta}, x, h) \\ \Omega_n^{(2)}(\boldsymbol{\beta}, x, h) \end{pmatrix} = 0, \tag{8}$$

where

$$\begin{aligned} \Omega_n^{(1)}(\boldsymbol{\beta}, x, h) &= \frac{1}{n}\sum_{t=1}^{n}\{\omega(Y_t; \beta_0 + h\boldsymbol{\beta}_1^T((X_t - x)/h)) \\ &\qquad \cdot K_h(X_t - x)\}, \\ \Omega_n^{(2)}(\boldsymbol{\beta}, x, h) &= \frac{1}{n}\sum_{t=1}^{n}\{\omega(Y_t; \beta_0 + h\boldsymbol{\beta}_1^T((X_t - x)/h)) \\ &\qquad \cdot [(X_t - x)/h]K_h(X_t - x)\}, \end{aligned} \tag{9}$$

with $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1^T)^T$ and the bandwidth $h > 0$. In general h is supposed to tend to 0 (seen as depending on n) as $n \to \infty$, in the literature. Here we may see this bandwidth h as a small positive number, an independent parameter not necessarily depending on n though there should be some relationship in our assumptions specified below as $n \to \infty$ and $h \to 0$. This perspective helps to make it easier for our proof below.

By solving the local maximum likelihood estimation above (see Fan, Farmen and Gijbels [9]), 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 (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 $\eta()$.

3. UNIFORM CONSISTENCY

In this section, we will derive the uniform consistency of the local fitting estimator $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\boldsymbol{\beta}}_1^T)^T = (\hat{f}(x), (\hat{f}'(x))^T)^T$ to $\boldsymbol{\beta}_0 = (\beta_{00}, \boldsymbol{\beta}_{01}^T)^T = (f(x), (f'(x))^T)^T$ (with f in $\boldsymbol{\beta}_0$ standing for the true function f in (2) at a slight confusion cost of notation) with respect to $x \in A$, a closed subset of $\mathbb{R}^d$. It is based on the general local estimating equations (8) and (9) given in Section 2.

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 [32]. However, for our concerned β-mixing time series, we give the theorems with proofs shown in this Section 3.

Before jumping into the results, we need the following assumptions.

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 $a \ge (\rho r - 2)r/(2 + \rho r - 4r)$; (ii) the joint probability density function $g_{X_{t_1},\cdots,X_{t_s}}(x_1, \cdots, x_s)$ is bounded uniformly for any $t_1 < \cdots < t_s$ and $1 \le s \le 2(r-1)$; (iii) $E|\omega(Y_t, f(X_t))|^{\rho r} < \infty$, $E|X_t|^{\rho r} < \infty$ for some real number $\rho > 4 - 2/r$, where $r \ge 1$ is some positive integer.

A2 The kernel $K(\cdot)$ is a bounded and symmetric density function on $\mathbb{R}^d$ with bounded support S_K, satisfying $\mu_{2K} = \int_{-\infty}^{\infty} uu^T K(u)du$ with $\|\mu_{2K}\| < \infty$, where $\|\cdot\|$ stands for the Euclidean norm for a vector or matrix. Furthermore, $|K(z) - K(x)| \le C\|z - x\|$ for $z, x \in S_K$ and some $0 < C < \infty$.

A3 The bandwidth $h \to 0$ and the sample size $n \to \infty$ satisfy the condition $\liminf_{n\to\infty, h\to 0} nh^{\frac{2(r-1)a+(\rho r-2)}{(a+1)\rho}} > 0$ for some integer $r \ge 3$. Furthermore, there exists a sequence of positive integers $s_n \to \infty$ such that $s_n = o((nh^d)^{1/2})$, $ns_n^{-b} \to 0$ and $s_n h^{\frac{2(\rho r-2)}{[2+b(\rho r-2)]}} > 1$ as $n \to \infty$ and $h \to 0$.

A4 For any function f given in (2), defined on a close set $A \subset \mathbb{R}^d$, we define its Lipschitz norm: For some $\psi > 0$, let $[\psi]$ be the largest integer not greater than ψ, and define (if it exists)

$$\begin{aligned} \|f\|_{\infty,\psi} &= \max_{0\le\kappa\le[\psi]} \sup_{x\in A} \|f^{(\kappa)}(x)\| \\ &\quad + \sup_{x\ne x'; x,x'\in A} \frac{\|f^{([\psi])}(x) - f^{([\psi])}(x')\|}{\|x - x'\|^{\psi-[\psi]}}, \end{aligned} \tag{10}$$

where $f^{(\kappa)}(x)$ stands for the κ-th derivative of $f(x)$ with respect to x. Define a functional space

$$C_c^{\psi}(A) := \{f : f\ is\ a\ continuous\ function\ from\ A\ to\ \mathbb{R}$$

(11) $\quad$ *with* $\|f\|_{\infty,\psi} \le c\}$,

where c is a positive constant.
We require $\boldsymbol{\beta}_0 = (\beta_{00}, \boldsymbol{\beta}_{01}^T)^T \in \mathbf{F} := C_c^\psi(A) \times (C_c^{\psi-1}(A))^d$ with $\beta_{00} = f \in C_c^\psi(A)$ and $\boldsymbol{\beta}_{01} = f' \in (C_c^{\psi-1}(A))^d$, with f here standing for a true function f in (2) (at a slight cost of notational confusion), for some $\psi \ge 2$ satisfying

$$d/[2(\psi-1)] < 1 - r/[b(r-1)].$$

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

(12) $$\Phi(x, z) = E[\omega(Y_t, z)|X_t = x].$$

- (i) $(x, z) \to \Phi(x, z) \cdot g(x)$ is three times continuously differentiable as a function from $\mathbb{R}^{d+1}$ to $\mathbb{R}$, where $g(x)$ is the marginal density of X_t, 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 the function $z \to \omega(y, z)$ is Lipschitz continuous on a compact set. For any compact $\tilde{C} \subset \mathbb{R}$ there is a function $\Omega^*(y)$ (depending on $\tilde{C}$) such that

  (13) $$|\omega(y, z) - \omega(y, \tilde{z})| \le \Omega^*(y) \cdot |z - \tilde{z}|,$$

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

Remark 3.1. *Assumption 1 shows a standard β-mixing process which is satisfied by many linear and non-linear time series models [11, 27]. The kernel is guaranteed to be bounded by Assumption 2, which is commonly seen in this type of problem [18, 41]. Assumption 3 is also standard in time series topics [12, 26] though we see the bandwidth h as an independent parameter in this paper. The Lipschitz norm conditions (Assumptions A4 and A5) are introduced to give a tighter bound than uniform norm [32]. In A4, if $d = 1$ and $\psi = 2$, then the condition $d/[2(\psi-1)] < 1 - r/[b(r-1)]$ is satisfied under $b > 2r/(r-1)$ imposed mildly on the mixing coefficient in A1. Note that we are concerned with $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$, which, as a function of x, are in $\mathbf{F} = C_c^\psi(A) \times (C_c^{\psi-1}(A))^d$. Under A4, the Lipschitz continuous norm is stronger than the uniform norm for a function in $\mathbf{F}$, i.e.*

(14) $$\|\boldsymbol{\beta}\|_{\mathbf{F}} = \|\boldsymbol{\beta}\|_\infty = \max_{i=1,2,\cdots,d+1} sup_{x\in A}|\beta_i(x)| \le \|\boldsymbol{\beta}\|_{\infty,\psi},$$

where $\beta_i(x)$ denotes the i-th component of $\boldsymbol{\beta}$. Thus, consistency in Lipschitz norm implies uniform consistency. Assumption A5 was introduced for a general case of local estimating equations (c.f., Nielsen [32]). Recalling that $f(X_t) = \eta(\mu_t)$ under canonical link function $\eta = \phi'^{-1}$ and (9), we have $\omega(Y_t, z) = Y_t - \phi'(z)$, $\Phi(x, z) = E[\omega(Y_t, z)|X_t = x] = E(Y_t|X_t = x) - \phi'(z) = \phi'(f(x)) - \phi'(z)$. Clearly $\Phi(x, f(x)) = 0$, where $f(x)$ is the true function defined in (2), also denoted as β_{00} below. Here assumption A5 holds automatically under assumption A1.

3.2 Theorems

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

Theorem 3.1. *Suppose the assumptions A1-A4 with model 2 are satisfied. Then, as $n \to \infty$ and $h \to 0$,*

$$E[\Omega_n(\boldsymbol{\beta}, x, h)] = (1 + o(1))\text{diag}(1, h\mathcal{I}_d)\Omega_0(\boldsymbol{\beta}, x),$$

where $o(1)$ is uniformly with respect to $x \in A$ and $\boldsymbol{\beta} \in \mathbf{F}$, $\mathcal{I}_d$ is a $d \times d$ identity matrix, and $\Omega_0(\boldsymbol{\beta}, x) = (\Omega_0^{(1)}(\boldsymbol{\beta}, x), (\Omega_0^{(2)}(\boldsymbol{\beta}, x))^T)^T$, with $\Omega_0^{(1)}(\boldsymbol{\beta}, x) = \Phi(x, \beta_0)g(x)$ and

$$\Omega_0^{(2)}(\boldsymbol{\beta}, x) = (\boldsymbol{\beta}_1\dot{\Phi}_z(x, \beta_0) + \dot{\Phi}_x(x, \beta_0))g(x) + \Phi(x, \beta_0)g'(x).$$

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

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

Further, $E[\Omega_n(\boldsymbol{\beta}, x, h)] = 0$ 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 [32].

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

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

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 $\|\boldsymbol{\beta}\|_{\infty,\phi} \le c$ 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):

$$\begin{aligned}
&E[\Omega_n^{(1)}(\boldsymbol{\beta}, x, h)] \\
&= E\left[\frac{1}{n}\sum_{t=1}^n [Y_t - \phi'(\beta_0 + \boldsymbol{\beta}_1^T(X_t - x))]K_h(X_t - x)\right] \\
&= E\left[E[\frac{1}{n}\sum_{t=1}^n [Y_t - \phi'(\beta_0 + \boldsymbol{\beta}_1^T(X_t - x))]]K_h(X_t - x)|X_t\right] \\
&= E\left[\frac{1}{n}\sum_{t=1}^n [E[Y_t|X_t] - \phi'(\beta_0 + \boldsymbol{\beta}_1^T(X_t - x))]K_h(X_t - x)\right]
\end{aligned}$$

$$= E\left[\frac{1}{n}\sum_{t=1}^{n}[\phi'(f(X_t)) - \phi'(\beta_0 + \boldsymbol{\beta}_1^T(X_t - x))]K_h(X_t - x)\right],$$

where $E[Y_t|X_t] = \phi'(f(X_t))$ follows from (2).

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

$$\begin{aligned} E[\Omega_n^{(1)}(\boldsymbol{\beta}, x, h)] &= E\Bigg[\frac{1}{n}\sum_{t=1}^{n}[\tilde{f}(f(X_t)) \\ &\quad - \tilde{f}(\beta_0 + \boldsymbol{\beta}_1^T(X_t - x))]K_h(X_t - x)\Bigg] \\ &= (1 + o(1))[\tilde{f}(f(x)) - \tilde{f}(\beta_0)]g(x), \end{aligned} \tag{16}$$

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, where $\omega(y; z) = y - \phi'(z)$ as indicated above, but for a general $\omega(y, z)$ under assumption A5, we can still establish (16) as in Nielsen [32]:

$$\begin{aligned} &E[\Omega_n^{(1)}(\boldsymbol{\beta}, x, h)] \\ &= E[\omega(Y_t; \beta_0 + \boldsymbol{\beta}_1^T(X_t - x))K_h(X_t - x)] \\ &= E[\Phi(X_t; \beta_0 + \boldsymbol{\beta}_1^T(X_t - x))K_h(X_t - x)] \\ &= \Phi(x, \beta_0)g(x) + O(h^2), \end{aligned}$$

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, $\Phi(x, \beta_0) = \tilde{f}(f(x)) - \tilde{f}(\beta_0)$.

Similarly from (9), as done above (Nielsen [32]),

$$\begin{aligned} E[\Omega_n^{(2)}(\boldsymbol{\beta}, x, h)] &= h\mu_{2K}[(\boldsymbol{\beta}_1\dot{\Phi}_z(x, \beta_0) \\ &\quad + \dot{\Phi}_x(x, \beta_0))g(x) + \Phi(x, \beta_0)g'(x)] + O(h^3), \end{aligned}$$

where corresponding to the local exponential family regression in Section 2, $\dot{\Phi}_x(x, \beta_0) = \tilde{f}'(f(x))f'(x) = f'(x)\phi''(f(x))$ and $\dot{\Phi}_z(x, \beta_0) = -\tilde{f}'(\beta_0) = -\phi''(\beta_0)$, with $\tilde{f}'(z) = \phi''(z)$ as defined above.

Thus we get:

$$E[\Omega_n^{(1)}(\boldsymbol{\beta}, x, h)] = \Omega_0^{(1)}(\boldsymbol{\beta}, x) + O(h^2) \tag{17}$$

and

$$E[\Omega_n^{(2)}(\boldsymbol{\beta}, x, h)] = h\Omega_0^{(2)}(\boldsymbol{\beta}, x) + O(h^3) \tag{18}$$

where

$$\begin{aligned} \Omega_0^{(1)}(\boldsymbol{\beta}, x) &= \Phi(x, \beta_0)g(x) \\ \Omega_0^{(2)}(\boldsymbol{\beta}, x) &= \mu_{2K}[(\boldsymbol{\beta}_1\dot{\Phi}_z(x, \beta_0) + \dot{\Phi}_x(x, \beta_0))g(x) \\ &\quad + \Phi(x, \beta_0)g'(x)]. \end{aligned}$$

Denote by $\boldsymbol{\beta}_0 = (\beta_{00}, \boldsymbol{\beta}_{01}^T)^T$ the solution to $\Omega_0(\boldsymbol{\beta}, x) = 0$, where $\Omega_0(\boldsymbol{\beta}, x) = (\Omega_0^{(1)}(\beta, x), (\Omega_0^{(2)}(\beta, x))^T)^T$. Then we have:

$$\begin{cases} \Phi(x, \beta_{00}) = 0 \\ \boldsymbol{\beta}_{01}(x) = -\frac{\dot{\Phi}_x(x, \beta_{00})}{\dot{\Phi}_z(x, \beta_{00})}, \end{cases} \tag{19}$$

which is actually unique correspondingly to our local general linear regression in Section 2, with $\beta_{00} = f(x)$ and $\boldsymbol{\beta}_{01} = f'(x)$ (at a slight cost of notational confusion with $f(x)$ for the true function here). The proof of Theorem 3.1 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, by the unique solution in (19) to $\Omega_0(\boldsymbol{\beta}, x) = 0$, for any $\delta > 0$, there exists some $\varepsilon > 0$, such that

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_\infty > \delta \Rightarrow \max_{i=1,2}|\Omega_0^{(i)}(\hat{\boldsymbol{\beta}}, x)| > \varepsilon, \quad \textit{for} \quad x \in A. \tag{20}$$

As done in Nielsen [32], we will assume that $C > 0$ has been chosen so that $\|\boldsymbol{\beta}_0(x)\| \leq C$ for any $x \in A$. Note that by the differentiability of $\boldsymbol{\beta}_0$ and the boundedness of A such a C exists. Further, the estimating function in equation (8) will typically be sufficiently smooth to guarantee (via the implicit function theorem) that the estimator defined by equation (8) is continuously differentiable and thus Lipschitz on A. In this case, minimising over Lipschitz functions is not a restriction, but it allows us to define an estimator $\hat{\boldsymbol{\beta}}$ even if equation (8) cannot be solved. Minimising over a bounded set of Lipschitz functions is a restriction, though, and one should be careful that C is chosen sufficiently large.

By (20), if we 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 $n \to \infty$ and $h \to 0$, then we have the uniform consistency as follows.

Theorem 3.2. *Suppose the assumptions A1-A5 are satisfied. Then* $\mathrm{diag}(1, h^{-1}\mathcal{I}_d)\Omega_n(\boldsymbol{\beta}, x)$ *converges uniformly in probability to* $\Omega_0(\boldsymbol{\beta}, x)$ *with respect to* $\boldsymbol{\beta} \in \mathbf{F}, x \in A$, *and further* $\hat{\boldsymbol{\beta}}_n(x) - \boldsymbol{\beta}_0(x) \to 0$ *uniformly for* $x \in A$, *as* $n \to \infty$ *and* $h \to 0$.

Proof. Let $D_n = \mathrm{diag}(1, h\mathcal{I}_d)$. In view of (20), we need the fact that $\tilde{\Omega}_n(\boldsymbol{\beta}, x) = D_n^{-1}\Omega_n(\boldsymbol{\beta}, x)$ converges in probability to $D_n\Omega_0(\boldsymbol{\beta}, x)$ uniformly with respect to $\boldsymbol{\beta} \in \mathbf{F}$ and $x \in A$, the proof of which is sketched below, under the given assumptions.

We notice from (8) and (19) that $\Omega_n^{(i)}(\hat{\boldsymbol{\beta}}(x), x) = 0$ and $\Omega_0^{(i)}(\boldsymbol{\beta}_0(x), x) = 0$, and hence

$$\begin{aligned} \Omega_0^{(i)}(\hat{\boldsymbol{\beta}}, x) &= \Omega_0^{(i)}(\hat{\boldsymbol{\beta}}, x) - \Omega_0^{(i)}(\boldsymbol{\beta}_0, x) \\ &= (\Omega_0^{(i)}(\hat{\boldsymbol{\beta}}, x) - \tilde{\Omega}_n^{(i)}(\hat{\boldsymbol{\beta}}, x)). \end{aligned} \tag{21}$$

Letting $Z_t = (Y_t, X_t^T)^T$, define the empirical process:

(22)
$$G_n(\boldsymbol{\beta}, x, h) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} (\omega^*(Z_t, \boldsymbol{\beta}, x, h) - E[\omega^*(Z_t, \boldsymbol{\beta}, x, h)]),$$

where, letting $\omega_t(\boldsymbol{\beta}, x) = \omega\Big(Y_t, \beta_0 + \boldsymbol{\beta}_1^T(X_t - x)\Big)$,

$$\omega^*(Z_t, \boldsymbol{\beta}, x, h) = \omega_t(\boldsymbol{\beta}, x) K((X_t - x)/h) \begin{bmatrix} 1 \\ (\frac{X_t - x}{h}) \end{bmatrix}.$$

Note that

(23)
$$\sqrt{nh^d} D_n^{-1}(\Omega_n(\boldsymbol{\beta}, x) - E\Omega_n(\boldsymbol{\beta}, x)) = h^{-d/2} G_n(\boldsymbol{\beta}, x, h).$$

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 $(nh^d)^{-1/2} h^{-d/2} G_n(\boldsymbol{\beta}, x, h)$.

The two components of $G_n(\boldsymbol{\beta}, x, h)$ are denoted by $G_n^{(i)}(\boldsymbol{\beta}, x, h)$, $i = 1, 2$. In this proof, we need to determine when $G_n(\boldsymbol{\beta}, x, h)$ converges uniformly in distribution to a multivariate Gaussian process $G(\boldsymbol{\beta}, x, h)$ indexed by $\vartheta \equiv (\boldsymbol{\beta}, x) \in \mathbf{F} \times A$ for any fixed $h > 0$. Note that we see h and n are two independent parameters in this paper, the perspective of which helps to make the proof easier here. 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 $h^{-d/2} G_n(\vartheta, h)$ to a multivariate Gaussian distribution $h^{-d/2} G(\vartheta, h)$ at any finite number of pairs of $\vartheta = (\boldsymbol{\beta}, x)$'s (c.f., Lu, Tjøstheim and Yao [27]), with the mean 0 and the covariance between $h^{-d/2} G(\vartheta, h)$ and $h^{-d/2} G(\vartheta', h)$, at $\vartheta, \vartheta' \in \mathbf{F} \times A$, equal to

$$\Gamma_h(\vartheta, \vartheta') = h^{-d} \sum_{j=-\infty}^{\infty} \operatorname{cov}(W^*(Z_t, \vartheta, h), W^*(Z_{t+j}, \vartheta', h)).$$

Note that under the mixing condition A1, as $h \to 0$,

(24)
$$\Gamma_h(\vartheta, \vartheta') \to \Gamma(\vartheta, \vartheta'),$$

where $\Gamma(\vartheta, \vartheta')$ equals zero if $\vartheta \neq \vartheta'$, and $\Gamma(\vartheta, \vartheta) =$

$$E(\omega_t^2(\vartheta) | X_t = x) \begin{bmatrix} \int_{R^d} K^2(u) du & 0 \\ 0 & \int_{R^d} u u^T K^2(u) du \end{bmatrix}.$$

The proof is routine because of the CLT for mixing processes based on the Bernstein blocking technique (see, e.g., Hallin, Lu and Tran [16] Theorem 3.1, and Lu and Linton [26]), and therefore the details are omitted. 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 $\eta > 0$, there is a $\delta > 0$ such that:

$$\limsup_{n \to \infty} P\Big(\sup_{\boldsymbol{\beta} \in \mathbf{F}, x \in A} \sup_{(\boldsymbol{\beta}', x') \in B((\boldsymbol{\beta}, x), \delta)} |G_n^{(i)}(\boldsymbol{\beta}'(\cdot), x', h)$$

(25)
$$- G_n^{(i)}(\boldsymbol{\beta}(\cdot), x, h)| > \epsilon \Big) < \eta.$$

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, Massart and Rio [7].

Lemma 3.1. *To prove the stochastic equicontinuity of the empirical process we need to check the following conditions*
(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) $E_p[\{\widetilde{\Omega}^*(Z_t)\}^{2r}] < \infty$, *where* $r > 1$ *in (a), and* $\widetilde{\Omega}^*(Z_t)$ *is the envelope of* $\mathcal{M} = \{\omega^*(\cdot, \boldsymbol{\beta}, x, h) : \boldsymbol{\beta} \in \mathbf{F}, x \in A\}$, *that is* $|\omega^*(\cdot, \boldsymbol{\beta}, x, h)| \leq \widetilde{\Omega}^*(\cdot)$ *for any* $\boldsymbol{\beta} \in \mathbf{F}, x \in A$.
(c) $\forall \varepsilon > 0$, $\log N_2(\epsilon, \mathcal{M}) = O(\varepsilon^{-2\eta})$ *for some* $\eta > 0$ *satisfying* $\eta < 1 - r/[b(r-1)]$, *for* b *and* r *as in (a) and (b), where* $N_2(\varepsilon, \mathcal{M})$ *is the* L_2*-bracketing cover number of* $\mathcal{M}$ *in (b)*

This lemma is an alternative statement of Application 4 in Doukhan et al. ([7], 1995, page 405).

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 $\|\boldsymbol{\beta}\| \leq C$ for a sufficiently large $C > 0$. Note that

$$\{(y, z) \longrightarrow \omega(y; \beta_0(x) + \boldsymbol{\beta}_1^T(x)(z - x) K(\frac{z - x}{h}) : x \in A, h > 0, \boldsymbol{\beta} \in \mathbf{F}, \|\boldsymbol{\beta}\|_{\mathbf{F}} < C\}$$

with the envelope

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

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

$$\{(y, z) \longrightarrow \omega(y; \beta_0(x) + \boldsymbol{\beta}_1(x)(z - x)) \frac{z - x}{h} K(\frac{z - x}{h}) : x \in A, h > 0, \boldsymbol{\beta} \in \mathbf{F}, \|\boldsymbol{\beta}\|_{\mathbf{F}} < C\}$$

with the envelope

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

Hence (b) holds by conditions A1 and A2.

(c) The proof can be done as in Lu, Tjøstheim and Yao [27] (page S26), so we only have a simple idea given here. 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, for the i-th component,

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

Similarly, by Lipschitz norm

$$
\begin{aligned}
&|\boldsymbol{\beta}_1(x)^T x - \widetilde{\boldsymbol{\beta}}_1(\widetilde{x})^T \widetilde{x}| \\
&\le \|\boldsymbol{\beta}_1(x)\| \cdot \|x - \widetilde{x}\| + \|\widetilde{x}\| \cdot \|\boldsymbol{\beta}_1(x) - \widetilde{\boldsymbol{\beta}}_1(\widetilde{x})\| \\
&\le C\|x - \widetilde{x}\| + \sup_{x \in A} \|x\| \cdot (\|\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\|_\infty + 2C\|x - \widetilde{x}\|).
\end{aligned}
$$

As $\mathbf{F} = C_c^{\psi}(A) \times (C_c^{(\psi-1)}(A))^d$ with $\psi \ge 2$, for $\forall \varepsilon > 0$, we can cover $\mathbf{F}$ by finite number, say N_1, of balls of radius ε with centres $\boldsymbol{\beta}_i, i = 1, .., N_1$, in $\mathbf{F}$, such that: $\forall \boldsymbol{\beta} \in \mathbf{F}, \exists \boldsymbol{\beta}_i$, such that

$$
\|\boldsymbol{\beta} - \boldsymbol{\beta}_i\|_\infty \le \frac{\varepsilon}{2C}.
$$

By **?**] (Theorem 2.7.1), it is known that $N_1 = N(\varepsilon, \mathbf{F}, \|\cdot\|_\infty)$ satisfies $\log N(\varepsilon, \mathbf{F}, \|\cdot\|_\infty) \le C\varepsilon^{-d/(\psi-1)} = C\varepsilon^{-2\eta}$ with $\eta = d/[2(\psi-1)] < 1 - r/[b(r-1)]$ by condition A4. Similarly, A is a closed subset in $\mathbb{R}^d$, for $\forall \varepsilon > 0$, we can cover A by finite number, $N_2 = C\varepsilon^{-d}$, of balls of radius ε with centres $x_j, j = 1, .., N_2$, in A, such that: $\forall x \in A, \exists x_i$, such that

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

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

$$
\omega\Big(Y_t, \beta_0 + \boldsymbol{\beta}_1(X_t - x)\Big) \cdot K((X_t - x)/h) \begin{bmatrix} 1 \\ (\frac{X_t - x}{h}) \end{bmatrix}
$$

is a continuous function of $(\boldsymbol{\beta}, x)$, $\omega^*(\cdot, \boldsymbol{\beta}, x, h)$ can be approximated by, say, $\omega^*(\cdot, \boldsymbol{\beta}_{i^*}, x_{j^*}, h)$ for some i^* and j^* for any $\boldsymbol{\beta} \in \mathbf{F}, x \in A$. Therefore we can cover $\mathcal{M} = \{\omega^*(\cdot, \boldsymbol{\beta}, x, h) : \beta \in \mathbf{F}, x \in A\}$ by $N_2(\varepsilon, M) \le N_1 \times N_2$ suitably defined balls as specified in (c) (**?**], Theorem 2.7.1). The details are omitted here (c.f., Lu, Tjøstheim and Yao [27]).

Thus $\{G_n(\boldsymbol{\beta}, x, h) : \boldsymbol{\beta} \in \mathbf{F}, x \in A\}$ converges in distribution to process $\{G(\boldsymbol{\beta}, x, h) : \boldsymbol{\beta} \in \mathbf{F}, x \in A\}$, and hence $\{h^{-d/2} G_n(\boldsymbol{\beta}, x, h) : \boldsymbol{\beta} \in \mathbf{F}, x \in A\}$ converges in distribution to the process $\{h^{-d/2} G(\boldsymbol{\beta}, x, h) : \boldsymbol{\beta} \in \mathbf{F}, x \in A\}$ for a fixed $h > 0$. Note that as $h \to 0$, $\{h^{-d/2} G(\boldsymbol{\beta}, x, h) : \boldsymbol{\beta} \in \mathbf{F}, x \in A\}$ converges to a Gaussian process indexed by $\vartheta = (\boldsymbol{\beta}, x) \in \mathbf{F} \times A$, with mean zero and variance-covariance equal to $\Gamma(\vartheta, \vartheta')$ defined in (24). Hence,

$$
\sup_{\|\boldsymbol{\beta}(x)\| \le C, x \in A} |h^{-d/2} G_n^{(i)}(\beta, x, h)| = Op(1), i = 1, 2. \tag{26}
$$

By (26) together with Equation (17) we have

$$
\begin{aligned}
&\sup_{\|\boldsymbol{\beta}(x)\| \le C, x \in A} |\Omega_n^{(1)}(\boldsymbol{\beta}(x), x, h) - \Omega_0^{(1)}(\boldsymbol{\beta}(x), x)| \\
&\le \frac{1}{\sqrt{nh^d}} \sup_{\|\boldsymbol{\beta}(x)\| \le C, x \in A, h > 0} |h^{-d/2} G_n^{(1)}(\boldsymbol{\beta}(x), x, h)| \\
&+ \sup_{\|\boldsymbol{\beta}(x)\| \le C, x \in A, h > 0} |E[\Omega_n^{(1)}(\boldsymbol{\beta}(x), x, h) - \Omega_0^{(1)}(\boldsymbol{\beta}(x), x)]| \\
&= Op(1/(\sqrt{nh^d})) + O(h^2) \xrightarrow{\mathrm{P}} 0; \qquad (27)
\end{aligned}
$$

and, similarly, by (26) together with (18),

$$
\begin{aligned}
&\sup_{\|\boldsymbol{\beta}(x)\| \le C, x \in A} |\frac{1}{h}\Omega_n^{(2)}(\boldsymbol{\beta}(x), x, h) - \Omega_0^{(2)}(\boldsymbol{\beta}(x), x)| \\
&\le \sup_{\|\boldsymbol{\beta}(x)\| \le C, x \in A} \|\frac{1}{h}[\Omega_n^{(2)}(\boldsymbol{\beta}(x), x, h) - E\Omega_n^{(2)}(\boldsymbol{\beta}(x), x)]\| \\
&+ \sup_{\|\boldsymbol{\beta}(x)\| \le C, x \in A} \|\frac{1}{h} E\Omega_n^{(2)}(\boldsymbol{\beta}(x), x, h) - \Omega_0^{(2)}(\boldsymbol{\beta}(x), x)\| \\
&\le h^{-1} \frac{1}{\sqrt{nh^d}} \sup_{\|\boldsymbol{\beta}(x)\| \le C, x \in A, h > 0} |h^{-d/2} G_n^{(2)}(\boldsymbol{\beta}(x), x, h)| \\
&+ \sup_{\|\boldsymbol{\beta}(x)\| \le C, x \in A} \|\frac{1}{h} E\Omega_n^{(2)}(\boldsymbol{\beta}(x), x, h) - \Omega_0^{(2)}(\boldsymbol{\beta}(x), x)\| \\
&= Op(\frac{1}{\sqrt{nh^{d+2}}}) + O(h^2). \qquad (28)
\end{aligned}
$$

Thus by (20) and (21) with $nh^{d+2} \to \infty$ and $h \to 0$, it follows that (21) converges in probability to zero uniformly with respect to $x \in A$, and $\|\hat{\boldsymbol{\beta}}(.) - \boldsymbol{\beta}_0(.)\|_{\mathbf{F}} = o_p(1)$ is proved. □

Based on Theorem 3.2, 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), 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.

Theorem 3.3. *Under the assumptions of Theorem 3.2 with a continuous link function η, we have*

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

in probability as $n \to \infty$ and $h \to 0$.

In practice, we can take the close set $A \subset \mathbb{R}^d$ 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 and the bandwidth h to zero.

Next, we provide a uniform convergence rate for $\hat{\boldsymbol{\beta}}(x) = (\hat{f}(x), (\hat{f}'(x))^T)^T$ over a closed set A by refining the argument in the proof of Theorem 3.2.

Theorem 3.4. *Suppose the assumptions A1-A4 with model (2) are satisfied. Then, uniformly for $x \in A$,*

$$
\dot{\Omega}_{\boldsymbol{\beta}_0}(\hat{\boldsymbol{\beta}}(x) - \boldsymbol{\beta}_0(x)) = \begin{bmatrix} O_p(\frac{1}{\sqrt{nh^d}} + h^2) \\ \mathbf{1}_d O_p(\frac{1}{\sqrt{nh^{d+2}}} + h^2) \end{bmatrix}, \tag{29}
$$

as $n \to \infty$ and $h \to 0$, where $\dot{\Omega}_{\boldsymbol{\beta}_0}$ is a $(1+d)\times(1+d)$ matrix partitioned into 4 blocks with the $(1,1)$th block a 1×1 sub-matrix whose element equal to $\dot{\Phi}_z(x,\beta_{00})g(x)$, the $(1,2)$th block being a $1\times d$ sub-matrix of elements 0's, the $(2,1)$th block being a $d\times 1$ sub-matrix $\mu_{2K}\{[\boldsymbol{\beta}_{01}\ddot{\Phi}_{zz}(x,\beta_{00})+\ddot{\Phi}_{xz}(x,\beta_{00})]g(x)+\dot{\Phi}_z(x,\beta_{00})g'(x)\}$, and the $(2,2)$th block being a $d\times d$ sub-matrix $\mu_{2K}\dot{\Phi}_z(x,\beta_{00})g(x)$, and $\mathbf{1}_d$ stands for a d-dimensional vector of elements equal to 1.

Proof. Note that the map $\boldsymbol{\beta}\mapsto\boldsymbol{\omega}(\boldsymbol{\beta},x)$ is continuously differentiable at point of $\boldsymbol{\beta}_0$ with the first derivative at $\boldsymbol{\beta}_0$ (recalling the notation $\boldsymbol{\beta}=(\beta_0,\boldsymbol{\beta}_1^T)^T$ and $\boldsymbol{\beta}_0=(\beta_{00},\boldsymbol{\beta}_{01}^T)^T$). Also take notice of the function $\Omega_0(\boldsymbol{\beta},x)$ defined in (17) and (18) with the derivate at $\boldsymbol{\beta}=\boldsymbol{\beta}_0$:

$$\begin{aligned}&\Omega_0^{(1)}(\boldsymbol{\beta},x)-\Omega_0^{(1)}(\boldsymbol{\beta}_0,x)\\&=\Phi(x,\beta_0)g(x)-\Phi(x,\beta_{00})g(x)\\&=\dot{\Phi}_z(x,\beta_{00})g(x)(\beta_0-\beta_{00})+O(\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_\infty^2),\end{aligned}\tag{30}$$

and

$$\begin{aligned}&\Omega_0^{(2)}(\boldsymbol{\beta},x)-\Omega_0^{(2)}(\boldsymbol{\beta}_0,x)\\&=\mu_{2K}[(\beta_2\dot{\Phi}_z(x,\beta_0)+\dot{\Phi}_x(x,\beta_0))g(x)+[\Phi(x,\beta_0)]g'(x)\\&-\mu_{2K}[(\boldsymbol{\beta}_{01}\dot{\Phi}_z(x,\beta_{00})+\dot{\Phi}_x(x,\beta_{00}))g(x)\\&\qquad+(\Phi(x,\beta_{00}))g'(x)]\\&=\mu_{2K}\{[\boldsymbol{\beta}_{01}\ddot{\Phi}_{zz}(x,\beta_{00})+\ddot{\Phi}_{xz}(x,\beta_{00})]g(x)\\&\qquad+\dot{\Phi}_z(x,\beta_{00})g'(x)\}\times(\beta_0-\beta_{00})\\&\quad+\mu_{2K}\dot{\Phi}_z(x,\beta_{00})g(x)(\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01})+O(\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_\infty^2).\end{aligned}$$

Hence,

$$\begin{aligned}&\Omega_0(\boldsymbol{\beta},x)-\Omega_0(\boldsymbol{\beta}_0,x)\\&=\dot{\Omega}_{\boldsymbol{\beta}_0}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)+O(\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_\infty)(\boldsymbol{\beta}-\boldsymbol{\beta}_0).\end{aligned}\tag{31}$$

Now let $\hat{\boldsymbol{\beta}}_n$ be the uniformly consistent estimator of $\boldsymbol{\beta}_0$ with assumption held. It then follows from the above that $\|\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0\|_\infty$ and $\Omega_0(\hat{\boldsymbol{\beta}}_n,x)-\Omega_0(\boldsymbol{\beta}_0,x)$ have the same convergence rates uniformly with respect to $x\in A$. We have, by noticing that $\Omega_n(\hat{\boldsymbol{\beta}},x,h)=0$ and $\Omega_0(\boldsymbol{\beta}_0,x)=0$,

$$\begin{aligned}&\Omega_0(\hat{\boldsymbol{\beta}}_n,x)-\Omega_0(\boldsymbol{\beta}_0,x)\\&=\Big\{D_n^{-1}\Omega_n(\hat{\boldsymbol{\beta}}_n,x,h)-\Omega_0(\hat{\boldsymbol{\beta}}_n,x)\Big\}\equiv I_n.\end{aligned}\tag{32}$$

It is noted that, uniformly for $x\in A$ and $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_{\mathbf{F}}\le\delta_n\to 0$, we have

$$\sup_{x\in A}\|I_n\|\le\sup_{x\in A}\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_{\mathbf{F}}\le\delta_n}\|\Big\{D_n^{-1}\Omega_n(\boldsymbol{\beta},x,h)-\Omega_0(\boldsymbol{\beta},x)\Big\}\|.$$

Note that

$$\begin{aligned}&D_n^{-1}\Omega_n(\boldsymbol{\beta},x,h)-\Omega_0(\boldsymbol{\beta},x)\\&\quad=D_n^{-1}[\Omega_n(\boldsymbol{\beta},x,h)-E\Omega_n(\boldsymbol{\beta},x,h)]\\&\quad+[D_n^{-1}E\Omega_n(\boldsymbol{\beta},x,h)-\Omega_0(\boldsymbol{\beta},x)]\end{aligned}\tag{33}$$

Now from equations (27) and (28) as well as Theorem 3.1, we have

$$I_n=D_n^{-1}\mathbf{1}O_P\{(1/(nh^d))^{1/2}\}+h^2\mathbf{1},$$

where $\mathbf{1}$ is a $(1+d)\times 1$ vector of elements being 1's. The proof is done. □

Before ending this section, we make a remark as a referee commented. Indeed, other mixing, such as α-mixing, dependence has been one of the popular dependence conditions in statistical and econometric literature; see Masry and Tjøstheim (1995) [29] and Lu (1998) [24] for example. The α-mixing is implied by the β-mixing condition that is guaranteed by the geometric ergodicity of time series models. In fact, under certain weak assumptions, many autoregressive and more general nonlinear time-series models are geometrically ergodic and β-mixing, and hence strongly mixing (i.e., α-mixing), with exponential mixing rates (c.f., Masry and Tjøstheim (1995) [29] and Lu (1998) [24]). See also Pham and Tran (1985) [35], Pham (1986) [34] and Tjøstheim (1990) [38] for more information. Note that in those mentioned references, the geometric ergodicity is studied for time series models (either linear or nonlinear), by which the β-mixing, and hence α-mixing, is derived with exponentially decreasing mixing coefficients. In this sense, from the time series modelling perspective, β-mixing is clearly an ideal mixing concept characterising the technical condition on the dependence of data as well. In this paper, we adopted this technical concept of β-mixing because we applied the empirical process theory from the reference Doukhan, Massart and Rio (1995, p.405, [7]), which needs β-mixing. Although there have been some references studying empirical process under α-mixing (c.f., Mohr (2020) [31] and the related references therein), their condition imposed on the covering/bracketing number $N_1(\varepsilon)=N(\varepsilon,\mathbf{F},\|\cdot\|_\infty)$ is of a form $\int_0^1 x^{-\gamma/(1+\gamma)}(N_1(x))^{1/Q}dx<\infty$ for some $\gamma>0$ and $Q\ge 2$, which cannot be satisfied by the $N_1(x)=O(1)\exp\{Cx^{-2\eta}\}$ for the class of functions $\mathbf{F}$ with $\eta>0$ as specified in the proof of Theorem 3.2 above. It can be conjectured that our uniform consistency holds true under α-mixing, which however needs further investigation. With different mixing concepts, the conditions on the mixing coefficients may change, reflecting the strength of dependence for two segments of a time series, which will result in a difference in Assumption A1. We leave all these questions for future research.

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. We hope to demonstrate that the proposed local linear method is robust for the exponential family.

4.1 Simulation

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

$$\begin{aligned} X_t &= \cos(2X_{t-1}) + \epsilon_t \\ Y_t &= I(X_t > 0), \end{aligned} \tag{34}$$

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:

$$\begin{aligned} p(x) &= P(Y_t = 1|X_{t-1} = x) \\ &= P(cos(2X_{t-1}) + \epsilon_t > 0|X_{t-1} = x) \\ &= P(\frac{\epsilon_t}{\sigma} > \frac{-cos(2x)}{\sigma}) \\ &= 1 - \Phi(-\frac{cos(2x)}{\sigma}) = \Phi(\frac{cos(2x)}{\sigma}) \end{aligned} \tag{35}$$

The corresponding log odds can be obtained from:

$$f(X_t) = log\frac{p_t}{1-p_t}, \tag{36}$$

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

$$\log L = \sum_{i=1}^{n}[log(p_i)\cdot Y_i + log(1-p_i)(1-Y_i)]K(\frac{X_i - x}{h}), \tag{37}$$

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

$$K(u) = \frac{3}{4}(1-u^2)I_{[-1,1]}(u), \tag{38}$$

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 [9]. Different criterion would also lead to different optimal h. In this paper, we are going to use cross validation based on log likelihood to select the best h within given data sample. Note that the $\log L$

$$\begin{aligned} logL = \sum_{i=1}^{n}[log(\frac{1}{1+e^{-(f+f'(X_i-x))}})\cdot Y_i \\ + \log(1-\frac{1}{1+e^{-(f+f'(X_i-x))}})\cdot(1-Y_i)]K(\frac{X_i-x}{h}). \end{aligned} \tag{39}$$

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:

$$\hat{p}_i^{h_{[-i]}} = \frac{1}{1+e^{-\hat{f}^h_{[-i]}(X_i)}}, \tag{40}$$

$$CV(h) = \sum_{i=1}^{n}[log(\hat{p}_i^{h_{[-i]}})Y_i + \log(1-\hat{p}_i^{h_{[-i]}})(1-Y_i)]. \tag{41}$$

Similarly, for other exponential family distribution, e.g., Poisson distribution, the log likelihood function (37) 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

$$SEE = \frac{1}{n_{est}}\sum_{j=1}^{n_{est}}(\hat{f}(x_j) - f(x_j))^2, \tag{42}$$

where $f(x_j) = log(\frac{p_j}{1-p_j})$ with $p_j = p(x_j)$ and $p(x)$ defined in (35). Here x_j's are the points of the partition of $[-1, 1]$ into small intervals of length 0.01 with $n_{est} = 201$.

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

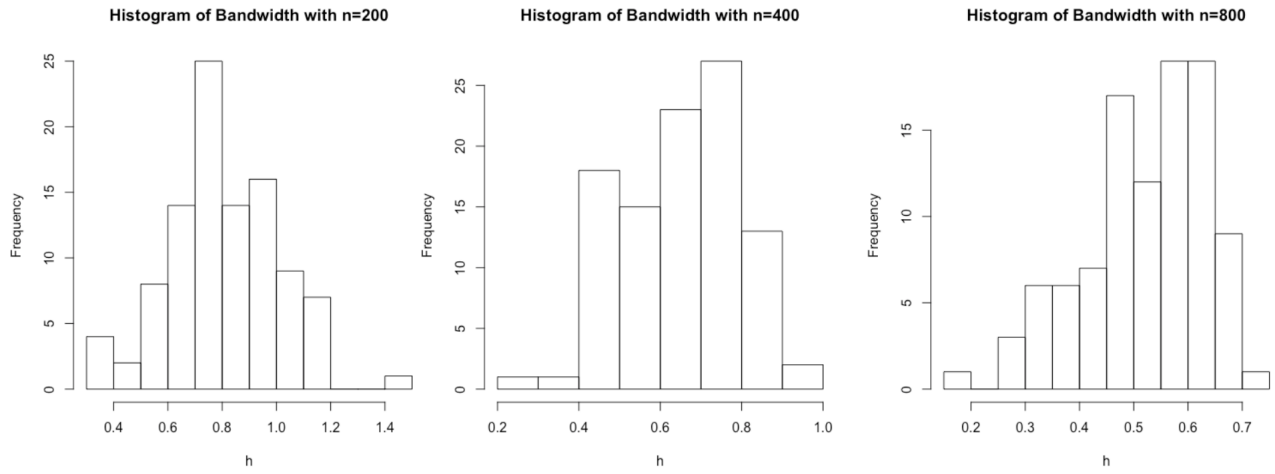

Figure 1. Bandwidth selected for sample size $n = 200$, $n = 400$ and $n = 800$ with 100 repetitions

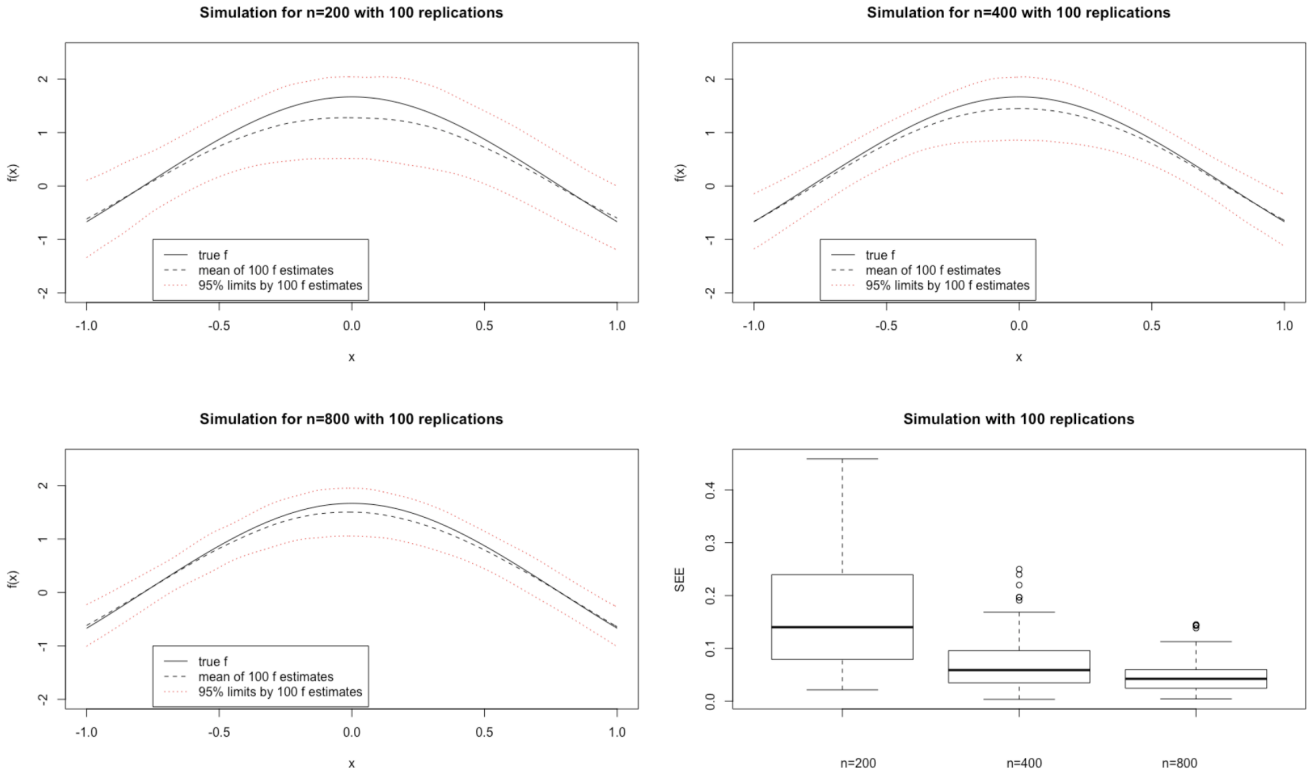

Figure 2. Estimation results of sample size $n = 200$, $n = 400$ and $n = 800$ with 100 repetitions

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 [39]. 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 nonparametric 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 paper. 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 [8], 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) [13]). 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.

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.

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) [20].

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.

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.

Indeed, the model and method developed in this paper can be easily adopted for low-dimensional covariates. The theoretical results are obtained for arbitrary numbers of dimensions of X_t, that we can estimate $E(Y_t|X_t := \{x_{1t}, \ldots, x_{dt}\})$ for any d. However, when the dimension d is high, a so-called "curse of dimensionality" shows that the performance of such nonparametric estimation will deteriorate with the dimension d increasing. This is an active research area in the field of statistics and econometrics, where the results of this paper can help to establish further research for semiparametric modelling in the case of discrete-valued response time series data in future; see, for example, Chen, Li, Linton and Lu (2018)[4] in the continuous valued response case and Peng and Lu (2021)[33] for the binary-valued response case. We leave this for more investigation in future.

5. CONCLUSION

In this paper, 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, an application to COVID-19 dataset is 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.

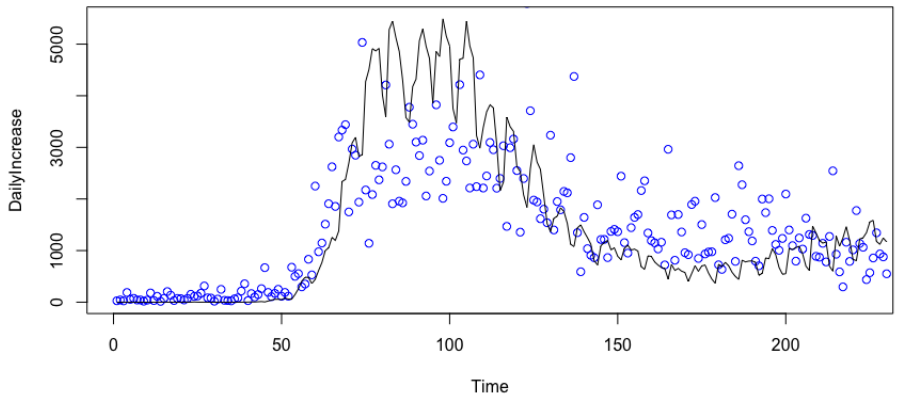

Figure 3. Estimated Daily Increase (Blue dots) based on EPU Index versus Actual Daily Increase (Black line)

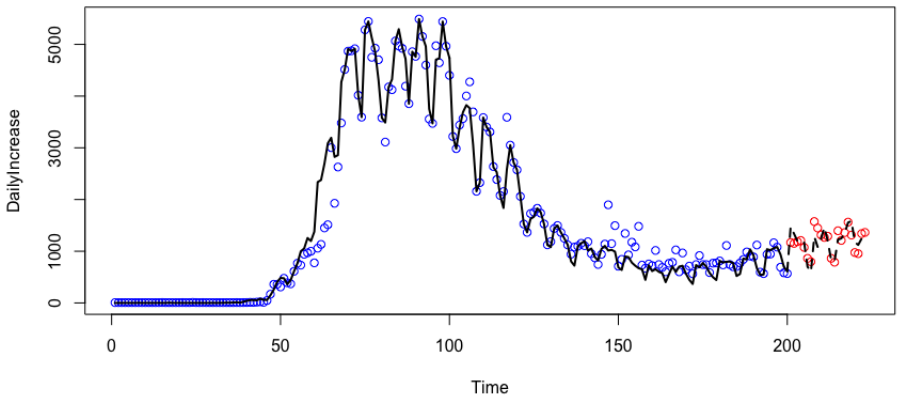

Figure 4. Estimated Daily Increase (blue dots) and Predicted Daily Increase (red dots) base on past information Y_{t-7} versus Actual Daily Increase (black line)

The investigation of nonparametric smoother for time series data is still an active area that can be applied to many disciplines. The results in this paper thus can further contributes to the studies related to count time series data.

ACKNOWLEDGEMENT

The authors are grateful to the Editor-in-Chief, Professor Ming-Hui Chen, and an Associate Editor, a Co Guest-Editor and two referees for their valuable and constructive comments and suggestions, which have greatly helped to improve the presentation of this paper. The research was partially supported by British Academy/Leverhulme Trust (No.SG162909) and NSFC (No.71971131), which are acknowledged.

REFERENCES

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

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

[3] 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.

[4] 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** 919–932.

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

[6] Davis, R. A., Holan, S. H., Lund, R. and Ravishanker, N.

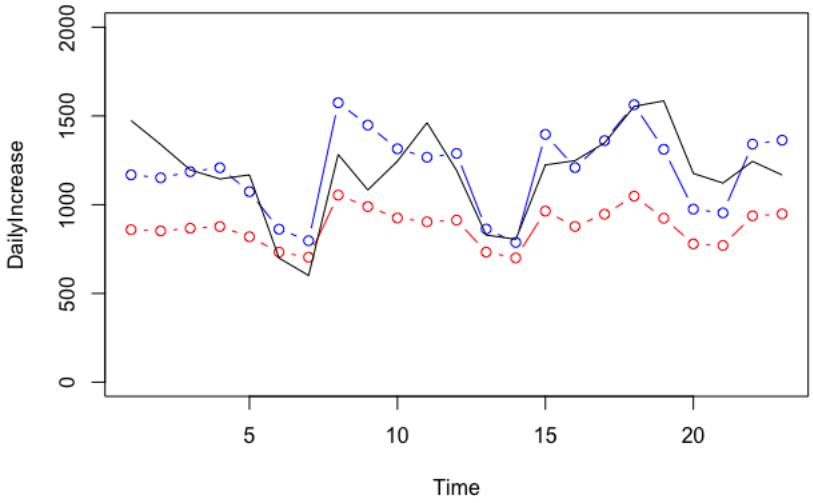

Figure 5. Predicted Daily Increase by Local Linear Regression (blue dots) and Generalised Linear Regression (red dots) based on past information $X_t = Y_{t-7}$, versus Actual Daily Increase (black line)

(2016). *Handbook of discrete-valued time series*. CRC Press.
[7] Doukhan, P., Massart, P. and Rio, E. (1995). Invariance principles for absolutely regular empirical processes. In *Annales de l'IHP Probabilités et statistiques* **31** 393–427.
[8] EconomicPolicyUncertainty (2021). UK Daily News Index. http://www.policyuncertainty.com/uk_daily.html.
[9] Fan, J., Farmen, M. and Gijbels, I. (1998). Local maximum likelihood estimation and inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **60** 591–608.
[10] Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications: monographs on statistics and applied probability 66* **66**. CRC Press.
[11] Fan, J. and Yao, Q. (2003). *Nonlinear time series: nonparametric and parametric methods*. Springer Science & Business Media.
[12] Fan, J., Yao, Q. and Cai, Z. (2003). Adaptive varying-coefficient linear models. *Journal of the Royal Statistical Society: series B (statistical methodology)* **65** 57–80.
[13] Farr, W. (1840). Progress of epidemics. *Second report of the Registrar General of England and Wales* 16–20.
[14] Fokianos, K., Rahbek, A. and Tjøstheim, D. (2009). Poisson Autoregression. *Journal of the American Statistical Association* **104:488** 1430-1439.
[15] Gao, J. (2007). *Nonlinear time series: semiparametric and nonparametric methods*. CRC Press.
[16] Hallin, M., Lu, Z. and Tran, L. T. (2004). Local linear spatial regression. *Annals of Statistics* **32** 2469-2500.
[17] Hansen, B. E. (2008). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory* 726–748.
[18] Hardle, W., Hall, P., Ichimura, H. et al. (1993). Optimal smoothing in single-index models. *The annals of Statistics* **21** 157–178.
[19] Kristensen, D. (2009). Uniform convergence rates of kernel estimators with heterogeneous dependent data. *Econometric Theory* 1433–1445.
[20] 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** 577–582.
[21] Li, D., Lu, Z. and Linton, O. (2012). Local linear fitting under near epoch dependence: uniform consistency with convergence rates. *Econometric Theory* **110** 935–958.
[22] Li, Q. and Racine, J. S. (2007). *Nonparametric econometrics: theory and practice*. Princeton University Press.
[23] Liebscher, E. (1996). Strong convergence of sums of α-mixing random variables with applications to density estimation. *Stochastic Processes and Their Applications* **65** 69–80.
[24] Lu, Z. (1998). On the geometric ergodicity of a non-linear autoregressive model with an autoregressive conditional heteroscedastic term. *Statistica Sinica* **8** 1205-1217.
[25] Lu, Z. (2001). Asymptotic normality of kernel density estimators under dependence. *Annals of the Institute of Statistical Mathematics* **53** 447–468.
[26] Lu, Z. and Linton, O. (2007). Local linear fitting under near epoch dependence. *Econometric Theory* **23** 37–70.
[27] Lu, Z., Tjøstheim, D. and Yao, Q. (2007). Adaptive varying-coefficient linear models for stochastic processes: asymptotic theory. *Statistica Sinica* **17** 177–198.
[28] Masry, E. (1996). Multivariate local polynomial regression for time series: uniform strong consistency and rates. *Journal of Time Series Analysis* **17** 571–599.
[29] Masry, E. and Tjøstheim, D. (1995). Nonparametric Estimation and Identification of Nonlinear ARCH Time Series Strong Convergence and Asymptotic Normality: Strong Convergence and Asymptotic Normality. *Econometric Theory* **11** 258-289.
[30] Mcdonald, D., Shalizi, C. and Schervish, M. (2011). Estimating beta-mixing coefficients. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* 516–524.
[31] Mohr, M. (2020). A weak convergence result for sequential empirical processes under weak dependence. *STOCHASTICS* **92** 140–164.
[32] Nielsen, S. F. (2005). Local linear estimating equations: Uniform consistency and rate of convergence. *Nonparametric Statistics* **17** 493–511.
[33] Peng, R. and Lu, Z. (2021). Semiparametric Averaging

of Nonlinear Marginal Logistic Regressions and Forecasting for Time Series Classification. *Econometrics and Statistics* **https://doi.org/10.1016/j.ecosta.2021.11.001.**
[34] PHAM, D. T. (1986). The mixing properties of bilinear and generalized random coefficient autoregressive models. *Stochastic Processes and their Applications* **23** 291–300.
[35] PHAM, D. T. and TRAN, L. T. (1985). Some strong mixing properties of time series models. *Stochastic Processes and their Applications* **19** 297–303.
[36] TERASVIRTA, T., TJOSTHEIM, D., GRANGER, C. W. et al. (2010). Modelling nonlinear economic time series. *OUP Catalogue*.
[37] TIBSHIRANI, R. and HASTIE, T. (1987). Local likelihood estimation. *Journal of the American Statistical Association* **82** 559–567.
[38] TJØSTHEIM, D. (1990). Nonlinear time series and Markov chains. *Advances in Applied Probability* **22** 587– 611.
[39] UKGOVERNMENT (2021). The official UK Government website for data and insights on Coronavirus. https://coronavirus.data.gov.uk.
[40] WONG, K. C., LI, Z., TEWARI, A. et al. (2020). Lasso guarantees for beta-mixing heavy-tailed time series. *Annals of Statistics* **48** 1124–1142.
[41] XIA, Y. and LI, W. (1999). On single-index coefficient regression models. *Journal of the American Statistical Association* **94** 1275–1285.

Rong Peng
Business School, University of Edinburgh, Edinburgh, EH8 9JS, and
School of Mathematical Sciences, University of Southampton, Southampton, SO17 1BJ
UK
E-mail address: rp2e13@southamptonalumni.ac.uk

Zudi Lu
S3RI and School of Mathematical Sciences, University of Southampton, Southampton, SO17 1BJ
UK
E-mail address: Z.Lu@soton.ac.uk