Received: 18 December 2020 | Accepted: 18 February 2022

DOI: 10.1111/rssa.12849

ORIGINAL ARTICLE

Nonlinear modal regression for dependent data with application for predicting COVID-19

Aman Ullah[1] | **Tao Wang**[1,2] | **Weixin Yao**[3]

[1]Department of Economics, University of California, Riverside, California, USA

[2]Department of Economics, University of Victoria, Victoria, British Columbia, Canada

[3]Department of Statistics, University of California, Riverside, California, USA

Correspondence
Tao Wang, Department of Economics, University of Victoria, Victoria, British Columbia V8W 2Y2, Canada.
Email: taow@uvic.ca

Abstract

In this paper, under the stationary α-mixing dependent samples, we develop a novel nonlinear modal regression for time series sequences and establish the consistency and asymptotic property of the proposed nonlinear modal estimator with a shrinking bandwidth h under certain regularity conditions. The asymptotic distribution is shown to be identical to the one derived from the independent observations, whereas the convergence rate ($\sqrt{nh^3}$ in which n is the sample size) is slower than that in the nonlinear mean regression. We numerically estimate the proposed nonlinear modal regression model by the use of a modified modal expectation–maximization (MEM) algorithm in conjunction with Taylor expansion. Monte Carlo simulations are presented to demonstrate the good finite sample (prediction) performance of the newly proposed model. We also construct a specified nonlinear modal regression to match the available daily new cases and new deaths data of the COVID-19 outbreak at the state/region level in the United States, and provide forward predictions up to 130 days ahead (from 24 August 2020 to 31 December 2020). In comparison to the traditional nonlinear regressions, the suggested model can fit the COVID-19 data better and produce more precise

We are grateful to the joint Editor James Carpenter, an anonymous Associate Editor, and two anonymous referees for helpful comments and suggestions, which have substantially improved an earlier version of the paper. We also thank seminar participants at the 7th Rimini Center for Economic Analysis Time Series Workshop and the 2021 NBER-NSF Time Series Conference for helpful comments.

[Correction added on 28 June 2022 after first online publication: Change in corresponding address.]

predictions. The prediction results indicate that there are systematic differences in spreading distributions among states/regions. For most western and eastern states, they have many serious COVID-19 burdens compared to Midwest. We hope that the built nonlinear modal regression can help policymakers to implement fast actions to curb the spread of the infection, avoid overburdening the health system and understand the development of COVID-19 from some points.

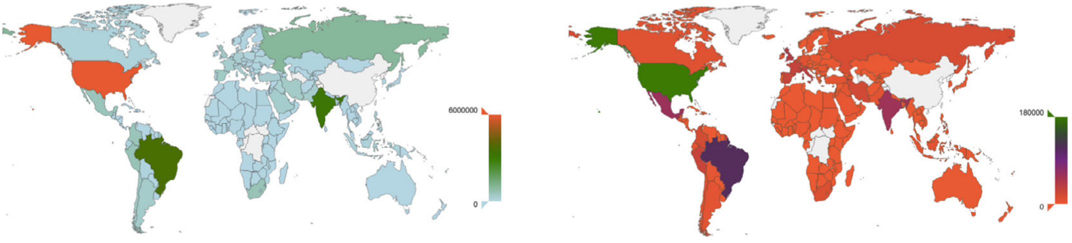
KEYWORDS

COVID-19, dependent data, MEM algorithm, modal regression, nonlinear, prediction

1 | INTRODUCTION

COVID-19 is caused by a coronavirus called SARS-CoV-2 and was identified in Wuhan, the capital city of Hubei province, China, for the very first time in December 2019. On 30 January 2020, the World Health Organization (WHO) declared the COVID-19 outbreak as a Public Health Emergency of International Concern (PHEIC). COVID-19 is a global threat spreading exponentially rather than linearly, that is, the number of new cases is proportional to the existing number of cases, which has been dramatically affecting the health and safety of people all over the world. Based on the information from the Johns Hopkins Coronavirus Resource Center (Dong et al., 2020), due to the extensive spread of COVID-19, there are more than 23 million cases of COVID-19 and more than 800 thousand deaths worldwide as of 23 August 2020 (Figure 1 shows that compared to other countries, the United States (US) has suffered from COVID-19 in a more severe way (Yancy, 2020)). In the US alone, since the first US case of COVID-19 infection was identified in Washington state on 20 January 2020, more than 5.6 million COVID-19 cases and 170 thousand COVID-19 deaths have been identified across the US up to 23 August 2020 (Figure 2 indicates the urgency and necessity of providing reliable predictions to understand the growth behaviour of COVID-19 in the US). WHO quotes 3.4% as the fatality rate (% people who contract the coronavirus and then die). The ongoing global outbreak of the COVID-19 pandemic, which was eventually classified as a pandemic on 11 March 2020 by WHO, poses serious challenges for countries/regions worldwide in designing tailored methods of epidemic control to provide effective and reliable health protection while allowing as much as possible societal and economic activity. It is unclear to anyone where this pandemic will lead us. In such an emergency situation without globally effective antiviral drugs for treating COVID-19 infections, a reliable prediction model for COVID-19 data is undoubtedly essential for policymakers to implement fast actions to curb the spread of the infection, avoid overburdening the health system and understand the dynamics of the COVID-19 spread.

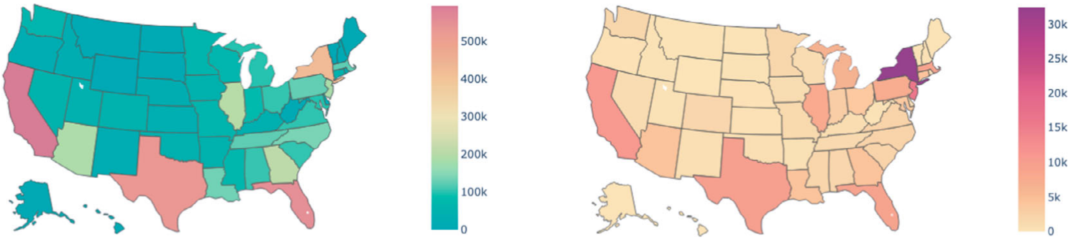
Most of the existing methods for predicting the incidence and prevalence of COVID-19 provided by researchers with backgrounds in epidemiology, biostatistics and economics focus on some mechanistic models, such as the Susceptible–Exposed–Infectious–Recovered (SEIR) model (Grimm et al., 2020; Hauser et al., 2020; Maugeri et al., 2020), the Institute for Health Metrics



(a) COVID-19 Cases

(b) COVID-19 Deaths

FIGURE 1 Visualization of the total number of cases and deaths in the world-23 August 2020; data source: Tencent News https://new.qq.com/ch/antip/ [Colour figure can be viewed at wileyonlinelibrary.com]



(a) COVID-19 Cases

(b) COVID-19 Deaths

FIGURE 2 Visualization of the total number of cases and deaths in the US-23 August 2020; data source: the GitHub repository managed by The New York Times https://github.com/nytimes/covid-19-data [Colour figure can be viewed at wileyonlinelibrary.com]

and Evaluation (IHME) model (IHME, 2020; Jewell et al., 2020) and the Risk-Based model (Barda et al., 2020; Pueyo, 2020), or some statistical models/distributions (Deb & Majumdar, 2020; Fenga, 2020; Linton et al., 2020; Lu et al., 2020; Verity et al., 2020), such as the time series ARMA model and the machine learning model, for the number of cumulative deaths or cases. However, the accuracy of prediction largely depends on the reliability of data, and it is a widespread opinion in the scientific community that the current official COVID-19 data are often noisy with outliers, biased, skewed and/or truncated (Linton et al., 2020; Rudnicki & Piliszek, 2020; Tuli et al., 2020). Therefore, the traditional statistical regression model built on mean might provide low accuracy and even misleading prediction results.

To meet the challenges of the noisy and skewed COVID-19 data, we propose a new statistical regression tool—nonlinear modal regression—that goes beyond the traditional regression models to investigate the dynamic of COVID-19 prevalence in different regions, where the dependent variable of our interest is the official number of daily new cases or new deaths of COVID-19 in a region that could be a state of the US (we concentrate on the daily change value as it is a more representative indicator of epidemic severity and an important metric for assessing the effectiveness of COVID-19 regulation). Note that the built model can be applied to conduct predictions for some regions which are still in the early stage of the COVID-19 pandemic or when the COVID-19 pandemic happens again in the future (there is a growing belief among epidemiologists that COVID-19 will behave similarly to the seasonal flu and re-emerge annually in the winter).

It is well-known that the independence assumption for observations is not always valid in empirical applications. There are many statistical/economics analysis problems with high-dimensional data or information network data, where the data exhibit some sort of dependency, such as Markovian chains, mixing sequences, long-range memory process and so on. In these cases, the statistical properties of the estimator presented in the papers considering independent identically distributed (i.i.d.) samples may change. Because of this, there has been an extensive literature concerning the estimator for dependent data (Bester et al., 2011; Cai & Ould-Said, 2003; Fan & Yao, 2008; Härdle et al., 1997; Pagan & Ullah, 1999; Robinson, 1984). Nevertheless, nearly all of the existing models/methods regarding dependent samples were considered from the mean or quantile regression and are especially useful when there is no outlier in the data, or the density is not very skewed. When the time series dataset contains many outliers (or aberrant observations) or the data are skewed resulting in non-normally estimated standardized residuals (or heavy-tailed error distributions), which is a common feature of financial/macroeconomics/panel time series data, the traditional mean or quantile regression may lose robustness/efficiency or have misspecification (Krief, 2017; Ullah et al., 2021). Thus, modal regression that focuses on the conditional mode, instead of the mean or quantile, of the response variable given the predictor may be more feasible for modelling processes in such cases. Furthermore, when the data are symmetrically distributed, where the modal regression line coincides with the mean regression line, modal regression can overcome the shortcoming of lack of robustness of mean regression to achieve robust and efficient estimators (Yao et al., 2012). To the best of our knowledge, besides Kemp et al. (2020) which considered the estimation of parametric vector autoregressive conditional mode models, there has not been any attempt to estimate modal regression for dependent samples. Substantially different to Kemp et al. (2020), in this paper, we fill the literature gap by focusing on the estimation of a nonlinear modal regression for stationary and weakly dependent samples under α-mixing condition, which is indeed omnipresent in time series econometrics and is less restrictive than other mixing conditions available in the literature. Due to the space constraint, we leave the nonlinear modal-based robust regression for dependent data derived from mode value in another research, which is based on but significantly different from the proposed nonlinear modal regression in the current paper; see Remark 3.

This paper is primarily aimed at applying nonlinear modal regression to understand the characteristics of dependent samples from a mode perspective and settle theoretical properties rigorously. For the simplicity of notations, in what follows, we let $\{(Y_t, X_t)_{t=1}^n\}$ be a stationary discrete-time random process, defined on the probability space $(\Omega, \mathcal{F}, \mathcal{P})$, where Ω denotes the sample space, $\mathcal{F}$ is the σ-algebra (the information) of events, and $\mathcal{P}$ is a probability measure. $\{(Y_t, X_t)_{t=1}^n\}$ has the same marginal distribution as (Y, X), where Y_t is the dependent variable of the main interest and $X_t \in \mathbb{R}^p$ denotes the covariates that may contain the lagged values of Y_t to reflect the dynamical characteristics of the underlying data generating mechanism. Let $f(Y \mid X)$ be the conditional density function of Y given X. The conventional regression model usually employs the mean (or the median) of $f(Y \mid X)$ to model the relationship between Y and X. For example, linear regression assumes that the mean or median of $f(Y \mid X)$ is a linear function of X. The main distinction of modal regression is to find the most probable value/scenario (i.e. mode) of a dependent variable Y given covariates X, which is defined as

$$Mode(Y \mid X) = \arg\max_{Y} f(Y \mid X). \tag{1}$$

When the dimension of X is not low, estimating (1) directly based on nonparametric kernel density estimation will raise many challenges due to the 'curse of dimensionality'. We in this paper avoid directly estimating the conditional density by imposing certain model assumptions on the conditional mode of the response given the covariates (assuming that the global mode is uniquely defined), that is, $Mode(Y \mid X)$; see Section 2 for more details. There is emerging literature on studying modal regression. Due to space limitations, we refer the interested readers to Yao and Li (2014), Chen (2018), Ullah et al. (2021), and the references therein for a comprehensive review of modal regression. Notice that Khardani and Yao (2017) extended the results in Kemp and Santos Silva (2012) to put forward a nonlinear modal regression for the independent samples. However, to the best of our knowledge, there is no existing literature investigating nonlinear modal regression under stationary α-mixing dependent samples using a kernel smoother, which is one of the main contributions of the current paper. It is noteworthy that compared to mean or median regression, modal regression has the following noticeable advantages (Ullah et al., 2021; Yao & Li, 2014): (a) better for reflecting the characteristics of skewed data; (b) better point prediction and narrower prediction intervals; (c) more robust to outliers and certain forms of measurement error; and (d) consistent estimation even for truncated data. Therefore, the modal regression can overcome the limitations of the traditional existing regressions and naturally provide reliable (prediction) models for the noisy COVID-19 data, which is the main innovation of the present paper contributing to the rapidly growing literature on predicting the spread of the current COVID-19 pandemic. We also show a new and interesting theoretical result that the asymptotic theorem for the proposed nonlinear modal estimator under stationary α-mixing dependent samples is the same as that for independent data under certain conditions, indicating the asymptotic negligence of the dependence. This remarkable result is intrinsic for nonparametric estimation for dependent samples and was already observed in the mean regression estimation (Cai & Ould-Said, 2003). Compared to the mean regression estimator, the modal regression estimator depends heavily on error term observations which are confined to the neighbourhood of a given point (i.e. zero) and will unexpectedly have a much slower convergence rate (the modal estimation requires a shrinking bandwidth h due to the use of a small portion of data around the mode), which is the price to be paid in order to estimate mode (Parzen, 1962). The proposed nonlinear modal estimator is relatively simple to implement, where we develop a computationally efficient MEM algorithm in conjunction with Taylor expansion to numerically estimate it.

Generally, most new confirmed cases are infected via contact with the new confirmed cases in recent days, indicating the necessity of incorporating lagged value for analysing and predicting COVID-19 new cases. Ho et al. (2020) introduced a flexible statistical model for the infections and deaths caused by COVID-19 in New Zealand, in which the growth rate of the cumulative number of cases depends on the current cumulative number of cases. Li and Linton (2021) developed a quadratic time trend model that was applied to the log of new cases and obtained satisfying results for the trajectory of the epidemic in most countries. Based on these observations, we apply the proposed nonlinear modal regression to model the log of new cases/deaths as a function of time (to capture the trend or bell-shaped curve) and its own one-step lagged value (to capture the dynamics by autoregressive fluctuations) based on the general structure of the effects and process of infection from a mode perspective; see Section 3 for more details on the model setting for COVID-19 data. Under the constraint imposed by a reasonable length of the paper, we compare the performance of nonlinear modal regression to that of nonlinear mean and median regressions for US COVID-19 data in the paper (for the sake of thoroughness, we also list the results associated with COVID-19 data obtained from the robust nonlinear mean regression with the bisquare weight in Online Appendix B. We emphasize that the outbreak spreads of COVID-19 are largely affected by the policies and social responsibilities of each state/region, it will be interesting in the future to compare the prediction results from the proposed model to some well-known predictions such as those from the IMHE model, SIR models in epidemiology, machine learning methods or other models that can take policy effects into account).

The results indicate that the newly proposed model has good fit performance for most states/regions in the US. We use mean squared error (*MSE*) and mean absolute percentage error (*MAPE*) to compare the out-of-sample prediction performance of the proposed nonlinear modal regression to that of nonlinear mean and median (and robust) regressions for the last 20 days of the samples, where we show that nonlinear modal regression can have considerably more precise predictions. We then apply the proposed nonlinear modal regression to predict COVID-19 new cases and new deaths in the US. Based on the prediction results up to the next 130 days (from 24 August 2020 to 31 December 2020), we can observe that there are systematic differences in spreading distributions across US states. Some states are showing a clear decreasing trend in the number of new cases and new deaths, such as Connecticut, Illinois, Maine, Maryland, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, among others, while others, such as Alabama, Arkansas, California, Florida, Georgia, Mississippi, Montana, North Carolina, North Dakota, Oregon, Texas, Utah, Washington and so on, are still in the first wave of the COVID-19 outbreak. Among the states, California, Florida, Texas and Georgia are the worst affected ones in terms of the number of predicted new cases and new deaths for the next 130 days. For most western and eastern states, they have many serious COVID-19 burdens compared to the Midwest. It is interesting to note that the prediction results may reflect the effect of different possible policy interventions, which can be interpreted as holding the current policies in place or under minimal interventions in each state/region. With the newly developed nonlinear modal regression, we hope that the prediction results can provide some timely information (i.e. turning point) to help policymakers to implement fast actions to curb the spread of the infection and avoid overburdening the health system.

The remainder of this paper is organized as follows. In Section 2, we introduce a nonlinear modal regression for dependent samples under the stationary α-mixing condition and develop an efficient modal estimation algorithm. We also present the asymptotic distributional theory for the resulting estimator under mild conditions, which gives guidelines for practically selecting a reliable bandwidth. Monte Carlo simulations are conducted to show the good finite sample performance of the proposed model. A specified nonlinear modal regression is introduced in Section 3. Based on the built model, we produce a modal multi-step-ahead point forecast framework for COVID-19 new cases and new deaths data, and present the out-of-sample prediction results of the behaviour of COVID-19 at the state/region level in the US. The paper is concluded with some remarks in Section 4. We put additional numerical results, list all figures related to the prediction results, and outline the proofs for the main theorems in the online appendix.

2 | NONLINEAR MODAL REGRESSION

In order to streamline the discussion, we start in this section with the nonlinear modal estimator for dependent samples, where the numerical solutions are obtained via a modified MEM algorithm (Li et al., 2007; Yao, 2013) with the help of a first-order Taylor expansion. Under the assumption of α-mixing, we then present the asymptotic property and optimal bandwidth.

2.1 | Nonlinear modal estimator

As mentioned previously, the traditional method of estimating (1) is to directly estimate the conditional density $f(Y \mid X)$ nonparametrically based on the multivariate kernel density estimation; see the related discussions in Chen et al. (2016). However, due to the 'curse of dimensionality', such a method is practically infeasible when the dimension of covariates is moderate or high, which also contributes to the lack of enough research interest in modal regression. In this paper, similar to mean or quantile regression, we propose estimating the modal regression (1) by imposing some model assumptions on $Mode(Y \mid X)$ directly (assuming that it is uniquely defined) to avoid the 'curse of dimensionality' of the fully nonparametric kernel method. In particular, we assume the following baseline model

$$\begin{cases} Y_t = r(X_t, \beta) + \epsilon_t, \\ Mode(Y_t \mid X_t) = r(X_t, \beta), \quad t = 1, \dots, n, \end{cases} \tag{2}$$

where t represents calendar day that equals to one for the first date of the data, $\beta \in \Theta$ is an unknown parameter vector with dimension p, Θ is the known compact parameter space, $r(\cdot) : \mathbb{R}^p \times \Theta \to \mathbb{R}$ is a parametric nonlinear function measurable on $\mathbb{R}^p$ for each β in Θ, and $\{\epsilon_t\}_{t=1}^n$ is a sequence of stochastic random variables with $Mode(\epsilon_t|\mathcal{F}_t) = 0$ almost surely (a.s.) for every t in which $\mathcal{F}_t$ is the σ-field generated by $\{X_s, \epsilon_s\}_{s\leq t}$. Different from the most existing regressions, we do not impose any second moment conditions on ϵ_t, thus it can be conditional homoscedastic or conditional heteroscedastic. It is worth pointing out that in order to illustrate the applicability of nonlinear modal regression for time series data in a more general setting, we focus on dependent observations. However, Equation (2) could also be an autoregressive time series model with finite order p, that is, $Mode(Y_t|Y_{t-l}) = r(\{Y_{t-l}\}_{l=1}^p, \beta)$, which characterizes the nonlinearity in terms of lags and could be considered as a special case of time series model in this section. The form of $r(\cdot)$ for analysing the COVID-19 data will be discussed in Section 3. Then, the modal parameter β can be estimated by maximizing the following kernel-based objective function given observations $\{(Y_t, X_t)\}_{t=1}^n$ and knowledge of $r(\cdot)$

$$Q_n(\beta) = \frac{1}{nh}\sum_{t=1}^{n} K\left(\frac{Y_t - r(X_t, \beta)}{h}\right), \tag{3}$$

where $K(\cdot)$ is a nonnegatively symmetric kernel function such as the Gaussian kernel (i.e. $K(t) = (2\pi)^{-1/2}\exp[-(1/2)t^2]$) that we will use by default in this paper, and $h := h(n)$ is a bandwidth that is assumed to go to 0 with n going to infinity (':=' denotes 'equals by definition'). To keep the notation simple, we however suppress n throughout the paper. Notice that $K(\cdot)$ is a function following the same rules as a probability density function, for example, it is positive and integrable. However, the role of bandwidth h (control mode) is different from that in nonparametric regression (control smoothness). According to Yao et al. (2012), the choice of kernel function is not very important in modal regression compared to the choice of bandwidth. We choose the Gaussian kernel in this paper for the sake of simplicity. In particular, we can obtain an explicit expression for the M-Step in Algorithm 1.

Remark 1 When $r(X_t, \beta) = \beta^*$, only an intercept term, $Q_n(\beta^*)$ is a kernel density estimate of Y, and thus the maximizer of Equation (3) is the estimated mode of $f(Y)$. Here, we extend this kernel-type objective function to estimate the modal regression parameter β in the regression setting. When $r(X_t, \beta) = X_t^T \beta$ in which T represents the transpose of a matrix or a vector, the modal regression (2) is simplified to the linear modal regression (Kemp & Santos Silva, 2012; Yao & Li, 2014). Note that if $K(t) = 2^{-1} I(|t| \leq h)$, a uniform kernel, then Equation (3) tries to find the curve $r(X_t, \hat{\beta})$ such that the band $r(X_t, \hat{\beta}) \pm h$ contains the largest number of response Y_t, where $\hat{\beta}$ is the modal estimator. Therefore, the modal regression provides more meaningful point predictions and shorter prediction intervals than the mean regression.

It is well-known that the estimation of nonlinear models is a notoriously difficult problem, especially for modal regression, as maximizing (3) does not have an explicit solution. We thus develop a modified MEM Algorithm 1 originally proposed by Li et al. (2007) and Yao (2013) to simplify the computations, which decomposes the optimization (3) into E-Step and M-Step. Given the initial value $\beta^{(0)}$ (e.g. nonlinear least squares (NLS) estimate), we shall repeat the two steps in the algorithm until it converges. Note that if $r(X_t, \beta)$ is a linear function of X_t, then M-Step is just a weighted LS estimation and has an explicit solution. To simplify the computation of M-Step for a general nonlinear function $r(\cdot)$, we approximate $r(X_t, \beta)$ by a first-order Taylor expansion around the current parameter estimate. It can be proved that each iteration of the above algorithm monotonically nondecreases the objective function (3) following the procedures in Yao and Li (2014), that is, at each iteration $Q_n(\beta^{(g+1)}) \geq Q_n(\beta^{(g)})$ and the equality holds if and only if $\beta^{(g+1)} = \beta^{(g)}$. Therefore, the algorithm is very stable and converges. However, for the bandwidth h with a small value, the objective function may have multiple maxima, and there is no guarantee that the MEM algorithm will converge to the global maximizer. Accordingly, it is important to try different starting points on each occasion to compare the values of the target function to choose the best optimal one (Yao & Li, 2014).

Algorithm 1 MEM Algorithm for Nonlinear Modal Regression

E-Step. Calculate the weight $\pi\left(t \mid \beta^{(g)}\right)$

$$\pi\left(t \mid \beta^{(g)}\right) = \frac{K\left(\frac{Y_t - r\left(X_t, \beta^{(g)}\right)}{h}\right)}{\sum_{t=1}^{n} K\left(\frac{Y_t - r\left(X_t, \beta^{(g)}\right)}{h}\right)} \propto K\left(\frac{Y_t - r\left(X_t, \beta^{(g)}\right)}{h}\right),$$

where g is the iteration indicator.

Expansion. Approximate $r(X_t, \beta)$ by a first order Taylor expansion around $\beta^{(g)}$

$$r(X_t, \beta) \approx r\left(X_t, \beta^{(g)}\right) + \frac{\partial r(X_t, \beta)}{\partial \beta^T}\Big|_{\beta = \beta^{(g)}} \left(\beta - \beta^{(g)}\right).$$

M-Step. Update $\beta^{(g+1)}$ by

$$\begin{aligned}
\beta^{(g+1)} &= \arg\max_{\beta} \sum_{t=1}^{n} \left\{ \pi\left(t \mid \beta^{(g)}\right) \log K\left(\frac{Y_t - r\left(X_t, \beta\right)}{h}\right) \right\} \\
&= \left[\sum_{t=1}^{n} \pi\left(t \mid \beta^{(g)}\right) \frac{\partial r\left(X_t, \beta^{(g)}\right)}{\partial \beta} \frac{\partial r\left(X_t, \beta^{(g)}\right)}{\partial \beta^T} \right]^{-1} \left[\sum_{t=1}^{n} \pi\left(t \mid \beta^{(g)}\right) \frac{\partial r\left(X_t, \beta^{(g)}\right)}{\partial \beta} Y_t^{(g)} \right],
\end{aligned}$$

where $Y_t^{(g)} = Y_t - r\left(X_t, \beta^{(g)}\right) + \frac{\partial r\left(X_t, \beta^{(g)}\right)}{\partial \beta^T} \beta^{(g)}$.

Based on the above algorithm, it can be seen that the major difference between the mean regression by the LS estimation and the modal regression lies in the weight $\pi(t|\beta^{(g)})$ used in E-Step. For the LS estimation, each observation has an equal weight $1/n$. On the other hand, for modal regression estimate, the weight $\pi(t|\beta^{(g)})$ calculated in E-step depends on how close Y_t is to the modal regression curve $r(X_t, \beta)$. This weighting scheme allows modal regression to reduce the effect of observations far away from the modal regression curve to achieve robustness, which is one of the advantages of modal regression over mean regression.

2.2 | Asymptotic property

Before proceeding to the asymptotic theorem for the estimator under the α-mixing assumption, it is convenient to introduce some notations that will be used in the remaining part of this section. We define $T_n(x) = T(x) + o_p(s_n)$ (or $O_p(s_n)$) uniformly for $x \in \mathcal{X}$ if $\sup_{x\in\mathcal{X}} |T_n(x) - T(x)| = o_p(s_n)$ $\big($or $O_p\,(s_n)\big)$, and use '$\xrightarrow{d}$' to represent convergence in distribution. We say that $f(n) = o(g(n))$ if for all $c > 0$, there exists some $k > 0$ (not depend on n) such that $0 \le f(n) < cg(n)$ for all $n \ge k$. Let $\|\cdot\|$ denote the Euclidean norm, that is, $\|A\| = [tr(AA^T)]^{1/2}$ in which $tr(A)$ is the trace of the matrix or vector A. For positive sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \asymp b_n$ if $a_n/b_n + b_n/a_n$ is bounded for all large n. To facilitate the derivation of the consistency and asymptotic theorem for the estimator from (3) in a general framework, we impose the following regularity conditions.

C1 The true value of parameter β_0 defined in (2) is in the interior of the known compact parameter space Θ, which is a subset of $\mathbb{R}^p$.

C2 The kernel function $K(\cdot)$ is a nonnegatively symmetric density function with bounded support and integrates to one. It is four times continuous differentiable with all derivatives bounded in absolute value. Furthermore, $\int t^{2+\delta}K^{2+\delta}(t)dt < \infty$ with probability one in which $\delta \in [0,1)$ is a constant.

C3 The regression function $r(\cdot)$ has at least a continuous first derivative on an open set that contains the true parameter point β_0. In addition, $n^{-1}\sum_{t=1}^{n}\{\partial r(X_t, \beta)/\partial\beta\}\{\partial r(X_t, \beta)/\partial\beta\}^T$ converges to a finite positive definite matrix at $\beta = \beta_0$.

C4 The conditional density of ϵ given X denoted by $q(\epsilon\,|\,X) : \mathbb{R} \to \mathbb{R}$ is bounded away from zero and infinity, and has the fourth continuous derivative. $q^{(c)}(\cdot|X)$ denotes the cth derivative of $q(\cdot|X)$. Furthermore, $q(\epsilon\,|\,X) < q(0\,|\,X)$ for all $\epsilon \neq 0$ and X, and the first derivative $q^{(1)}(\epsilon\,|\,X) = 0$.

C5 $\{(Y_t, X_t)\}$ is a stationary α-mixing process, and the mixing coefficient $\rho(n) = \sup_{A\in\mathcal{F}^0_{-\infty}, B\in\mathcal{F}^\infty_t} |P(A\cap B) - P(A)P(B)|$ tending to zero for $n \to \infty$ satisfies $\sum_{n\ge1} n^\gamma(\rho(n))^{\delta/(2+\delta)} < \infty$ for some $\gamma > \delta/(2+\delta)$, where δ is a constant given in C2 and $\mathcal{F}$ is the σ-algebra of events generated by the random variables $\{(Y_t, X_t)\}$. Moreover, there is a sequence of positive integers d_n such that $d_n \to \infty$, $d_n h \to 0$, and $h^4\sum_{k=d_n}^{n}[\rho(k)]^{\delta/(2+\delta)} = o(nh^{-3})$.

C6 As $n \to \infty$, $n^{-1}\sum_{t=1}^{n} q^{(2)}(0\,|\,X_t)\left\{\frac{\partial r(X_t,\beta)}{\partial\beta}\right\}\left\{\frac{\partial r(X_t,\beta)}{\partial\beta}\right\}^T$ converges in probability to a negative definite matrix.

Most of the above conditions have been used in Kemp and Santos Silva (2012), Yao and Li (2014) and Ullah et al. (2021). *Condition C1* is a common condition and can be easily satisfied in practice, as there are no constraints on β. The bounded support in *Condition C2* imposed on the kernel function $K(\cdot)$ is for the brevity of proofs, and may be relaxed somewhat if we impose certain restrictions on the tail of the kernel function; for example, the Gaussian kernel is allowed (Ullah et al., 2021), which is the default kernel used in this paper. *Condition C3* is a commonly used condition on the smoothness of the nonlinear function and the information matrix to ensure the existence of the asymptotic mean and variance for the proposed nonlinear modal estimator, as the modal estimator $\hat{\beta}$ must satisfy $-\frac{1}{nh^2}\sum_{t=1}^{n} K^{(1)}\left(\frac{Y_t - r(X_t,\hat{\beta})}{h}\right) r^{(1)}(X_t,\hat{\beta}) = 0$ where $K^{(1)}(\cdot)$ and $r^{(1)}(\cdot)$ are the first derivatives of $K(\cdot)$ and $r(\cdot)$ respectively. *Condition C4* implies a certain smoothness of $q(\epsilon_t|X_t)$ in the neighbourhood of zero, which is necessary for identification. It imposes that the conditional density of ϵ has a well-defined global mode at zero (Kemp & Santos Silva, 2012; Ullah et al., 2021). It is to be conceded that this assumption is used for simple illustration; when the population is not homogeneous, the proposed method could also be applied to the multimode setting to capture different modal regression lines, under which the newly developed nonlinear modal regression can reveal the possible heterogeneity of COVID-19 development patterns across different states/regions. *Condition C5* is a condition on the data generating process that permits, and is the standard requirement for the α-mixing process, which is used to control the dependence between two random variables as the time distance increases. It is reasonably weak and is known to be satisfied by many stochastic processes, such as the stationary Markov process and the stationary autoregressive-moving average process. A sufficient condition for the mixing coefficient $\rho(n)$ to satisfy *Condition C5* is to set $\rho(n) = O(n^{-d})$ for some $d > 2(\gamma + 1)/\gamma$ (Cai & Ould-Said, 2003). When $\{(Y_t, X_t)\}_{t=1}^{n}$ are independent in which δ=0, the results in this paper also hold. *Condition C6* is the classic rank condition placing restrictions on the moments of covariates, which is necessary for deriving the asymptotic property of the proposed nonlinear modal estimator. All conditions related to bandwidth h are specified for each of the theorems stated below.

We are now in a position where we can state the main asymptotic results for the proposed nonlinear modal estimator. The results are presented in the following Theorems 1 and 2, where the modal convergence rate $\sqrt{nh^3}$ can be considered as a new one in the literature of nonlinear regression models for dependent samples.

Theorem 1 *Under the regularity conditions C1–C6, with probability approaching one, as* $n \to \infty$, $h \to 0$, *and* $nh^5 \to \infty$, *there exists a consistent maximizer* $\hat{\beta}$ *of* (3) *such that*

$$\|\hat{\beta} - \beta_0\| = O_p((nh^3)^{-1/2} + h^2).$$

Theorem 2 *With* $nh^7 = O(1)$, *under the same conditions as Theorem* 1, *the parameter satisfying the consistency result in Theorem* 1 *has the following asymptotic result*

$$\sqrt{nh^3}\left[\hat{\beta} - \beta_0 - \frac{h^2}{2}J^{-1}M\{1 + o_p(1)\}\right] \xrightarrow{d} N\left\{0, \int t^2K^2(t)dtJ^{-1}LJ^{-1}\right\}.$$

If we allow $nh^7 \to 0$, *the asymptotic theorem becomes*

$$\sqrt{nh^3}(\hat{\beta} - \beta_0) \xrightarrow{d} N\left\{0, \int t^2K^2(t)dtJ^{-1}LJ^{-1}\right\},$$

where $J = \mathbb{E}\left[q^{(2)}(0\,|\,X_t)\left\{\frac{\partial r(X_t,\beta)}{\partial\beta}\right\}\left\{\frac{\partial r(X_t,\beta)}{\partial\beta}\right\}^T\right]$, $M = \mathbb{E}\left[q^{(3)}(0\,|\,X_t)\left\{\frac{\partial r(X_t,\beta)}{\partial\beta}\right\}\right]$, *and* $L = \mathbb{E}\left[q(0\,|\,X_t)\left\{\frac{\partial r(X_t,\beta)}{\partial\beta}\right\}\left\{\frac{\partial r(X_t,\beta)}{\partial\beta}\right\}^T\right]$ *at* $\beta = \beta_0$.

The proofs of the above two theorems are outlined in Online Appendix C. For Theorem 1, the first term $(nh^3)^{-1/2}$ in the convergence rates characterizes the magnitude of the estimation variance, while the second term h^2 characterizes the magnitude of the estimation bias. It is necessary to emphasize that these results are consistent with those in Yao and Li (2014) and Ullah et al. (2021) for the i.i.d. data. Theorem 2 shows that the asymptotic bias term is mainly determined by the bandwidth and can be successfully removed under certain undersmoothing conditions. However, the asymptotic mean squared error (*AMSE*) optimal bandwidth h satisfies $h \asymp n^{-1/7}$, which does not meet the condition that $nh^7 \to 0$. Hence, undersmoothing is required, that is, $\lim_{n\to\infty} \sqrt{nh^7} \to 0$, which will be incorporated into this paper when selecting bandwidth in practice. We remark that the asymptotic results hold for both i.i.d. data and dependent samples under mild conditions including strongly mixing (α-mixing). The asymptotic negligence of dependence with a large sample size is intrinsic to nonparametric estimation for dependent samples and it was already observed in the mean regression estimator; see Cai and Ould-Said (2003). This should be expected as a heuristic principle for nonlinear modal regression as well due to the fact that under the α-mixing process, the covariance between random variables ϵ_t and ϵ_j such that $\epsilon_t, \epsilon_j \in (\epsilon - h, \epsilon + h)$ is dominated by the variance of ϵ_t through the conditions imposed on the smoothing parameter; see Lemma 1 in Online Appendix C. Thus, the dependence between the random variables ϵ_t and ϵ_j in a short interval is of 'short memory' which makes them behave as if they were independent (Härdle et al., 1997).

Remark 2 The convergence rate of the proposed nonlinear modal estimator $\hat{\beta}$, $n^{2/7}$ with the MSE-optimal bandwidth, is slower than the root-n convergence rate of the traditional NLS estimator, which is the cost we need to pay in order to estimate the conditional mode (Parzen, 1962). How to improve the convergence rate of the nonlinear modal estimator needs to be researched further in the future. For example, we may assume a certain analytical relationship among mean, median and mode to help estimate the nonlinear modal regression line. Nevertheless, for skewed data with moderate sample size, the modal regression usually provides better prediction performance than the mean and median regressions, as the mode is trying to capture the most likely data points; see the Monte Carlo simulation results in Yao and Li (2014) for cross-sectional data and Ullah et al. (2021) for fixed effects panel data. Our analysis of COVID-19 time series data in Section 3 also shows the superior prediction performance of the proposed nonlinear modal regression over the nonlinear mean and median (and robust) regressions.

Remark 3 It is observed that the proposed nonlinear modal regression focuses on asymmetric data to reveal the characteristics of the data that have been neglected by mean or quantile regression. In practice, it is also common to encounter symmetric data with outliers/aberrant observations or heavy tails. In such a case, we might still be interested in estimating the mean regression, while the proposed modal regression may not be directly applicable owing to the slower convergence rate and the traditional LS estimator is not robust to outliers or heavy tailed data. One way in the literature to handle this kind of data is to utilize robust regression models, like M-estimation, which will lose efficiency for normal errors. We can then supplement the robust regression literature by demonstrating that with symmetric data having only one mode at the centre and a heavy-tailed distribution, the nonlinear modal regression can be used alternatively as a robust regression to achieve robustness and efficiency. Compared to the proposed nonlinear modal regression, the main feature of the nonlinear modal-based robust regression is that we treat bandwidth h as a constant, which does not depend on sample size. Under suitable conditions, we can establish the asymptotic normality for the proposed modal-based robust estimator with $\sqrt{n}$ consistency, and demonstrate that the modal-based robust estimator could be more efficient than the NLS estimator with a heavy-tailed distribution, or as efficient as the NLS estimator with a normal distribution. Due to the space constraint, we leave the detail of the nonlinear modal-based robust regression for dependent data derived from mode value in another research.

2.3 | Optimal bandwidth

Compared to the bandwidth selection method for density estimation in order to estimate modes, it is more challenging to calculate the optimal bandwidth for modal regression, as the value of bandwidth can strongly affect the regression estimates. Particularly, if bandwidth is large enough, the modal regression will instead capture the mean estimate; see Remark 3. In addition, bandwidth plays an important role in estimating dependent observations, as the dependency can be controlled with the observations in a small window. There exist some methods for selecting the optimal bandwidths for nonparametric estimation of conditional modes based on kernel density estimation; see Chen (2018) and Zhou and Huang (2019). However, the methods for bandwidth selection in modal regression by directly imposing structural assumptions on $Mode(Y|X)$ are rather limited. One of them is related to the plug-in bandwidth selection method for linear modal regression, which was presented in Yao and Li (2014) and Ullah et al. (2021) by replacing the unknown quantities with the corresponding estimates. Nevertheless, such a plug-in method places a heavy burden on calculation. In this part, we discuss the asymptotic optimal bandwidth for h and suggest a simple data adaptive method to obtain the bandwidth.

To derive the asymptomatically optimal bandwidth, we minimize the *AMSE* of the proposed nonlinear modal estimator, that is,

$$\mathbb{E}\left\{(\hat{\beta}-\beta_0)^T W(\hat{\beta}-\beta_0)\right\} \approx M^T J^{-1} W J^{-1} M h^4/4 + (nh^3)^{-1}\mathrm{tr}(J^{-1}LJ^{-1}W)\int t^2K^2(t)dt, \quad (4)$$

where the symbol '$c_n \approx d_n$' indicates that $c_n/d_n \to 1$ as $n \to \infty$ and W is a weight function, such as an identity matrix, reflecting which coefficient is more important in inference. Therefore, the asymptotically optimal bandwidth h is

$$\hat{h}_{opt} = \left[\frac{3\int t^2K^2(t)dt\,\mathrm{tr}(J^{-1}LJ^{-1}W)}{M^T J^{-1} W J^{-1} M}\right]^{1/7} n^{-1/7}. \quad (5)$$

If $W = (J^{-1}LJ^{-1})^{-1}$, which is proportional to the inverse of the asymptotic variance of $\hat{\beta}$, then $\mathrm{tr}(J^{-1}LJ^{-1}W) = p$, and we can have

$$\hat{h}_{opt} = \left[\frac{3p\int t^2K^2(t)dt}{M^T L^{-1} M}\right]^{1/7} n^{-1/7}. \quad (6)$$

The optimal bandwidth in the above equation depends on the unknown density $q(\cdot)$ in a complicated manner, which is not available in practice. However, the expression can give some guidelines on how to select the optimal data-driven bandwidth in practice. To simplify the calculations, we can follow Kemp and Santos Silva (2012) to choose the bandwidth, and let $\hat{h} = 1.6\text{MAD}n^{-0.143}$($-0.13$ comes from the rate $-1/7$ and undersmoothing requirement) be a normalized median absolute deviation (MAD) estimate, where

$$MAD = med_j\{|(Y_j - r(X_j, \hat{\beta}_m)) - med_t(Y_t - r(X_t, \hat{\beta}_m))|\}, \tag{7}$$

$\hat{\beta}_m$ represents the corresponding mean estimate, and *med* representing the median value. Besides the above procedure, researchers could also follow the cross-validation method or the weighted integrated squared error method developed in Zhou and Huang (2019) to select the bandwidth.

2.4 | Monte Carlo experiments

To illustrate that the asymptotic result investigated in the above subsection provides a good approximation of the finite sample behaviour of the proposed nonlinear modal estimator, we present two numerical examples based on Monte Carlo experiments (one is shown in Online Appendix A). We mainly focus on asymmetric data and use DGP to represent the data generating process in what follows. For comparison, both nonlinear modal regression and mean regression are considered to estimate parameters with $M = 200$ replications and sample size $n \in \{200, 400, 600, 1000\}$. We examine how estimators behave in finite samples in terms of bias, standard error and MSE,

$$MSE(\hat{\beta}) = \frac{1}{M}\sum_{j=1}^{M} ||\hat{\beta}^{(j)} - \beta||^2$$

in which $\hat{\beta}^{(j)}$ is the estimate in the jth replication and β is the true value. In order to validate the asymptotic normality property, we present the shape of the empirical density of the standardized (recentred and rescaled) modal estimate. The coverage probabilities to assess the prediction performance of the proposed nonlinear modal regression are reported as well.

DGP 1 We generate the dependent data from the following model

$$Y_t = X_{1,t} + \exp(2X_{2,t}) + X_{1,t}\epsilon_t,$$

where $X_{1,t} = -0.3X_{1,t-1} + u_{1,t}$, $u_{1,t} \overset{i.i.d.}{\sim} \mathcal{N}(0, 0.8^2)$, $X_{2,t} = 0.4X_{2,t-1} + u_{2,t}$, $u_{2,t} \overset{i.i.d.}{\sim} \mathcal{N}(0, 0.5^2)$, and $\epsilon_t \overset{i.i.d.}{\sim} 0.5\mathcal{N}(-1, 2.5^2) + 0.5\mathcal{N}(1, 0.5^2)$ with $\mathbb{E}(\epsilon_t) = 0$ and $Mode(\epsilon_t) = 1$ (Ullah et al., 2021; Yao & Li, 2014). We then have

$$\begin{cases} \text{Mean Regression: } \mathbb{E}(Y_t \,|\, X_{1,t}, X_{2,t}) = X_{1,t} + \exp(2X_{2,t}), \\ \text{Modal Regression: } Mode(Y_t \,|\, X_{1,t}, X_{2,t}) = 2X_{1,t} + \exp(2X_{2,t}). \end{cases}$$

Notice that the median value of ϵ_t is around 0.67, which indicates that the nonlinear median regression line is $Median(Y_t \,|\, X_{1,t}, X_{2,t}) = 1.67X_{1,t} + \exp(2X_{2,t})$. For space considerations, we do not present results for median estimates, but they are available upon request.

The simulation results are summarized in Table 1 (β_m represents the coefficients of mean regression), from which we can see that the proposed nonlinear estimation procedure can recover

TABLE 1 Results of simulations—DGP 1

	Modal estimation				Mean estimation			
Sample size	β_1 (SE)	MSE(β_1)	β_2 (SE)	MSE(β_2)	$\beta_{m,1}$ (SE)	MSE($\beta_{m,1}$)	$\beta_{m,2}$ (SE)	MSE($\beta_{m,2}$)
$n = 200$	1.9329 (0.2288)	0.0566	2.0078 (0.0574)	0.0033	1.0073 (0.2371)	0.0560	1.9974 (0.0402)	0.0016
$n = 400$	1.9604 (0.0924)	0.0101	1.9995 (0.0313)	0.0010	1.0000 (0.1816)	0.0328	2.0008 (0.0241)	0.0006
$n = 600$	1.9620 (0.0817)	0.0081	2.0003 (0.0237)	0.0006	0.9956 (0.1440)	0.0207	1.9999 (0.0193)	0.0004
$n = 1000$	1.9620 (0.0603)	0.0051	1.9985 (0.0178)	0.0003	0.9886 (0.1043)	0.0110	1.9990 (0.0153)	0.0002
True value	$\beta_1 = 2$		$\beta_2 = 2$		$\beta_{m,1} = 1$		$\beta_{m,2} = 2$	

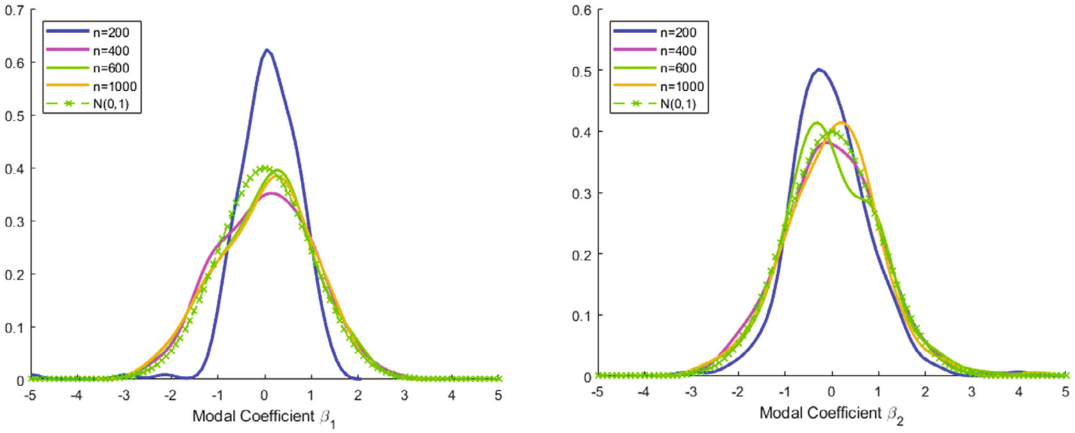


FIGURE 3 Empirical density of the standardized estimate [Colour figure can be viewed at wileyonlinelibrary.com]

modal coefficients well with finite samples. Also, when the sample size increases, the performance of all estimators improves as expected, both in terms of biases and standard errors. With skewed data where the mean, median and mode differ by a location shift, it is necessary to perform the nonlinear modal regression to complement the nonlinear mean or quantile regression and capture the most likely effect that the existing regressions cannot directly reveal.

We present the shape of the empirical density of the standardized modal estimate in Figure 3 to examine the asymptotic normality property of the nonlinear modal estimator. In accordance with the theoretical findings, most of the results manifest asymptotic normality as the sample size n increases. It is noticed that the performance of the asymptotic normality approximation is not perfectly good. We attribute it to the value of the bandwidth selected, which has a substantial effect on the estimation of parameters. How to develop a more efficient way to select the optimal bandwidth needs to be carefully researched in the future.

To show the advantage of the proposed nonlinear modal regression in prediction, we follow Yao and Li (2014) and Ullah et al. (2021) to report the coverage probabilities of prediction intervals of three different lengths ($0.1\sigma, 0.2\sigma, 0.5\sigma, \sigma = \sqrt{\text{Var}(\epsilon_t)} \approx 2$). We use the same DGP procedure as

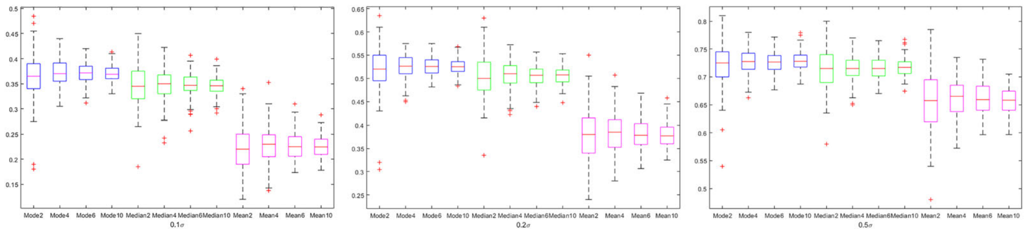


FIGURE 4 Boxplots of average of coverage probabilities: the numbers 2, 4, 6 and 10 represent the values of $n = 200$, 400, 600 and 1000 respectively [Colour figure can be viewed at wileyonlinelibrary.com]

before but implement the out-of-sample prediction with 200 repetitions for the additional n data points. The representative results of the coverage probabilities of the proposed nonlinear modal regression model and the nonlinear mean and median regression models are reported in Figure 4, which shows that in comparison to the nonlinear mean and median regressions, the nonlinear modal regression tends to have superior predictive performance by providing the highest coverage probabilities. Although median regression outperforms mean regression due to the skewness of the error distribution, its performance is worse than that of modal regression. As expected, the nonlinear modal regression and mean and median regressions would have closer coverage probabilities with the increase in the interval length. These simulation findings encourage the use of nonlinear modal regression in prediction.

3 | NONLINEAR MODAL REGRESSION FOR COVID-19 DATA

The prediction advantage of modal regression illustrated in the above section provides underlying support for building a nonlinear modal regression to predict COVID-19. We in this section develop a nonlinear modal regression based on the general structure of the effects and process of infection from a mode view and use it to predict COVID-19 new cases and new deaths in the US, which are the key quantities that determine the epidemic peak. We aim to investigate how well the proposed model could be used to guide the modelling of the dynamic of the spread.

3.1 | Model framework

We first discuss the choice of a nonlinear modal function $r(X_t, \beta)$ according to the transmission characteristics of COVID-19. It has been shown that the COVID-19 spread follows an exponential distribution, and the number of new cases/deaths does not follow a standard distribution like Gaussian or Exponential due to the large number of outliers and noise; see the related literature summarized in Tuli et al. (2020). In addition, Tuli et al. (2020) showed that the COVID-19 cases/deaths data follow the generalized inverse Weibull (GIW) distribution better than the Gaussian, which has the following probability density function (de Gusmão et al., 2011)

$$f(y) = abc^b y^{-(b+1)} \exp\left[-a\left(\frac{c}{y}\right)^b\right], \quad y > 0 \tag{8}$$

with three parameters $a \in \mathbb{R} > 0$, $b \in \mathbb{R} > 0$ and $c \in \mathbb{R} > 0$. It can be easily proven that (8) is a probability density function by substituting $u = -ac^b y^{-b}$. Instead of considering probability distribution, Tuli et al. (2020) treated (8) as a regression function and used it to establish a mean regression model of cross countries COVID-19 prediction between a dependent variable Y_t and time trend t, which is expressed as follows

$$Y_t = abc^b t^{-(b+1)} \exp\left[-a\left(\frac{c}{t}\right)^b\right], \tag{9}$$

where $t > 0$ is the time in the number of days from the first case. Tuli et al. (2020) introduced a machine learning-based iterative weighting strategy to fit (9) with the number of cases data and compared the prediction performance with the Gaussian fitting by *MSE*, *MAPE* and R^2, where they showed that the proposed GIW model performs significantly better.

By coincidence, the same phenomenon, that is, the data with a large number of noise follow a GIW-type shape, appears when we plot the new cases/deaths data against time for most states in the US, which motivates us to develop a regression model for the COVID-19 data in the US based on (9). This paper however does not use this mean regression model directly as it only depends on time t and cannot capture the dynamics of COVID-19. Since previous studies have suggested that the log of new cases/deaths is more suitable to be the dependent variable (Deb & Majumdar, 2020; Li & Linton, 2021; Schüttler et al., 2020; Wang et al., 2020), as the logarithm value can weight more evenly values close to the maximum of the objective function and disregard other values, we then instead take the logarithm on both sides of (9), from which we can see that $\log(Y_t)$ is linearly associated with $\log(t)$ and t^δ (δ is a constant number). We emphasize that it is reasonable to use log value due to the increase in the number of new cases/deaths rose by multiple orders of magnitude in a short period of time and sensible to include t^δ to capture the fact that most states in the US experience a decreasing trend after approaching the peak number of cases/deaths per day (in the early stage of the COVID-19 epidemic, the data usually show an exponential growth trend. After a period of time, as the number of uninfected people decreases, the growth rate starts to decelerate and the number of cases keeps rising until reaching a peak. Subsequently, the number of new infections begins to decline). Furthermore, to incorporate the time series structure of the data and the fact that each infected person will create a chain of new infections, we include the lag variable $\log(Y_{t-1})$ in the model. Along with the above arguments, we propose the following nonlinear modal regression for modelling COVID-19 data by taking into consideration the effect of progress evolving over time

$$\log(Y_t) = \alpha + \beta\log(t) + \eta\log(Y_{t-1}) + \gamma t^\delta + \epsilon_t, \quad t = 2, \dots, n, \tag{10}$$

where error term $\{\epsilon_t\}_{t=2}^n$ is a sequence of stochastic random variables with $Mode(\epsilon_t \mid \mathcal{F}_{t-1}) = 0$ almost surely (a.s.) for model identification in which $\mathcal{F}_{t-1}$ is the σ-field generated by $\{Y_{t-1-s}\}_{s=0}^\infty$. Therefore, the nonlinear modal regression line is defined as

$$r(X_t, \theta) = \alpha + \beta\log(t) + \eta\log(Y_{t-1}) + \gamma t^\delta, \tag{11}$$

where $X_t = (1, \log(t), \log(Y_{t-1}), t)$ and $\theta = (\alpha, \beta, \eta, \gamma, \delta)^T$. Compared to (9), the proposed nonlinear modal regression model can better incorporate other covariates into the mode structure, such as the lag variable or social distance variables. In addition, the new model uses the conditional mode instead of mean or quantile to model the nonlinear relationship among variables. We also note that, although not presented here, the model developed in (10) performs better in terms of *MSE* and *MAPE* than using a polynomial regression for t, the model (11) without the lag variable (Y_{t-1}), and the model (11) with two lag variables (Y_{t-1} and Y_{t-2}).

Different from the mean or median regression, we propose estimating the modal regression (10) using the following kernel-based objective function

$$Q_n(\theta) = \frac{1}{(n-1)h}\sum_{t=2}^{n} K\left(\frac{\log(Y_t) - \alpha - \beta\log(t) - \eta\log(Y_{t-1}) - \gamma t^{\delta}}{h}\right), \tag{12}$$

whose estimation relies on the choice of the regularization parameter—the bandwidth h. We propose to choose the bandwidth according to Kemp and Santos Silva (2012), where we minimize MSE and $MAPE$ for a grid of 50 values of h between $50MAD$ and $0.5MAD(n-1)^{-0.143}$ with $MAD = med_t\{|\log(Y_t) - r_m(X_t, \hat{\theta}_m) - med_t(\log(Y_t) - r_m(X_t, \hat{\theta}_m))|\}$ in which $\hat{\theta}_m(\cdot)$ representing the corresponding NLS estimate.

With the available parameter estimate $\hat{\theta} = (\hat{\alpha}, \hat{\beta}, \hat{\eta}, \hat{\gamma}, \hat{\delta})^T$ obtained from Algorithm 1, we can formulate a k-step ahead prediction to capture the dynamic behaviour of COVID-19 by fitting the nonlinear modal regression (10) recursively for the entire horizon

$$\hat{Mode}(\log(\hat{Y}_{t+k|t})|t+k, \log(\hat{Y}_{t+k-1})) \approx \hat{\alpha} + \hat{\beta}\log(t+k) + \hat{\eta}\log(\hat{Y}_{t+k-1}) + \hat{\gamma}(t+k)^{\hat{\delta}}, \tag{13}$$

where $\log(\hat{Y}_{t+k|t})$ represents the estimate of $\log(Y_{t+k})$ based on the data $\log(Y_1), \ldots, \log(Y_t)$, $\log(\hat{Y}_{t+1}), \ldots, \log(\hat{Y}_{t+k-1})$. Particularly, we pretend the pre-step estimate was the true value of Y_t at the corresponding step and use it as part of the input variable for predicting the next step. To graphically present the prediction procedure, we have the following roadmap

$$Y_t \xrightarrow{(Y_t,\ t+1)} \hat{Y}_{t+1} \xrightarrow{(\hat{Y}_{t+1},\ t+2)} \hat{Y}_{t+2} \xrightarrow{(\hat{Y}_{t+2},\ t+3)} \hat{Y}_{t+3} \quad \cdots \quad \xrightarrow{(\hat{Y}_{t+k-1},\ t+k)} \hat{Y}_{t+k}.$$

Remark 4 To reduce the computation time, we apply the same modal estimates with the bandwidth h, which are constructed using samples $\{Y_t\}_{t=1}^n$ and the corresponding time sequence for all predictions. However, the prediction performance can be improved if we dynamically reestimate modal parameters each time to incorporate the substantial information contained in the intermediate variables $Y_{t+1}, \ldots, Y_{t+k-1}$ about the conditional mode when the pre-stage estimated forecast is added to the samples (e.g. we estimate θ with the data $\{Y_t, t\}_{t=1}^n$ and use the corresponding estimate to predict the value of Y_{n+1}. After that, we use the data $(\{Y_t, t\}_{t=1}^n, \hat{Y}_{n+1})$ to reestimate θ and use the corresponding estimate to predict Y_{n+2}. Iterative this procedure until we achieve all predictions). Although the suggested recursive prediction procedure performs well for COVID-19 data in this paper, we notice that the accuracy of the predictions may deteriorate when k is too large, which is due to the accumulation of errors with the predicting horizon. Therefore, compared to the long-term prediction, the proposed model is better to be used for the short-term prediction.

Remark 5 It is noticed that there is a basic assumption for (13) such that the predicted value $\hat{Y}_{t+k-1}$ performs almost the same as the true value Y_{t+k-1} with $Mode(\epsilon_{t+1} \mid \hat{Y}_{t+k-1}) = 0$, which is the main reason we use '$\approx$' sign in (13). How to release this assumption to provide a more reliable prediction for modal regression needs to be researched further. However, compared to the prediction procedure of the k-step-ahead predictions based only on the observed data, as is standard in macro settings, our procedure should be more reliable. For instance, as mode does not have the additive property, it is difficult to guarantee that $Mode(\eta\epsilon_t + \epsilon_{t+1} \mid Y_{t-1}) = 0$ with equation $\log(Y_{t+1}) = \alpha + \beta\log(t+1) + \eta(\alpha + \beta\log(t) + \eta\log(Y_{t-1}) + \gamma t^{\delta}) + \gamma(t+1)^{\delta} + \eta\epsilon_t + \epsilon_{t+1}$.

Remark 6 We in this paper model the new cases and new deaths datasets with Equation (10), separately, which indicates that the predicted new deaths and new cases do not appear to be linked to each other. Such a univariate model may ignore possible comovements with other available time series. In practice, it is extremely likely that new cases and new deaths are collectively impactful on observable trends, that is, there is a dependency nature in the series. Thus, it is possible to improve predictions and the explanatory power of the model by jointly predicting these two through a nonlinear vector autoregressive modal regression by extending the results in Kemp et al. (2020), that is, $Y_{jt} = r(Y_{-jt}, \{Y_{jt-l}\}_{l=1}^{L}, X_{jt}, \gamma) + e_{jt}$ with finite order L for $j = 1, 2$ in which Y_{-jt} collects all but the jth observation at time t and X_{jt} includes all possible factors that affect both cases and deaths. With the stationary condition and $Mode(e_{jt}|\mathcal{F}_{t-1}) = 0$ in which $\mathcal{F}_{t-1}$ is the σ-filed generated by $\{Y_{-jt}, \{Y_{jt-l}\}_{l=1}^{L}\}$, it can be shown that the estimator of γ is identified and asymptotically normally distributed. In addition, due to the computation burden, we do not compute the confidence interval for predictions. This should be easily carried out based on the bootstrapped modal regression method introduced in Ullah et al. (2021), where we independently draw bootstrapped pseudo samples of residuals from the estimated regression, use the pseudo residual to minus the corresponding mode value to ensure the mode of residual is zero, and then follow the standard procedure as in mean regression to get the modal confidence interval. Future studies could fruitfully explore these issues further.

3.2 | Modal prediction results

We use publicly available COVID-19 data on the daily number of reported cases and deaths to fit the proposed model (we use the case and death data from each state/region to fit the model (11) and fully expect that the parameters vary across the states/regions, as different states/regions are at different stages of the epidemic cycle and have taken different approaches to managing it), and perform an out-of-sample prediction analysis for all states/regions in the US (including the District of Columbia and Puerto Rico) to predict the number of daily new cases and deaths. We remark that the daily data are superior for short-term/medium tactical predicting and are more informative than weekly or monthly data, as they can reflect the turning point of the curve timely and encourage policymakers and people to take flexible actions at any moment. The data of aggregated US COVID-19 cases/deaths we use are from the GitHub repository managed by The New York Times (https://github.com/nytimes/covid-19-data), which was accessed on 24 August 2020 and used to calculate the daily new cases and new deaths data through the differencing transformation (the last date for the data in this paper is 23 August 2020), that is, $\text{New Cases} = \text{Cases}_t - \text{Cases}_{t-1}$ and $\text{New Deaths} = \text{Deaths}_t - \text{Deaths}_{t-1}$. We set all negative values in the new dataset to be zero for calculation. Due to space limitation, we do not put the results of the descriptive statistics of data here, but they are available upon request. Note that this dataset automatically updates every day with new information.

The accuracy and reliability of a model can be tested by comparing the actual values with the predicted values. Following Tuli et al., 2020, we use performance metrics—*MSE* and *MAPE* (lower values indicate better fit)—to determine the residuals between predictions and actual values in order to compare the out-of-sample prediction validity of the proposed nonlinear modal regression and mean and median (and robust) regressions for the last 20 days of the samples (they are treated as validation data, while the other data are used for training)

$$MSE = \frac{1}{20}\sum_{t}(\log(Y_t) - \log(\hat{Y}_t))^2, \tag{14}$$

$$MAPE = \frac{1}{20}\sum_{t}\frac{|\log(Y_t) - \log(\hat{Y}_t)|}{\log(Y_t)} \times 100, \quad t \in \text{ last 20 days.} \tag{15}$$

The model comparison results are summarized in Table 2 (and Table 7 in Online Appendix B), with the best performing model highlighted in bold font. As we can see from Table 2 (and Table 7 in Online Appendix B), the proposed nonlinear modal regression succeeds in predicting the new cases/deaths for 20 days ahead with better accuracy compared to the nonlinear mean and median (and robust) regressions for most states/regions. It has more precise predictions with lower *MSEs* and *MAPEs* for most states with the observed data. Overall, we can see that the proposed nonlinear modal regression model outperforms other competing models in terms of prediction accuracy and can give reliable guidance on the trend of the epidemic in the future. There is no special reason for comparing model predictions in terms of *MSE* and *MAPE* over 20 days, which was chosen arbitrarily. To show the results robust to choosing alternative time horizons, we also compared the prediction performance for the last 30 days of the samples, which does not reveal the large difference in prediction or comparison results.

We then apply the proposed nonlinear modal regression to predict the number of new cases and new deaths for up to 130 days (24 August 2020–31 December 2020) to show how the epidemic has evolved over time, which has some differences from many other papers focusing on the long-term trajectory of COVID-19 using mean regression. To conduct the prediction for the latest 130 days, we use the same bandwidth obtained from the training data (when comparing the model prediction performance) to reestimate nonlinear modal regression with a full sample for each state/region (training data+validation data), and then apply the suggested recursive prediction procedure with the new parameter estimates. We remark that there exist large variations in the parameter fittings, which indicates that long-term predictions are complicated. However, the long-term prediction in comparison with the short-term prediction can provide the pattern of the epidemic. Also, there is an underlying assumption for the prediction results, which is that the data used are reliable and the outbreak will continue to follow the past pattern in the future (Petropoulos & Makridakis, 2020). We acknowledge that this assumption is actually the key issue for predicting the transmission of COVID-19, and it is necessary to update predictions by the suggested model when new information/data is available. However, one advantage of modal regression is that it can cope with some forms of measurement errors. Thus, applying modal regression to predict COVID-19 still has an advantage compared to traditional regressions.

To clearly show the dynamic of the COVID-19 spread in the US, we divide the prediction period into four stages, which are 24 August–30 September, 1 October–31 October, 1 November–31 November and 1 December–31 December. The prediction results are presented in Table 3, from which we can observe that the COVID-19 outbreak in the US is dynamic both in time and across different states/regions. Some states/regions are showing a clear decreasing trend in the number of new cases and new deaths (it is tempting to speculate that this result is due to

TABLE 2 Model comparison results

	New cases-mode		New deaths-mode		New cases-mean		New deaths-mean		New cases-median		New deaths-median	
State/region	***MSE***	***MAPE***	***MSE***	***MAPE***	***MSE***	***MAPE***	***MSE***	***MAPE***	***MSE***	***MAPE***	***MSE***	***MAPE***
AL	0.1712	4.5563	**0.6943**	**35.5300**	0.1734	5.0252	0.6953	36.3657	**0.1384**	**4.4649**	0.9213	50.8288
AK	**0.1569**	**7.5698**	**0.1095**	**62.6997**	0.7199	17.8641	0.1198	79.8023	2.3017	33.9916	0.1682	89.3547
AZ	**0.3095**	**6.4026**	**0.7599**	**24.2938**	0.9339	12.2357	0.8226	24.8236	1.5375	17.1057	3.2440	32.2442
AR	0.1280	4.4861	**0.4956**	**23.8808**	0.1595	4.9279	0.6668	28.4478	**0.0868**	**3.7454**	0.9910	34.4291
CA	**0.2040**	**4.3817**	**0.5272**	13.5211	0.2576	4.6241	0.5313	**13.4917**	12.0594	36.3353	0.6386	13.9525
CO	**0.1226**	**4.7919**	**1.2861**	**63.6700**	0.1576	5.7275	1.5153	72.3817	0.4542	10.7173	2.0659	66.9498
CT	**5.0400**	**44.6875**	**1.1784**	**133.0601**	5.4543	51.1498	2.1439	179.9120	5.2566	48.3522	6.5791	171.9621
DE	**0.2388**	7.7782	**0.4570**	85.1702	0.2454	8.3739	0.8193	125.5970	0.2779	**5.6415**	0.7644	**36.6193**
DC	**0.4210**	**14.2005**	**0.4204**	82.0909	0.7385	19.4940	1.6516	175.4791	0.4884	15.3523	0.8240	**55.8350**
FL	**0.1081**	**3.1617**	**0.2720**	**8.9638**	0.2255	4.4887	0.4822	11.2948	0.1124	3.4112	0.3507	9.4744
GA	**0.0933**	**3.1374**	1.17771	22.4311	0.0953	3.2290	1.3582	24.5763	0.2448	5.4379	**0.6647**	**16.6072**
HI	**0.8501**	**15.0967**	**0.3821**	**73.1096**	4.5141	36.6774	0.5695	95.8169	8.1706	51.3927	0.6140	97.5388
ID	**0.1737**	**6.1939**	**0.9100**	**43.6542**	0.4780	10.1412	1.1558	49.4825	1.5833	20.1322	2.6503	91.1905
IL	**0.2880**	**5.9163**	2.0747	**45.7357**	0.3712	7.1152	2.8030	523.2636	0.5619	9.2363	**1.6185**	48.3631
IN	0.1591	5.0592	**0.5063**	**29.2463**	0.3958	8.5265	1.5668	40.3907	**0.1087**	**4.2200**	1.1215	51.2706
KS	**1.3407**	19.9173	**0.6090**	**42.7460**	1.3489	19.4022	0.6648	44.1607	1.5040	**18.6129**	0.6261	42.8812
KY	**0.2305**	**5.9564**	0.4734	**21.5414**	0.3371	8.3026	0.8958	36.7502	0.2675	7.3576	**0.4541**	23.6336
LA	**6.5341**	14.2872	**2.1630**	**29.0675**	6.5771	**11.4450**	2.9178	42.0480	6.5429	19.8462	2.6146	37.7073
IA	**0.1406**	**5.1942**	1.0872	39.7959	0.1445	5.5619	1.2304	44.0175	0.2799	7.1513	**0.4754**	**23.9231**
ME	0.7938	**23.4699**	**0.1436**	**43.9784**	0.8567	25.3532	0.1676	91.5552	**0.5190**	31.6516	0.3118	124.5080

TABLE 2 (Continued)

	New cases-mode		New deaths-mode		New cases-mean		New deaths-mean		New cases-median		New deaths-median	
State/region	***MSE***	***MAPE***	***MSE***	***MAPE***	***MSE***	***MAPE***	***MSE***	***MAPE***	***MSE***	***MAPE***	***MSE***	***MAPE***
MD	**0.0368**	**2.5438**	**0.8368**	**33.7025**	0.1471	5.1483	1.1255	41.9215	0.0374	2.5885	1.6134	62.2135
MA	**0.1779**	**6.1324**	**0.4949**	**27.5922**	0.4097	9.5495	1.6480	42.7498	4.0086	84.2134	2.1522	61.7372
MI	**0.4083**	**7.6109**	**2.0608**	**50.8842**	0.86664	13.7508	3.3902	76.3392	1.1114	15.5525	2.4700	59.1702
MN	**0.0797**	**3.0202**	2.1624	62.5585	0.1180	4.0501	2.8030	73.1131	0.7755	12.7442	**0.2617**	**23.5088**
MS	**0.2104**	**5.5623**	**0.4832**	21.6170	0.2119	5.9219	0.5972	24.6271	0.3400	7.6840	0.4983	**21.1192**
MO	**0.1533**	**4.8664**	**0.4962**	**17.8628**	0.2882	7.1869	0.9254	30.6017	2.0424	19.4914	0.7243	26.3912
MT	**1.0004**	**19.3589**	**0.4924**	60.1072	1.6448	26.1567	0.4806	**59.7919**	5.3211	48.6101	0.7768	95.9944
NE	**0.1410**	**5.4894**	**0.4809**	37.3174	0.4029	9.7425	0.4927	**36.9213**	0.5193	11.0812	0.5364	38.3479
NV	0.0630	**3.0931**	**0.8165**	**32.1520**	**0.0603**	3.0970	1.2392	36.1752	0.2832	7.4281	1.3796	36.8643
NH	**0.2892**	**15.0186**	0.6415	125.4310	0.5377	20.4409	0.7696	142.2721	0.4894	19.4310	**0.3534**	**22.4967**
NJ	**1.0870**	**17.8764**	**0.7045**	**25.4733**	2.0952	25.1485	1.0478	40.9881	1.3815	20.2057	7.0072	153.6538
NM	**0.2643**	**8.5568**	**0.1674**	**21.8560**	0.3769	10.7114	0.3389	28.0368	0.3004	9.3521	0.2964	32.2350
NY	**0.0449**	**3.0657**	**0.6965**	**17.1554**	0.1438	5.3248	1.2468	38.8174	1.1538	15.5820	2.7453	64.9087
NC	0.3222	7.0517	0.5711	27.0946	0.4167	8.2013	**0.5678**	**26.9968**	**0.2933**	**6.6726**	0.6438	28.1160
ND	**0.3865**	**10.4769**	**0.2729**	**26.3197**	0.4928	12.2427	0.5751	72.0739	0.8129	15.8381	0.3398	42.7843
OH	**0.0573**	2.8443	**0.8720**	**46.1271**	0.0643	**2.8045**	1.2982	46.2033	0.3008	7.2960	2.1708	73.6845
OK	**0.1821**	**5.2804**	**1.0089**	**39.1482**	0.2220	5.9597	1.5633	48.5069	1.0923	15.2377	1.4307	46.3320
OR	**0.0407**	**2.7310**	**0.6189**	48.0773	0.0434	2.8604	0.6585	46.4323	1.2971	16.7406	0.9339	**38.1471**
PA	**0.0485**	**2.6820**	2.0029	44.4196	0.1419	4.8088	3.6654	60.8070	0.2093	5.9761	**1.7553**	**42.5175**
PR	**1.1170**	**13.2338**	**1.3052**	**48.8159**	1.4482	15.9853	2.7094	75.2085	2.1473	21.9538	3.3145	83.7110

(Continues)

TABLE 2 (Continued)

	New cases-mode		New deaths-mode		New cases-mean		New deaths-mean		New cases-median		New deaths-median	
State/region	*MSE*	*MAPE*	*MSE*	*MAPE*	*MSE*	*MAPE*	*MSE*	*MAPE*	*MSE*	*MAPE*	*MSE*	*MAPE*
RI	**5.0243**	**41.8085**	**0.3811**	**51.8988**	7.0779	61.0051	1.9419	164.4378	9.9050	70.2124	0.9570	69.1517
SC	**0.0835**	**3.5217**	**0.4014**	**16.4416**	0.2432	6.1175	0.6080	18.1253	0.1153	4.0323	0.7424	20.3135
SD	**0.1802**	**6.7710**	**0.2033**	**21.1258**	0.3197	9.9167	0.2088	23.0946	1.0279	18.9569	0.7436	100.0031
TN	**0.1058**	**3.6718**	**0.3449**	**16.2219**	0.1475	4.2999	0.6678	21.8297	0.2820	6.3490	0.3492	16.3966
TX	**0.0595**	**2.1483**	**0.2397**	**8.5471**	0.3448	5.6411	0.7228	14.0341	0.1099	2.6563	0.2809	9.1692
UT	**0.2180**	6.6142	0.5060	33.4387	0.5707	11.4507	**0.5043**	**33.1763**	0.2272	**6.3151**	0.7189	45.1290
VA	**0.0619**	**2.7185**	**0.9535**	**33.2575**	0.0958	3.7875	1.0224	38.1446	0.1015	3.4359	1.6647	55.0330
WA	0.1298	4.8612	**0.7005**	**26.9027**	0.3073	7.4457	0.7100	27.0861	**0.1055**	**4.0141**	1.0449	31.2122
WV	**0.0965**	**5.6020**	**1.4237**	**84.2442**	0.3392	10.2841	1.5321	87.9995	1.0661	20.3992	2.3734	112.7576
WI	**0.0695**	**3.3494**	**0.6409**	**33.9752**	0.0697	3.3731	1.1918	45.2079	0.1367	4.6163	1.1583	44.6260
VT	**0.2033**	**18.0705**	–	–	0.7570	36.6008	–	–	1.8541	65.3600	–	–
WY	**0.7836**	**11.2654**	–	–	0.7930	12.4550	–	–	0.9138	16.4421	–	–

Notes: When the dataset has zero values, we instead use $\log(Y_t + 1)$ transformation for the whole data. When calculating *MAPE*, we eliminate all log(1)=0 values. For VT and WY, the existing death data are not sufficient for predicting (most values are zero). Thus, we do not have the predicted new deaths results for these two states. The bold numbers represent the best results among modal, mean and median regressions. Particularly, for *MSE* and *MAPE*, bold numbers represent the smallest values. For the sake of thoroughness, we also list the results obtained from the robust nonlinear regression with the bisquare weight in Table 7 in Online Appendix B, where modal regression still shows some advantages. Because R^2 does not translate well to modal and median regressions, we do not report it here for model comparison. However, we calculate the modified R^2 for modal regression using a kernel-based objective function, and the results indicate good fit performance of modal regression.

TABLE 3 Modal prediction results

	Predictions of modal regression 24 August 2020–21 December 2020			Predictions of modal regression 24 August 2020–21 December 2020	
	09/30 10/31 11/30 12/31	**09/30 10/31 11/30 12/31**		**09/30 10/31 11/30 12/31**	**09/30 10/31 11/30 12/31**
State/region	**Total new cases**	**Total new deaths**	**State/region**	**Total new cases**	**Total new deaths**
AL	45570/46151/52293/61847	600/538/545/583	AK	2759/2735/3025/3503	3/3/3/4
AZ	2772/712/737/945	456/528/634/789	AR	38586/45110/56629/73506	288/272/290/326
CA	296260/380190/446670/530800	2295/3086/3740/4693	CO	10461/73826/63510/58704	9/0/0/0
CT	83/0/0/0	0/0/0/0	DE	2360/1704/1450/1329	0/0/0/0
DC	1569/1107/946/881	0/0/0/0	FL	270170/323480/384240/473570	2738/2860/3256/3870
GA	101620/108200/125330/151120	1315/1137/1121/1170	HI	8046/7494/8178/9367	1/1/1/1
ID	8692/9657/11487/14158	45/41/43/47	IL	45842/14954/4759/1301	46/0/0/0
IN	21824/16657/14641/13683	65/0/0/0	KS	7895/6849/6749/7039	55/35/24/17
KY	17470/17932/20270/23934	134/85/61/44	LA	18705/15715/15186/15651	385/230/167/131
IA	19712/16834/16233/16550	93/27/0/0	ME	508/300/206/148	0/0/0/0
MD	15219/8544/5695/4027	18/0/0/0	MA	1050/0/0/0	134/0/0/0
MI	17657/13613/12608/12549	20/0/0/0	MN	32478/32085/35103/40228	63/0/0/0
MS	34407/35406/40023/47283	649/592/599/642	MO	35374/35535/39682/46351	123/52/18/0
MT	12332/15629/17901/21389	12/12/13/15	NE	4193/2000/1059/636	122/115/121/135
NV	22775/24441/28663/34930	279/249/250/267	NH	295/85/19/0	0/0/0/0
NJ	3104/1123/523/264	0/0/0/0	NM	8305/7397/7328/7681	85/31/5/0
NY	12006/6420/4858/4076	68/0/0/0	NC	82554/81715/88581/100270	538/528/579/665
ND	5588/5533/6095/7036	34/28/26/25	OH	45069/47630/52908/61855	311/142/69/28

(Continues)

TABLE 3 (Continued)

	Predictions of modal regression 24 August 2020–21 December 2020			Predictions of modal regression 24 August 2020–21 December 2020	
	09/30 10/31 11/30 12/31	09/30 10/31 11/30 12/31		09/30 10/31 11/30 12/31	09/30 10/31 11/30 12/31
State/region	Total new cases	Total new deaths	State/region	Total new cases	Total new deaths
OK	24683/26744/31587/38723	122/97/87/83	OR	13887/14217/15953/18724	113/101/101/108
PA	15625/10309/8261/7179	20/0/0/0	PR	16332/18541/22792/29017	87/75/77/84
RI	82/0/0/0	0/0/0/0	SC	70597/91851/120560/163670	847/815/871/977
SD	2663/2052/1859/1799	64/63/70/80	TN	73717/77510/89298/107170	672/669/735/847
TX	994910/1679200/2255200/3082600	3222/3464/4222/5328	UT	19034/19507/20285/22113	90/87/92/104
VA	35771/30929/30338/31505	474/371/340/332	WA	44721/63229/81864/108910	498/473/492/540
WV	4456/4308/4623/5217	14/7/4/2	WI	35719/36597/40896/47835	75/29/8/0
VT	89/53/37/28	–	WY	1509/1448/1555/1756	–

Notes: The results represent the total number of modal predicted new cases and new deaths between 24 August and 30 September, between 1 October and 31 October, between 1 November and 30 November, and between 1 December and 31 December, separately.

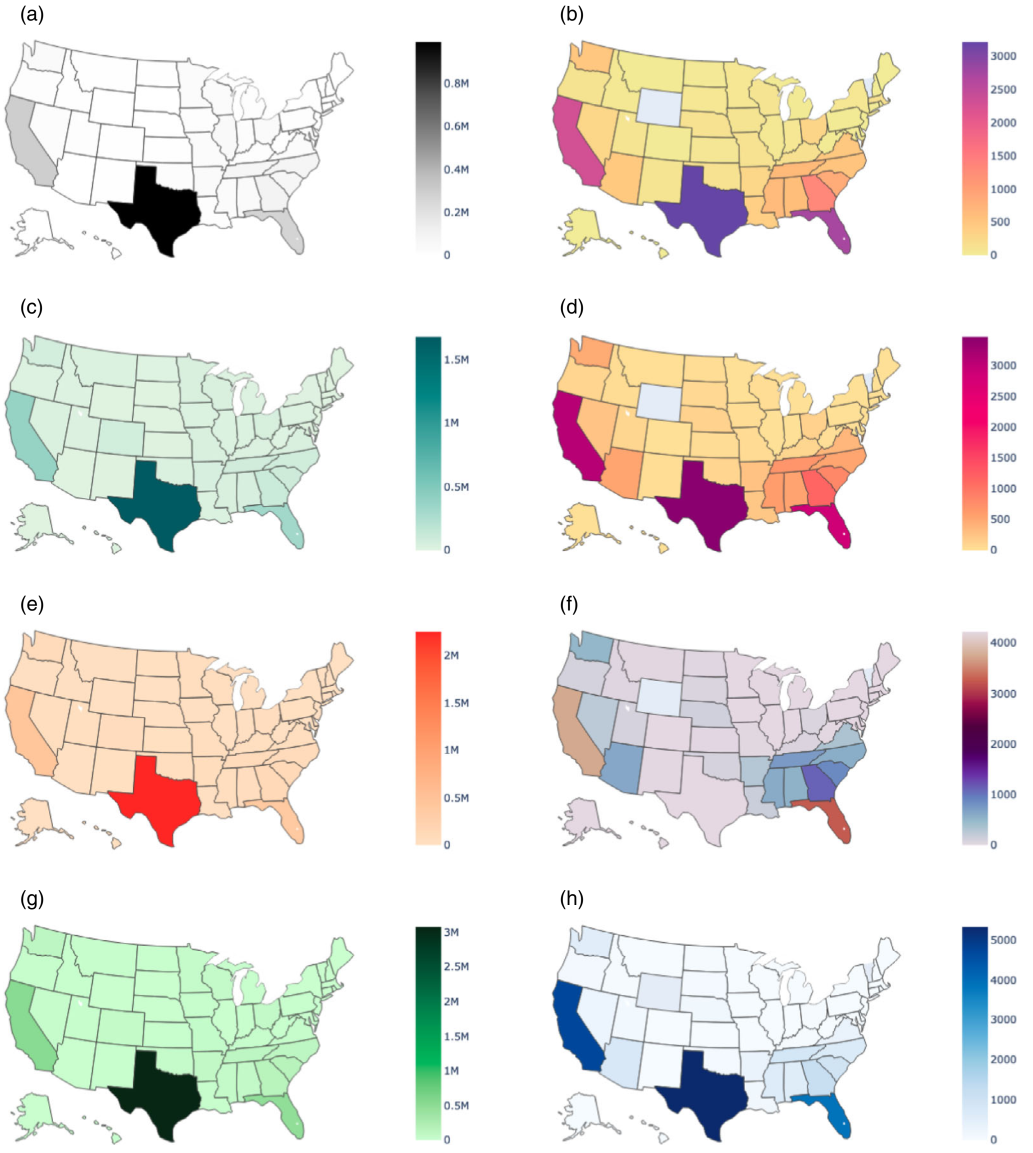


FIGURE 5 Visualization of the Total Number of Modal Predicted New Cases and New Deaths across the US. (a) Predicted new cases 24 August–30 September; (b) Predicted new deaths 24 August–30 September; (c) Predicted new cases 1 October–31 October; (d) Predicted new deaths 1 October–31 October; (e) Predicted new cases 1 November–30 November; (f) Predicted new deaths 1 November–30 November; (g) Predicted new cases 1 December–31 December; (h) Predicted new deaths 1 December–31 December [Colour figure can be viewed at wileyonlinelibrary.com]

the rapid imposition of alert levels and ever tighter lockdowns for these states. The detailed analysis of the effect of lockdowns and social distancing policies on the transmission of COVID-19 is beyond the scope of this paper; see the related discussions in Section 4), for example, Connecticut, Illinois, Maine, Maryland, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, among others, while some other states/regions are still in the first wave of the COVID-19 outbreak with an increase in the number of new cases and new deaths, for example,

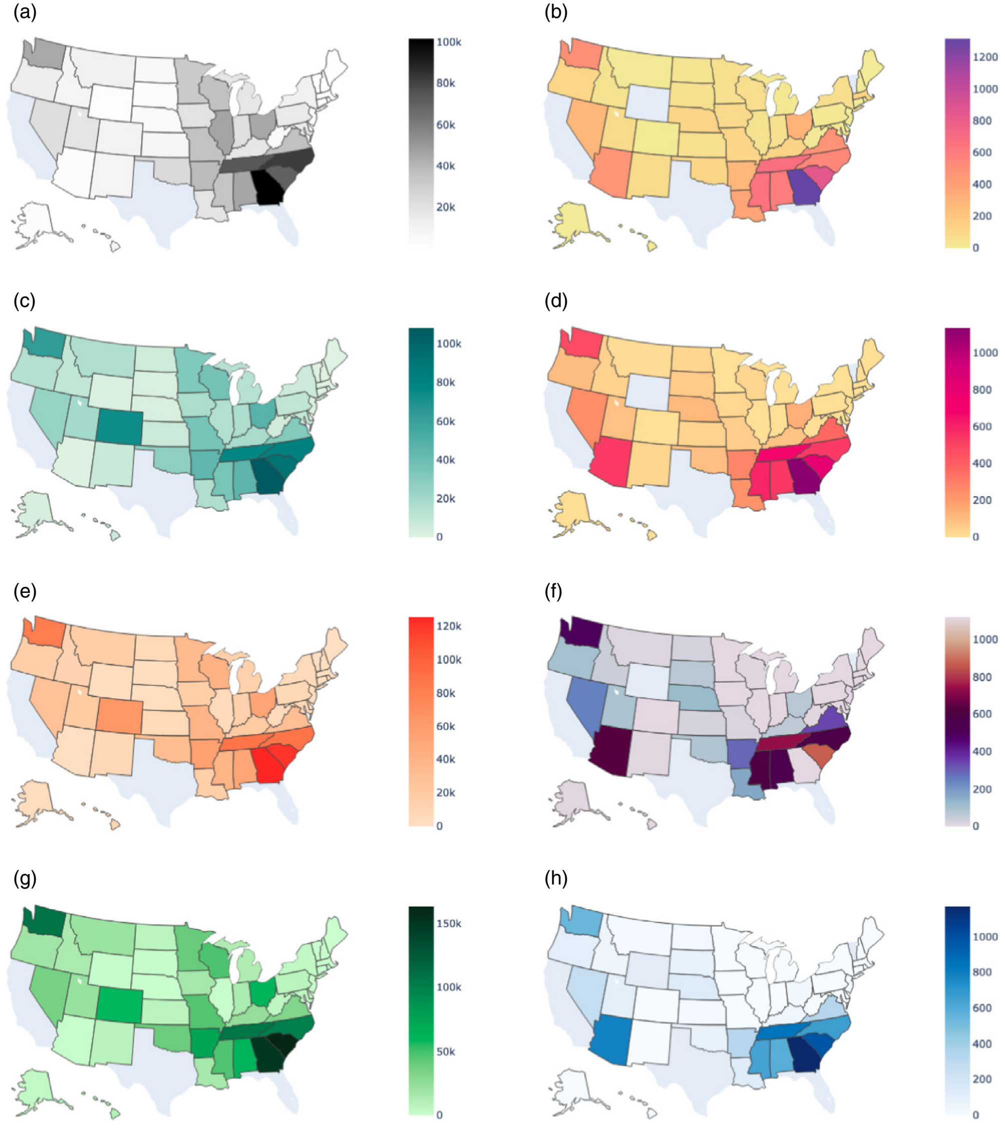


FIGURE 6 Visualization of the Total Number of Modal Predicted New cases and New Deaths across the US after Removing CA, TX and FL. (a) Predicted new cases 24 August–30 September; (b) Predicted new deaths 24 August–30 September; (c) Predicted new cases 1 October–31 October; (d) Predicted new deaths 1 October–31 October; (e) Predicted new cases 1 November–30 November; (f) Predicted new deaths 1 November–30 November; (g) Predicted new cases 1 December–31 December; (h) Predicted new deaths 1 December–31 December [Colour figure can be viewed at wileyonlinelibrary.com]

Alabama, Arkansas, California, Florida, Georgia, Mississippi, Montana, North Carolina, North Dakota, Oregon, Texas, Utah, Washington and so on. Furthermore, as Figure 5 shows (the darker the colour, the more severe the infection), it is clear that there are systematic differences in spreading distributions among states/regions (heterogeneous across the states/regions). In particular, for the next 130 days, California, Florida, Texas and Georgia are the most severe states in terms of the number of predicted new cases and new deaths, which indicates the urgency for these states to take actions to keep social distancing and necessary precautions.

It should be noted that the number of predicted new cases and new deaths across different states/regions has orders of magnitude differences, resulting in the almost uniform colour in Figure 5 for other states/regions having small numbers. To better reveal the situations of other states/regions from visualization, we remove the first three states with the largest numbers of predicted new cases and new deaths from Figure 5. The new visualizing results are presented in Figure 6 which shows a stark heterogeneity across states/regions. We find that for most western and eastern states, the total numbers of new cases and new deaths are incredibly large for the next 130 days based on the prediction results, and these states are in fact experiencing significantly more serious COVID-19 burdens compared to the Midwest (under the stress of economic stagnation, many states/regions have reopened their economies. However, based on the analysis of modal prediction results, it is clear that the outbreak has not been sufficiently controlled in many states up to the date of this paper).

Last but not the least, Online Appendix B contains the nonlinear modal prediction figures (Figure 9) for each state/region (including the District of Columbia and Puerto Rico) in terms of new cases and new deaths, which further demonstrates that the trend of daily confirmed new cases and new deaths is being nicely captured (except for some noisy fluctuations) and the significant new trend is detected by the proposed nonlinear modal regression. Based on these figures, we can also observe that for some states/regions, they have already arrived at a saturation stage and show a decreasing trend for the number of new cases and new deaths, for example, Colorado, Connecticut, Delaware, Maine, Massachusetts, New Hampshire and Pennsylvania, while for some other states/regions, such as Alabama, Arkansas, California, Florida, Idaho, Nebraska, Tennessee and Texas, they will still be at the initial phase of the epidemics and show an increase of the trend for the number of new cases and new deaths if the control and intervention policy is not implemented more effectively. We also list the prediction results for the nonlinear mean regression, median regression, and robust regression (including performance metrics) in Online Appendix B (Figures 10–15 and Tables 5–8), although we have shown that nonlinear modal regression is of higher prediction quality than nonlinear mean and median regressions. The results indicate that there are systemic prediction differences among these models.

4 | CONCLUDING REMARKS

The outbreak of COVID-19 has been unprecedentedly affecting the health and safety of people all over the world, which implies the urgency and importance of accurate prediction. In this paper, we propose a new model, namely parametric nonlinear modal regression for dependent samples, which is particularly useful for handling noisy, skewed, or truncated data (such as the COVID-19 data) and can complement the existing mean or quantile regression. The new model uses the conditional mode instead of mean or quantile to model the nonlinear relationship among variables. We employ a kernel-based objective function to simplify the computation and numerically estimate the proposed model by virtue of a modified MEM algorithm. The asymptotic theorem and the optimal bandwidth are investigated under mild conditions. We then use the proposed nonlinear modal regression to predict the COVID-19 outbreak in the US at the state/region level. We compare the predictions for this novel model with the predictions for nonlinear mean and median (and robust) regressions, and show that the proposed modal regression model can quantify the observed dynamics and provide more precise predictions. Although the outbreak spreads of COVID-19 are largely affected by the policies and social responsibilities of each state/region, we hope that the newly proposed model can be applied to analyse and classify the characteristics of

COVID-19 in the US to provide more timely information to help policymakers to implement fast actions to curb the spread of the infection, avoid overburdening the health system, and understand the development of COVID-19 from some points.

This work paves the way for a number of exciting research directions in the analysis of modal regression and COVID-19. In this paper, we focus on parametric nonlinear modal regression. As pointed out by a referee, the results could be extended to the nonparametric modal regression for dependent samples under α-mixing without imposing any kind of structural assumptions on the data generating process. Specifically, we can employ a kernel-based objective function with the local linear approximation. For our case, as we have both discrete (time variable) and continuous regressors in the model, we need to smooth the discrete variable using discrete kernels such that $\Lambda_\lambda(Z_i, z_0) = \prod_{j=1}^{q} \lambda_j^{\mathrm{I}\{Z_{i,j} \neq z_{0,j}\}}$, where I(.) denotes the usual indicator function, Z_i is a q-dimensional discrete random vector, and $\lambda = (\lambda_1, \ldots, \lambda_q)^T$, $\lambda_j \in [0, 1]$ is the bandwidth for the jth discrete covariate $Z_{i,j}$. We can then make the bandwidths in the discrete kernel be a vector of zeros, and the model will be reduced to the local linear modal regression, which splits the full sample into several subsamples according to different values of the discrete variables. Nevertheless, such a naive sample-splitting method may increase the estimation variance (Li & Racine, 2004). How to derive asymptotic properties and provide asymptotic analysis on the selection of optimal bandwidths for the nonparametric modal regression with mixed discrete and continuous data would be an interesting but challenging future research topic.

Furthermore, in the current paper, we focus on the new cases and new deaths in the US. However, the proposed model could be easily generalized to other countries, say country-level data, and other quantities of interest, for example, cumulative recorded cases and deaths, or the number of people needing hospitalization in an intensive care unit (ICU) each day for a set of regions. Different from the existing research about COVID-19 data, we can also use the proposed model to predict the unconditionally most likely (mode) value of new cases/deaths, which is one of the most important variables/factors when fighting the COVID-19 pandemic. When new cases/deaths reach their mode value, the healthcare system may have the biggest pressure and the largest chance of being overwhelmed, which could in turn affect the death rate. The importance of the mode value can also be seen by noticing that even if the total number of cases is fixed, if we could spread the cases over time and reduce the mode value, the healthcare system can function much better and thus reduce the fatality rate. In addition, it is important to note that the model for COVID-19 presented in this paper has certain limitations (we have to take such predictions reticently, as the prediction error will accumulate over time), as it does not account for any mitigation measures and policy changes. To understand the factors that contribute to the spread of COVID-19, in the future, we could include many other covariates into the model, that is, the factors that might affect new cases/deaths such as social distancing measures as well as the timing of their implementation, the demographics and health condition of the population, the state of the epidemic, the capacity of the healthcare system, the population density and so on. We can then model how the number of cases/deaths depends on the above collected covariates to find out whether there are some clusters of countries having a similar relationship between dependent variable and covariates. Also, it will be interesting to study the spatiotemporal pattern in the spread of COVID-19 by incorporating the spatial correlation into the modal regression.

ORCID

Weixin Yao https://orcid.org/0000-0001-5925-5081

REFERENCES

Barda, N., Riesel, D., Akriv, A., Levy, J., Finkel, U., Yona, G. et al. (2020) Developing A COVID-19 mortality risk prediction model when individual-level data are not available. *Nature Communication*, 11, 4439.

Bester, C.A., Conley, T.G. & Hansen, C.B. (2011) Inference with dependent data using cluster covariance estimators. *Journal of Econometrics*, 165, 137–151.

Cai, Z. & Ould-Said, E. (2003) Local *M*-estimator for nonparametric time series. *Statistics & Probability Letters*, 65, 433–449.

Chen, Y.C. (2018) Modal regression using kernel density estimation: a review. *Wiley Interdisciplinary Reviewers: Computational Statistics*, 10, e1431.

Chen, Y.C., Genovese, C.R., Tibshirani, R.J. & Wasserman, L. (2016) Nonparametric modal regression. *The Annals of Statistics*, 44(2), 489–514.

Deb, S. & Majumdar, M. (2020) A time series method to analyze incidence pattern and estimate reproduction number of COVID-19. *arXiv:2003.10655*.

Dong, E., Du, H. & Gardner, L. (2020) An interactive web-based dashboard to track Covid-19 in real time. *The Lancet Infectious Diseases*, 20(5), 533–534.

Fan, J. & Yao, Q. (2008) *Nonlinear time series: nonparametric and parametric methods*. Berlin/Heidelberg: Springer Science & Business Media.

Fenga, L. (2020) Forecasting the Covid-19 diffusion in Italy and the related occupancy of intensive care units. *medRxiv*. Available from: https://doi.org/10.1101/2020.03.30.20047894

Grimm, V., Mengel, F. & Schmidt, M. (2020) Extensions of the Seir model for the analysis of tailored social distancing and tracing approaches to cope with Covid-19. *medRxiv*. Available from: https://doi.org/10.1101/2020.04.24.20078113

de Gusmão, F.R.S., Ortega, E.M.M. & Cordeiro, G.M. (2011) The generalized inverse Weibull distribution. *Statistical Papers*, 52(3), 591–619.

Härdle, W., Lutkepohl, H. & Chen, R. (1997) A review of nonparametric time series analysis. *International Statistical Review*, 65(1), 49–72.

Hauser, A., Counotte, M.J., Margossian, C.C., Konstantinoudis, G., Low, N., Althaus, C.L. et al. (2020) Estimation of Sars-Cov-2 mortality during the early stages of an epidemic: a modelling study in Hubei, China and Northern Italy. *Plos Medicine*. Available from: https://doi.org/10.1371/journal.pmed.1003189

Ho, P., Lubik, T.A. & Matthes, C. (2020) Forecasting the COVID-19 epidemic: the case of New Zealand. *New Zealand Economic Papers*, doi:10.1080/00779954.2020.1842795

IHME. (2020, August) Institute for Health Metrics and Evaluation Covid-19 estimate. Available from: http://www.healthdata.org/covid

Jewell, N.P., Lewnard, J.A. & Jewell, B.L. (2020) Caution warranted: using the institute for health metrics and evaluation model for predicting the course of the Covid-19 pandemic. *Annals of Internal Medicine*, 173(3), 226–227.

Kemp, G.C.R. & Santos Silva, J.M.C. (2012) Regression towards the mode. *Journal of Econometrics*, 170(1), 92–101.

Kemp, G.C.R., Parente, P.M.D.C. & Santos Silva, J.M.C. (2020) Dynamic vector mode regression. *Journal of Business & Economic Statistics*, 38(3), 647–661.

Khardani, S. & Yao, A.F. (2017) Non-linear parametric mode regression. *Communications in Statistics-Theory and Methods*, 46(6), 3006–3024.

Krief, J.M. (2017) Semi-linear mode regression. *Econometrics Journal*, 20, 149–167.

Li, S. & Linton, O. (2021) When will the COVID-19 pandemic Peak? *Journal of Econometrics*, 220(1), 130–157.

Li, Q. & Racine, J. (2004) Cross-validation local linear nonparametric regression. *Statistica Sinica*, 14, 485–512.

Li, J., Ray, S. & Lindsay, B.G. (2007) A nonparametric statistical approach to clustering via mode identification. *Journal of Machine Learning Research*, 8(8), 1687–1723.

Linton, N.M., Kobayashi, T., Yang, Y., Hayashi, K., Akhmetzhanov, A.R., Jung, S.M. et al. (2020) Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: a statistical analysis of publicly available case data. *Journal of Clinical Medicine*, 9(2), 538.

Lu, F., Nguyen, A., Link, N. & Santillana, M. (2020) Estimating the prevalence of Covid-19 in the United States: three complementary approaches. *medRxiv*. Available from: https://doi.org/10.1101/2020.04.18.20070821

Maugeri, A., Barchitta, M., Battiato, S. & Agodi, A. (2020) Modeling the novel coronavirus (SARS-CoV-2) outbreak in Sicily, Italy. *International Journal of Environmental Research and Public Health*, 17, 4964.

Pagan, A. & Ullah, A. (1999) *Nonparametric econometrics*. Cambridge: Cambridge University Press.

Parzen, E. (1962) On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33, 1065–1076.

Petropoulos, F. & Makridakis, S. (2020) Forecasting the novel coronavirus COVID-19. *Plos One*, 15(3), e0231236.

Pueyo, T. (2020) Coronavirus: Why you must act now. *Politicians, Community Leaders and Business Leaders: What Should You Do and When*.

Robinson, P.M. (1984) Robust nonparametric autoregression. In: Franke, J. (Eds.) *Lecture notes in statistics*, vol. 26. New York: Springer, pp. 247–255.

Rudnicki, W.R. &Piliszek, R. (2020) Estimate of Covid-19 prevalence using imperfect data. *medRxiv*. Available from: https://doi.org/10.1101/2020.04.14.20064840

Schüttler, J., Schlickeiser, R., Schlickeiser, F. & Kröger, M. (2020) Covid-19 predictions using a Gauss model, based on data from April 2. *Physics*, 2, 197–212.

Tuli, S., Tuli, S., Tuli, R. & Gill, S.S. (2020) Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing. *Internet of Things*, 1, 1–16.

Ullah, A., Wang, T. & Yao, W. (2021) Modal regression for fixed effects panel data. *Empirical Economics*, 60, 261–308.

Verity, R., Okell, L.C., Dorigatti, I., Winskill, P., Whittaker, C., Imai, N. et al. (2020). Estimates of the severity of coronavirus disease 2019: a model-based analysis. *The Lancet*, 20(6), 669–677.

Wang, L., Wang, G., Gao, L., Li, X., Yu, S., Kim, M. et al. (2020) Spatiotemporal dynamics, nowcasting and forecasting of COVID-19 in the United States. *arXiv:2004.14103*.

Yancy, C.W. (2020) COVID-19 and African Americans. *The Journal of the American Medical Association*, 323, 1891–1892.

Yao, W. (2013) A note on EM algorithm for mixture models. *Statistics and Probability Letters*, 83, 519–526.

Yao, W. & Li, L. (2014) A new regression model: modal linear regression. *Scandinavian Journal of Statistics*, 41, 656–671.

Yao, W., Lindsay, B.G. & Li, R. (2012) Local modal regression. *Journal of Nonparametric Statistics*, 24(3), 647–663.

Zhou, H. & Huang, X. (2019) Bandwidth selection for nonparametric modal regression. *Communications in Statistics-Simulation and Computation*, 48(4), 968–984.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Ullah, A., Wang, T. & Yao, W. (2022) Nonlinear modal regression for dependent data with application for predicting COVID-19. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 185(3), 1424–1453. Available from: https://doi.org/10.1111/rssa.12849