Received: 29 November 2020 | Accepted: 15 July 2021

DOI: 10.1111/rssb.12453

ROYAL STATISTICAL SOCIETY
Series B Statistical Methodology

DISCUSSION PAPER

Modelling the COVID-19 infection trajectory: A piecewise linear quantile trend model*

Feiyu Jiang[1] | **Zifeng Zhao**[2] | **Xiaofeng Shao**[3]

[1]Department of Statistics, School of Management, Fudan University, Shanghai, China

[2]Department of Information Technology, Analytics, and Operations, Mendoza College of Business, University of Notre Dame, Notre Dame, Indiana, USA

[3]Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, Illinois, USA

Correspondence
Xiaofeng Shao, Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL, 61820, USA.
Email: xshao@illinois.edu

Funding information
China Scholarship Council, Grant/Award Number: 201906210093; National Science Foundation, Grant/Award Number: DMS-20-14053, DMS-18-07032 and DMS-20-14018

Abstract

We propose a piecewise linear quantile trend model to analyse the trajectory of the COVID-19 daily new cases (i.e. the infection curve) simultaneously across multiple quantiles. The model is intuitive, interpretable and naturally captures the phase transitions of the epidemic growth rate via change-points. Unlike the mean trend model and least squares estimation, our quantile-based approach is robust to outliers, captures heteroscedasticity (commonly exhibited by COVID-19 infection curves) and automatically delivers both point and interval forecasts with minimal assumptions. Building on a self-normalized (SN) test statistic, this paper proposes a novel segmentation algorithm for multiple change-point estimation. Theoretical guarantees such as segmentation consistency are established under mild and verifiable assumptions. Using the proposed method, we analyse the COVID-19 infection curves in 35 major countries and discover patterns with potentially relevant implications for effectiveness of the pandemic responses by different countries. A simple change-adaptive two-stage forecasting scheme is further designed to generate short-term prediction of COVID-19 cumulative new cases and is shown to deliver accurate forecast valuable to public health decision-making.

KEYWORDS

change-point detection, forecasting, quantile regression, self-normalization, time series

*Read before The Royal Statistical Society at the Society's 2021 Annual Conference held in Manchester on Wednesday, September 8th, 2021, the President, Professor Sylvia Richardson, in the Chair. 'This discussion paper is linked to two other papers https://doi.org/10.1111/rssa.12875 and https://doi.org/10.1111/rssa.12874 published with discussion contributions and authors' replies in JRSSA 185:4'.

1 | INTRODUCTION

Since the initial outbreak of the novel coronavirus in Wuhan, China in early January 2020, the COVID-19 pandemic has rapidly spread across the world and brought enormous disruption to the economy and society. Various emergency measures, such as social distancing, school closures and economic shutdowns, have been taken by different countries to contain the first waves of the pandemic. To balance saving lives and the economy, a condition-based phased approach has since been adopted in much of the world, including United States and European Union, where the strictness of public health measures is set to adapt in accordance with the present epidemic condition. The success of such a phased approach critically depends on the accurate assessment of the current and near-future status of the pandemic.

One natural and fundamental indicator of the epidemic condition is the trajectory of daily new cases (i.e. the infection curve), as it reflects the transmission rate of the coronavirus and signals the potential impact on the public health system in coming days. This makes the modelling and forecast of the COVID-19 infection curve an important statistical task that can help provide valuable information for public health decision-making of the phased approach. A distinctive feature of the trajectory of new cases is its ever-changing growth rate, as evidenced by the multiple peaks and troughs of the curve (see Figure 1 in Section 4 and Figure S2 in the supplement for trajectories of nine representative countries). Another salient characteristic of the curve is its heteroscedasticity and the presence of notable outliers.

To accurately capture these important features, in this paper, we propose to study the trajectory of new cases (in log scale) via a piecewise linear quantile trend model. Specifically, let $Y_t = \log(R_t + 1)$, where R_t denotes the daily new cases on time t. Here, the addition of 1 in the log-transformation is to handle the case of $R_t = 0$. For a given set of quantile levels $\boldsymbol{\tau}^M = \{\tau_1, \tau_2, \ldots, \tau_M\} \subseteq (0, 1)$, we assume that for all $\tau \in \boldsymbol{\tau}^M$, the τth quantile of $\{Y_t\}_{t=1}^n$ follows

$$\begin{aligned} &Q_\tau(Y_t) = \beta_{0t}(\tau) + \beta_{1t}(\tau) t/n, \quad t = 1, \ldots, n, \\ &(\beta_{0t}(\tau), \beta_{1t}(\tau))^\top = \boldsymbol{\beta}(\tau)^{(i)} = (\beta_0(\tau)^{(i)}, \beta_1(\tau)^{(i)})^\top, \quad \lfloor nq_{i-1} \rfloor < t \le \lfloor nq_i \rfloor, \quad i = 1, \ldots, m_0 + 1, \end{aligned} \tag{1}$$

where $(\beta_{0t}(\tau), \beta_{1t}(\tau))$ is the linear trend (intercept and slope) of the τth quantile $Q_\tau(Y_t)$ at time t and $\mathbf{q} = \{q_1, \ldots, q_{m_0}\} \subseteq (0, 1)$ denotes the $m_0 \ge 0$ change-points with the convention that $q_0 = 0$ and $q_{m_0+1} = 1$. Denote $\boldsymbol{\beta}^{(i)} = (\boldsymbol{\beta}(\tau_1)^{(i)}, \ldots, \boldsymbol{\beta}(\tau_M)^{(i)})$, we require $\boldsymbol{\beta}^{(i)} \ne \boldsymbol{\beta}^{(i+1)}$ for $i = 1, \ldots, m_0 + 1$. Due to the log transformation, the slope $\beta_{1t}(\tau)$ naturally measures the growth rate of the new case at the τth quantile. Note that within each segment, the linear trend $\boldsymbol{\beta}(\tau)^{(i)}$ is allowed to vary across quantile levels and thus naturally accommodates heteroscedasticity. We provide more detailed discussion on the model assumption in Section 3.1.

The piecewise linear quantile trend model is intuitive, interpretable and is useful for tracking the epidemic condition as the change-points naturally segment the trajectory of new cases into phases with (approximately) the same growth rate. The slope of the last segment sheds light on the current pandemic status and helps guide the short-term forecast. The quantile-based approach is more natural and advantageous than the mean-based counterpart due to its built-in robustness to outliers and heteroscedasticity, and its ability to deliver both point and interval forecasts across multiple quantile levels $\boldsymbol{\tau}^M$.

An important part in the estimation of model (1) is to recover the unknown number m_0 and location $\mathbf{q}$ of the change-points. Note that model (1) can be seen as a piecewise quantile regression with deterministic covariates $(1, t/n)$. Though change-point estimation for piecewise linear regression has been extensively studied in the literature, see excellent reviews in Perron

(2006) and Aue and Horváth (2013), piecewise linear quantile regression has not received much attention with only a few exceptions. A sub-gradient testing-based algorithm is studied in Oka and Qu (2011) and a model selection-based procedure is derived in Aue et al. (2014, 2017). However, for methodological and theoretical simplicity, both methods impose temporal independence or martingale difference assumptions on the quantile error process, which can be restrictive for time series applications, see Jiang et al. (2020) for evidence of significant serial dependence in COVID-19 data. Wu and Zhou (2018) study change-point testing in a general M-estimation framework and allow for non-stationary and temporally dependent errors; however, the bootstrap-based testing procedure seems difficult to be extended to change-point estimation.

Based on the self-normalization (SN) idea in Shao (2010, 2015), we propose a novel SN-based change-point detection procedure for the piecewise linear quantile trend model (1) that is robust to temporal dependence and heteroscedasticity. The essential idea of SN is to form an inconsistent variance estimator (i.e. self-normalizer) based on recursive subsample estimates, which is proportional to both the unknown density function at the given quantile levels (due to the non-smooth objective function in quantile regression) and the long run variance (due to temporal dependence). The proposed change-point detection algorithm couples the SN tests computed on nested windows with a local scanning procedure and performs favourably in numerical studies. Due to the use of SN, theoretical results such as segmentation consistency are derived based on a novel non-standard technical argument new to the change-point literature.

Using the piecewise linear quantile trend model and SN-based segmentation method, we analyse the infection curves of COVID-19 (in log scale) in 35 major countries. We find that the spread of coronavirus in each country can typically be segmented into several phases with distinct growth rates and countries with geographical proximity share similar spread patterns, which is particularly evident for continental European countries and developing countries in Latin America. A change-adaptive two-stage forecasting scheme is further designed to generate both point and interval prediction of one-week and two-week ahead cumulative new cases, which delivers accurate forecast valuable to public health decision-making.

Related literature: There are a few recent works in statistics and econometrics literature on COVID-19 modelling. Li and Linton (2021) model the infection and death curves via a quadratic trend function and aim to estimate the inflection point of the pandemic. This work seems less suitable for the current coronavirus trajectories as the pandemic clearly exhibits multiple waves. Liu et al. (2021) analyse country-level infection curves via panel models with piecewise linear trends allowing for one known break-point. Jiang et al. (2020) model the mean trajectory of cumulative cases via a linear trend model allowing for unknown number of change-points, where an SN-based test statistic combined with NOT (a novel segmentation algorithm proposed in Baranowski et al., 2019) is used for change-point estimation. However, no segmentation consistency is provided.

Smooth time-varying coefficient models have also been proposed to capture the changing growth rate of the pandemic. For example, Harvey and Kattuman (2020) introduce the dynamic Gompertz model. Dong et al. (2020) and Gu et al. (2020) propose time-varying coefficient panel and SEIR models respectively to analyse the initial outbreak of the pandemic. In comparison, a change-point model may be more suitable for providing decision-making support for the phased approach to containing the epidemic as it naturally partitions the infection curve into segments with distinct growth rates (thus different severity). Indeed, in real data analysis, we compare with the quantile trend filtering proposed in Brantley et al. (2020), a non-parametric quantile regression method, and demonstrate the advantages of the piecewise linear quantile trend model.

Notably, all the aforementioned works on COVID-19 focus on the modelling of mean. However, as discussed above, the quantile-based approach we adopt here may be more advantageous given its built-in robustness to outliers and ability to capture heteroscedasticity (commonly found in COVID-19 infection curves) and deliver both point and interval forecasts with minimal assumptions.

The rest of the paper is organized as follows. Section 2 proposes the SN-based change-point detection algorithm for the piecewise linear quantile trend model. Section 3 studies the theoretical guarantees of the algorithm. Section 4 conducts real data analysis and demonstrates the promising utility of the proposed methodology for public health decision-making. Section 5 concludes. Simulation studies and technical proofs can be found in the supplementary material.

Some notations used throughout the paper are defined as follows. Given a vector $x = (x_1, ..., x_d)^\top$, denote $\|x\|^q = \sum_{i=1}^d |x_i|^q$ and denote $x^{\otimes 2} = xx^\top$ where $x^\top$ is the transpose of x. For a random vector $X \in \mathbb{R}^d$, denote $\|X\|_q = E(\|X\|^q)^{1/q}$ and $\|X\| = \|X\|_2$. We write $X \in \mathcal{L}^q$ if $\|X\|_q < \infty$. For $a \in \mathbb{R}$, define $\lfloor a \rfloor$ as its integer part.

2 | METHODOLOGY

In this section, we address the key task of estimating the unknown number m_0 and location $\mathbf{q}$ of the change-points in the piecewise linear quantile trend model (1). Section 2.1 proposes a subsample-based SN statistic for change-point testing. Built on the SN test, a novel segmentation algorithm, GOALS, is proposed in Section 2.2 for multiple change-point estimation at a single quantile level τ. Section 2.3 further extends the segmentation algorithm to change-point detection across multiple quantile levels $\boldsymbol{\tau}^M$.

We proceed by introducing two important quantities involved in the proposed algorithm: the global trimming parameter $\varepsilon \in (0, 1/2)$ and the local trimming parameter $\delta \in (0, \varepsilon/2)$. The global trimming parameter ε takes a small value such as $\varepsilon = 0.05, 0.1, 0.15$ and we require that $min_{1 \le i \le m_0+1}(q_i - q_{i-1}) \ge \epsilon$, which is a common assumption on the minimum spacing between change-points in the literature, see Andrews (1993), Bai and Perron (2003), Oka and Qu (2011) and Aue et al. (2014). The local trimming parameter $\delta \in (0, \varepsilon/2)$ is needed for the stability of the subsample-based SN statistic. See more detailed discussions in Sections 2.1 and 2.2.

2.1 | An SN-based test statistic

For the ease of presentation, we first focus on the piecewise linear quantile trend model (1) at a single quantile level τ (i.e. $M = 1$) and leave the extension to multiple quantile levels $\boldsymbol{\tau}^M$ to Section 2.3. Denote $\widehat{\boldsymbol{\beta}}_{t_1,t_2}(\tau)$ as the estimated linear trend parameters based on the subsample $\{Y_t\}_{t=t_1}^{t_2}$ via quantile regression at level τ. For notational simplicity, we omit τ and write $\widehat{\boldsymbol{\beta}}_{t_1,t_2} = \widehat{\boldsymbol{\beta}}_{t_1,t_2}(\tau)$ when no confusion arises.

Given a subsample $\{Y_t\}_{t=t_1}^{t_2}$ and a location $k \in (t_1, t_2)$, to assess the possibility of k being a change-point, it is natural to consider the CUSUM-type contrast statistic $D_n(t_1, k, t_2)$ where

$$D_n(t_1, k, t_2) = \frac{(k - t_1 + 1)(t_2 - k)}{(t_2 - t_1 + 1)^{3/2}} (\widehat{\boldsymbol{\beta}}_{t_1,k} - \widehat{\boldsymbol{\beta}}_{k+1,t_2}). \tag{2}$$

However, the asymptotic distribution of $D_n(t_1, k, t_2)$ depends on the long run variance for the influence function due to unknown temporal dependence and the unknown density function at the quantile level τ, and the consistent estimation of both quantities involves tuning parameters that are notoriously difficult to choose in practice, especially under the presence of change-points.

To bypass these issues, we utilize the SN technique for change-point testing (Shao, 2010; Shao & Zhang, 2010; Zhang & Lavitas, 2018). Specifically, for a subsample $\{Y_t\}_{t=t_1}^{t_2}$ and $k \in (t_1, t_2)$ such that $min(t_2 - k, k - t_1) \geq \lfloor n\epsilon \rfloor$, we define the self-normalizer $V_{n,\delta}(t_1, k, t_2) = L_{n,\delta}(t_1, k, t_2) + R_{n,\delta}(t_1, k, t_2)$, where

$$\begin{aligned} L_{n,\delta}(t_1, k, t_2) &= \sum_{i=t_1+\lfloor n\delta \rfloor}^{k-\lfloor n\delta \rfloor} \frac{(i-t_1+1)^2(k-i)^2}{(k-t_1+1)^2(t_2-t_1+1)^2} (\widehat{\boldsymbol{\beta}}_{t_1,i} - \widehat{\boldsymbol{\beta}}_{i+1,k})^{\otimes 2}, \\ R_{n,\delta}(t_1, k, t_2) &= \sum_{i=k+\lfloor n\delta \rfloor}^{t_2-\lfloor n\delta \rfloor} \frac{(i-1-k)^2(t_2-i+1)^2}{(t_2-t_1+1)^2(t_2-k)^2} (\widehat{\boldsymbol{\beta}}_{i,t_2} - \widehat{\boldsymbol{\beta}}_{k+1,i-1})^{\otimes 2}. \end{aligned} \tag{3}$$

The local trimming parameter $\delta \in (0, \varepsilon/2)$ is introduced to ensure that all the subsample estimates in the self-normalizer $V_{n,\delta}(t_1, k, t_2)$ are constructed with a subsample of size being a positive fraction of n. Intuitively, local trimming removes estimators based on small samples (thus large variance) and helps improve the stability of the self-normalizer. Theoretically, local trimming is needed for the uniform convergence of the deterministic design matrix and weak convergence of the recursive estimates based process $\{\widehat{\boldsymbol{\beta}}_{1,\lfloor nr \rfloor}, r \in [\delta, 1]\}$, see for example Zhou and Shao (2013) and Rho and Shao (2015).

Based on the contrast statistic $D_n(t_1, k, t_2)$ and the self-normalizer $V_{n,\delta}(t_1, k, t_2)$, we define the subsample SN statistic $T_{n,\delta}(t_1, k, t_2)$ such that

$$T_{n,\delta}(t_1, k, t_2) = D_n(t_1, k, t_2)^{\top} V_{n,\delta}(t_1, k, t_2)^{-1} D_n(t_1, k, t_2).$$

Intuitively, a large $T_{n,\delta}(t_1, k, t_2)$ indicates evidence of k being a change-point (see more discussions in Section 2.2). Asymptotically, the presence of $V_{n,\delta}(t_1, k, t_2)$ removes the effects of the unknown temporal dependence and density function on $T_{n,\delta}(t_1, k, t_2)$ and thus helps avoid the estimation of the two quantities. A formal statement of this phenomenon is given by Theorem 1 in Section 3.2. Note that SN can be viewed as a way of prepivoting (Beran, 1987) such that the distribution of SN-based test statistic is pivotal in large sample and does not depend on weak temporal dependence. To conserve space, we refer to Shao (2015) for a more detailed explanation.

2.2 | The GOALS algorithm

To extend change-point testing to multiple change-point estimation, a classical approach in the literature is to combine a change-point test statistic with binary segmentation (Scott & Knott, 1974). However, as pointed out in Baranowski et al. (2019), such a strategy breaks down under the presence of linear trend. In this section, we propose a new segmentation method, the *GlObAl testing + Local Scanning* (GOALS) algorithm, to couple with the SN-based test statistic for multiple change-point estimation in $\{Y_t\}_{t=1}^{n}$ at a single quantile level τ.

Given the global trimming parameter ε, define the window size $h = \lfloor \varepsilon n \rfloor$. For each $k = 1, ..., n$, we define its nested window set $G_n(k)$ where

$$G_n(k) = \{(t_1, t_2) \,|\, t_1 = k - ih + 1,\ i = 1, \ldots, \lfloor k/h \rfloor, t_2 = k + jh,\ j = 1, \ldots, \lfloor (n-k)/h \rfloor\}.$$

Note that for $k < h$ and $k > n - h$, by definition, we have $G_n(k) = \emptyset$. As suggested by its name, GOALS consists of two components.

Step 1. Global-testing: For each $k = 1, ..., n$, we define a global SN test statistic $T_{n,\varepsilon,\delta}(k)$ based on the subsample SN test $T_{n,\delta}(t_1, k, t_2)$ computed on its nested window set $G_n(k)$ such that

$$T_{n,\epsilon,\delta}(k) = \max_{(t_1,t_2)\in G_n(k)} T_{n,\delta}(t_1, k, t_2) = \max_{(t_1,t_2)\in G_n(k)} D_n(t_1, k, t_2)^\top V_{n,\delta}(t_1, k, t_2)^{-1} D_n(t_1, k, t_2),$$

where we set $\max_{(t_1,t_2)\in\emptyset} T_{n,\delta}(t_1, k, t_2) := 0$.

Step 2. Local-scanning: For each $k = 1, ..., n$, we define k as an h-local maximizer if

$$T_{n,\epsilon,\delta}(k) \geq T_{n,\epsilon,\delta}(j), \text{ for all } j \in [k - h + 1, k + h] \cap [1, n].$$

Given a properly chosen threshold ζ_n, we estimate the change-points via

$$(\widehat{k}_1, \ldots, \widehat{k}_{\widehat{m}}) = \{k \,|\, k \text{ is an } h\text{-local maximizer and } T_{n,\epsilon,\delta}(k) > \zeta_n\}.$$

The scale of the maximal SN statistic $T_{n,\varepsilon,\delta}(k)$ computed in the global-testing step reflects the likelihood of k being a change-point. The intuition is as follows. For a non-change-point k, if the nested window $(t_1, t_2) \in G_n(k)$ contains no change-point, its subsample SN statistic $T_{n,\delta}(t_1, k, t_2)$ is expected to be small. If $(t_1, t_2) \in G_n(k)$ contains change-points, the self-normalizer $V_{n,\delta}(t_1, k, t_2)$ is expected to experience inflation as $L_{n,\delta}(t_1, k, t_2)$ and $R_{n,\delta}(t_1, k, t_2)$ are based on contrast statistics and could significantly inflate due to the existence of change-points within (t_1, k) or $(k + 1, t_2)$. Thus, $V_{n,\delta}(t_1, k, t_2)$ inflates along with $D_n(t_1, k, t_2)$, which in turn keeps $T_{n,\delta}(t_1, k, t_2)$ small. Together, the maximal SN statistic $T_{n,\varepsilon,\delta}(k)$ is expected to be small for a non-change-point k. On the other hand, with a sufficiently small global trimming parameter ε such that $\min_{1\leq i\leq m_0+1}(q_i - q_{i-1}) \geq \epsilon$, for any true change-point k, there exists at least one nested window $(\tilde{t}_1, \tilde{t}_2) \in G_n(k)$ which contains k as the only change-point, thus the maximal statistic $T_{n,\varepsilon,\delta}(k)$ is expected to be large thanks to $T_{n,\delta}(\tilde{t}_1, k, \tilde{t}_2)$. Note that the nested window set $G_n(k)$ is discretized to lower computational cost.

The local-scanning step further exploits the notion that $\min_{1\leq i\leq m_0+1}(q_i - q_{i-1}) \geq \epsilon$, since for any change-point k, k is expected to be the h-local maximizer (with high probability) as $[k - h + 1, k + h]$ contains k as the only change-point. The local scanning step avoids sequential estimation of change-points and thus greatly simplifies both numerical computation and theoretical analysis. With a properly chosen threshold ζ_n, the local scanning step achieves segmentation consistency. More detailed discussion on the theoretical and practical choices of the threshold ζ_n is provided in Section 3.2.

The local-scanning component of GOALS is inspired by the Screening and Ranking algorithm (SaRa) in Niu and Zhang (2012) and Hao et al. (2013), which is designed for change-point detection in the mean of a univariate sequence with i.i.d. error. However, SaRa is a *pure* local approach

in that its CUSUM-based test statistic is computed solely on local-windows around k. In contrast, GOALS is a *hybrid* approach as the SN statistic $T_{n,\varepsilon,\delta}(k)$ is computed on a *global* nested window set $G_n(k)$. This nested nature of $G_n(k)$ helps $T_{n,\varepsilon,\delta}(k)$ adaptively retain more power when a change-point k is far away from other change-points by utilizing larger windows that cover k. The substantial power gain from the *hybrid* approach of GOALS over the *pure* local approach of SaRa is further confirmed in unreported simulation studies. A detailed comparison of GOALS with existing works of change-point estimation in quantile regression by Oka and Qu (2011) and Aue et al. (2014) is further given in Section 3.2.

Remark 1 A key element for the success of the *hybrid* approach by GOALS is the self-normalizer $V_{n,\delta}(t_1, k, t_2)$ in the SN statistic. As discussed above, the inflation of the self-normalizer helps keep $T_{n,\delta}(t_1, k, t_2)$ small for a non-change-point k even when (t_1, t_2) contains change-points. Thus, the global SN statistic $T_{n,\varepsilon,\delta}(k)$ achieved by change-point k and non-change-point k are asymptotically well separated, facilitating the selection of the threshold ζ_n. Such a phenomenon does not hold for CUSUM-based tests under the presence of linear trend, see Figure 1 in Baranowski et al. (2019) and the discussion therein.

2.3 | Multi-quantile GOALS

In this section, we propose M-GOALS, a straightforward extension of GOALS that conducts change-point estimation for the piecewise linear quantile trend model (1) simultaneously across multiple quantile levels $\boldsymbol{\tau}^M = (\tau_1, \ldots, \tau_M)$.

For $i = 1, \ldots, M$, denote $\widehat{\boldsymbol{\beta}}_{t_1,t_2}(\tau_i) = (\widehat{\beta}_{0;t_1,t_2}(\tau_i), \widehat{\beta}_{1;t_1,t_2}(\tau_i))^\top$ as the estimated linear trend parameters based on the subsample $\{Y_t\}_{t=t_1}^{t_2}$ via quantile regression at the quantile level τ_i. Furthermore, define $\widehat{\boldsymbol{\beta}}^M_{t_1,t_2} = (\widehat{\beta}_{1;t_1,t_2}(\tau_1), \ldots, \widehat{\beta}_{1;t_1,t_2}(\tau_M))^\top$, which collects the slope estimators across $\boldsymbol{\tau}^M$. As discussed in Section 1, the main task of our analysis is to segment the infection curves into phases with different growth rates, which is captured by the slope parameters $\widehat{\boldsymbol{\beta}}^M_{t_1,t_2}$ of the quantile regression. Thus, we construct the SN statistic based on $\widehat{\boldsymbol{\beta}}^M$ and propose the multi-quantile GOALS (M-GOALS) algorithm.

Step 1. Global-testing: For $k = 1, \ldots, n$, we compute a global SN test statistic $T^M_{n,\epsilon,\delta}(k)$ such that

$$T^M_{n,\epsilon,\delta}(k) = \max_{(t_1,t_2)\in G_n(k)} D^M_n(t_1,k,t_2)^\top V^M_{n,\delta}(t_1,k,t_2)^{-1} D^M_n(t_1,k,t_2), \tag{4}$$

where $D^M_n(t_1, k, t_2)$ and $V^M_{n,\delta}(t_1, k, t_2)$ are defined as in Equations (2)–(3) by replacing $\widehat{\boldsymbol{\beta}}_{i,j}$ with $\widehat{\boldsymbol{\beta}}^M_{i,j}$.

Step 2. Local-scanning: Given a properly chosen threshold ζ^M_n, we estimate the change-points via

$$(\widehat{k}_1, \ldots, \widehat{k}_{\widehat{m}}) = \{k \,|\, k \text{ is a } h\text{-local maximizer and } T^M_{n,\epsilon,\delta}(k) > \zeta^M_n\}.$$

Given the estimated change-points $(\widehat{k}_1, \ldots, \widehat{k}_{\widehat{m}})$ by M-GOALS, quantile regression can be conducted on each estimated segment $\{Y_t\}_{t=\widehat{k}_i+1}^{\widehat{k}_{i+1}}$ to estimate its linear trend parameters for each

quantile level $\tau \in \boldsymbol{\tau}^M$. To ensure the non-crossing constraint across quantiles, we employ the non-crossing quantile regression in Bondell et al. (2010) to simultaneously estimate the linear quantile trend parameters $(\widehat{\beta}_0(\tau)^{(i)}, \widehat{\beta}_1(\tau)^{(i)})$ across all quantile levels $\tau \in \boldsymbol{\tau}^M$.

Choice of trimming parameters (ε, δ): The trimming parameter ε commonly appears in the change-point literature under various context, see for example Andrews (1993), Bai and Perron (2003), Oka and Qu (2011), and Yau and Zhao (2016). Since the value of ε in part reflects the minimum spacing between true change-points, it is typically set at a small value such as 0.05, 0.1, 0.15 in the literature to ensure $\min_{1\le i\le m_0+1}(q_i - q_{i-1}) > \epsilon$. The choice of δ is less essential and we find the performance of (M-)GOALS stable across $\delta \in (0, \varepsilon/2)$. Throughout the paper, we set $\varepsilon = 0.1$ and $\delta = 0.02$, which ensures decent estimation quality of the subsample SN statistics under the moderate sample size n of the COVID-19 data (n ranging from 200 to 300) and is found to perform well in both simulation studies and real data analysis.

In practice, the optimal choice of (ε, δ) may depend on the specific real data application, and different choices of (ε,δ) could lead to different segmentation results by (M-)GOALS. A sensitivity analysis conducted in the supplement suggests that (M-)GOALS gives robust performance across a wide range of (ε, δ) under strong signal-to-noise ratio (SNR), however, its performance may tend to vary across different (ε, δ) under weak SNR. Thus, it is desirable to have a fully data-driven procedure that automatically selects a suitable (ε, δ) based on the observed data.

To this end, in Section S4 of the supplement, we further propose *multi-scanning* M-GOALS, which augments M-GOALS with a model selection based post-processing step and automatically consolidates estimated change-points from M-GOALS with different trimming parameters (ε, δ) via minimizing a quantile regression BIC function adapted from Lee et al. (2014). The multi-scanning M-GOALS is seen to perform well in the simulation studies, matching or exceeding the best performance by M-GOALS across a wide range of (ε, δ). In Section S6 of the supplement, we further re-examine the COVID-19 data using the multi-scanning M-GOALS, and it is seen that the analysis results in Section 4 given by M-GOALS with $(\varepsilon, \delta) = (0.1, 0.02)$ can be reproduced by multi-scanning M-GOALS to a remarkable degree, providing further support for the robustness of our empirical findings about the COVID-19 infection trajectories later presented in the real data section. To conserve space, we refer to Section S4 of the supplement for more details of multi-scanning M-GOALS.

3 | ASYMPTOTIC THEORY

In this section, we establish theoretical guarantees of GOALS and M-GOALS. To keep the presentation clear and intuitive, we derive the theoretical results under the framework of a location-scale model (5), which is an important special case of the piecewise linear quantile trend model (1). The theoretical guarantees for the general model (1) can be established via the same arguments with additional technical assumptions. Specifically, denote $X_t=(1,t/n)^\top$, we assume the observations $\{Y_t\}_{t=1}^n$ is generated from

$$\begin{aligned} &Y_t = X_t^\top \theta_t + (X_t^\top \boldsymbol{\gamma}_t)\varepsilon_t, \quad t = 1, \ldots, n, \\ &(\theta_t^\top, \boldsymbol{\gamma}_t^\top)^\top = (\theta^{(i)\top}, \gamma^{(i)\top})^\top, \lfloor nq_{i-1} \rfloor < t \le \lfloor nq_i \rfloor, \text{ for } i = 1, \ldots, m_0+1, \end{aligned} \tag{5}$$

where $\theta^{(i)} = (\theta_0^{(i)}, \theta_1^{(i)})^\top$, $\boldsymbol{\gamma}^{(i)} = (\boldsymbol{\gamma}_0^{(i)}, \boldsymbol{\gamma}_1^{(i)})^\top$, $\{\varepsilon_t\}$ is a stationary and weakly dependent error process and $\inf_{1\le t\le n} X_t^\top \boldsymbol{\gamma}_t \ge \underset{!!}{c} > 0$.

Essentially, model (5) is an extension of the widely used classical location-scale model with i.i.d. error (Koenker, 2005) to accommodate temporal dependence and structural breaks. Model (5) indicates that the linear trend parameter $\boldsymbol{\beta}(\tau)^{(i)}$ of the τth quantile can be written as $\boldsymbol{\beta}(\tau)^{(i)} = \theta^{(i)} + \boldsymbol{\gamma}^{(i)} Q_\tau(\varepsilon_t)$. Note that model (5) exhibits heteroscedasticity if $\boldsymbol{\gamma}_1^{(i)} \neq 0$, which implies different growth rate $\boldsymbol{\beta}_1(\tau)^{(i)} = \theta_1^{(i)} + \boldsymbol{\gamma}_1^{(i)} Q_\tau(\varepsilon_t)$ across different quantile level τ.

3.1 | Regularity conditions

We regulate the temporal dependence via the physical dependence measure in Wu (2005). Specifically, we assume the error process $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ is stationary and admits the causal representation such that

$$\varepsilon_t = G(e_t, e_{t-1}, \ldots), \quad t \in \mathbb{Z},$$

where $\{e_t\}_{t\in\mathbb{Z}}$ are i.i.d. random variables and $G(\cdot)$ is a measurable function. Let $\{e_t'\}_{t\in\mathbb{Z}}$ be an i.i.d. copy of $\{e_t\}_{t\in\mathbb{Z}}$, we define $\mathcal{F}_t = (\ldots, e_{t-1}, e_t)$ and define $\mathcal{F}_t' = (\mathcal{F}_{-1}, e_0', e_1, \ldots, e_t)$. For a real valued function $H(\cdot)$ such that $H(\mathcal{F}_t) \in \mathcal{L}^p$, we define

$$\delta_H'(j,p) = \|H(\mathcal{F}_j) - H(\mathcal{F}_j')\|_p, \quad p \geq 1, j \in \mathbb{N}.$$

Following Wu (2005), we can view $\mathcal{F}_t$ and $H(\mathcal{F}_t)$ as the input and output of a physical system and the dependence measure $\delta_H'(j, p)$ quantifies the dependence of $H(\mathcal{F}_j)$ on e_0 by measuring the distance between $H(\mathcal{F}_j)$ and its coupled version $H(\mathcal{F}_j')$.

Let $F(\cdot)$ and $f(\cdot)$ be the distribution function and density function of ε_t respectively. For each $\tau \in \boldsymbol{\tau}^M$, we define $\Gamma_\varepsilon(\tau) = f(Q_\tau(\varepsilon))^{-2} \sum_{t=-\infty}^{\infty} \mathrm{Cov}\{\psi_\tau(\varepsilon_0), \psi_\tau(\varepsilon_t)\}$ where $\psi_\tau(u) = \tau - \mathbf{1}(u < Q_\tau(\varepsilon))$. Denote $v_t^M = (\frac{\psi_{\tau_1}(\varepsilon_t)}{f(Q_{\tau_1}(\varepsilon))}, \ldots, \frac{\psi_{\tau_M}(\varepsilon_t)}{f(Q_{\tau_M}(\varepsilon))})^\top$ and define $\Gamma^M = \lim_{n\to\infty} \mathrm{Var}(\frac{1}{\sqrt{n}} \sum_{t=1}^n v_t^M)$. In addition, let $\phi_{\tau;k}(s; \mathcal{F}_t) = \tau - F(\varepsilon_{t+k} + s \,|\, \mathcal{F}_t)$ and let $\phi_{\tau;k}^{(l)}(s; \mathcal{F}_t)$ be the lth derivative of $\phi_{\tau;k}(s; \mathcal{F}_t)$. For $m_0 \geq 1$, define $\mathbf{b}_i(\tau) = \boldsymbol{\beta}(\tau)^{(i+1)} - \boldsymbol{\beta}(\tau)^{(i)}$ for $i = 1, \ldots, m_0$. We introduce some mild regularity conditions.

Assumption 1 (τ)(i) The distribution function $F(\cdot)$ admits a continuous density function $f(\cdot)$ that is bounded away from 0 and ∞ at $Q_\tau(\varepsilon)$; (ii) $F(s + Q_\tau(\varepsilon)) - F(Q_\tau(\varepsilon)) = sf(Q_\tau(\varepsilon)) + O(s^2)$ as $s \to 0$.

Assumption 2 (τ)There exists $s_0 > 0$ such that, (i)

$$\sup_{|s_1|, |s_2| \leq s_0, s_1 \neq s_2} (\phi_{\tau;1}(s_1; \mathcal{F}_t) - \phi_{\tau;1}(s_2; \mathcal{F}_t))/|s_1 - s_2| \in \mathcal{L}^1,$$

(ii) for $l = 0, 1$, we have $\sup_{|s|\leq s_0} \|\phi_{\tau,1}^{(l)}(s; \mathcal{F}_t)\|_4 < \infty$ and

$$\sum_{t=0}^{\infty} \sup_{|s| \leq s_0} \left\| E[\phi_{\tau;1}^{(l)}(s; \mathcal{F}_t) | \mathcal{F}_0] - E[\phi_{\tau;1}^{(l)}(s; \mathcal{F}_t') | \mathcal{F}_0'] \right\|_4 < \infty.$$

Assumption 3 (i) $\sum_{k=1}^{\infty} k^3 \delta_G'(k, 4) < \infty$; (ii) (GOALS) $\Gamma_\varepsilon(\tau) \in (0, \infty)$; (ii)′ (M-GOALS) Γ^M is positive definite.

Assumption 4 (i) $\min_{1\leq i\leq m_0+1} (q_i - q_{i-1}) > \epsilon$; (ii) (GOALS) we have $\mathbf{b}_i(\tau) = c_i(\tau)\kappa$ where $c_i(\tau) \in \mathbb{R}^2/\{(0, 0)^\top\}$ for $i = 1, \ldots, m_0$; (ii)′ (M-GOALS) let $\mathbf{b}_{1,i}(\tau) \in \mathbb{R}$ be the second element

of $\mathbf{b}_i(\tau)$ (i.e. change in slope parameters), we have $\mathbf{b}_{1,i}(\tau) = c_{1,i}(\tau)\kappa$ for each $\tau \in \boldsymbol{\tau}^M$, where $c_{1,i}(\tau) \in \mathbb{R}$ and for each $i = 1, \ldots, m_0$, there exists at least one $\tau \in \boldsymbol{\tau}^M$ such that $c_{1,i}(\tau) \neq 0$. Here, $\kappa = \kappa(n) > 0$ is a scalar dependent on n that measures the magnitude of changes.

Assumption 1 is a standard regularity condition in the quantile regression literature. Assumptions 2–3 regulate the error process $\{\varepsilon_t\}$ and guarantee the uniform Bahadur representation of the subsample estimates $\{\widehat{\boldsymbol{\beta}}_{\lfloor r_1 n \rfloor, \lfloor r_2 n \rfloor}(\tau), 0 < r_1 < r_2 < 1, \ |r_1 - r_2| \geq \eta\}$ for any $\eta > 0$. Assumptions 2–3 are adapted from Zhou and Shao (2013), where the authors extend the SN methodology in Shao (2010) from stationary time series to regression models with deterministic covariates and weakly dependent stationary errors. In comparison, our framework allows the error process to exhibit both heteroscedasticity and temporal dependence. Moreover, our theoretical argument hinges on the asymptotic analysis of subsample estimates $\widehat{\boldsymbol{\beta}}_{t_1,t_2}(\tau)$ allowing for the presence of change-points in the interval $[t_1, t_2]$, and thus is substantially different from that in Zhou and Shao (2013).

Assumption 4(i) requires the minimum spacing of the change-points to be larger than the global trimming parameter ε. Assumption 4(ii) and (ii)′ require the magnitudes of the changes are of the same order κ. See similar assumptions in Oka and Qu (2011) and Aue et al. (2014).

3.2 | Consistency of GOALS and M-GOALS

We first introduce some notations before presenting the consistency result. Denote $X(s) = (1, s)^\top$ and denote $B_X(r) = \int_0^r X(s)\, dB(s)$ where $B(\cdot)$ is a standard Brownian motion. For $i = 1, \ldots, m_0 + 1$, define $\Sigma_i(r) = \int_0^r [X(s)^\top \boldsymbol{\gamma}^{(i)}]^{-1} X(s) X(s)^\top \, ds$, $\Lambda_i(r_1, r_2) = [\Sigma_i(r_2) - \Sigma_i(r_1)]^{-1}[B_X(r_2) - B_X(r_1)]$, and define $\Lambda_i^M(r_1, r_2) = \int_{r_1}^{r_2} \{e_2^\top [\Sigma_i(r_2) - \Sigma_i(r_1)]^{-1} X(s)\}\, dB^M(s)$ where $e_2 = (0, 1)^\top$ and $B^M(\cdot)$ is an M-dimensional Brownian motion with independent entries.

For $u \in (\varepsilon, 1 - \varepsilon)$, define the scaled limit of $G_n(k)$ by

$$G_\epsilon(u) = \left\{(u_1, u_2) | u_1 = u - i\epsilon, \ i = 1, \ldots, \lfloor u/\epsilon \rfloor, \ u_2 = u + j\epsilon, \ j = 1, \ldots, \lfloor (1-u)/\epsilon \rfloor \right\}. \quad (6)$$

Furthermore, for $u \in (\varepsilon, 1 - \varepsilon)$ and $(u_1, u_2) \in G_\epsilon(u)$, define

$$\begin{aligned} D(u_1, u, u_2) &= \frac{(u-u_1)(u_2-u)}{(u_2-u_1)^{3/2}} \{\Lambda_1(u_1, u) - \Lambda_1(u, u_2)\}, \\ V_\delta(u_1, u, u_2) &= \int_{u_1+\delta}^{u-\delta} \frac{(r-u_1)^2(u-r)^2}{(u_2-u_1)^2(u-u_1)^2} \{\Lambda_1(u_1, r) - \Lambda_1(r, u)\}^{\otimes 2} \, dr \\ &\quad + \int_{u+\delta}^{u_2-\delta} \frac{(r-u)^2(u_2-r)^2}{(u_2-u_1)^2(u_2-u)^2} \{\Lambda_1(r, u_2) - \Lambda_1(u, r)\}^{\otimes 2} \, dr, \end{aligned}$$

and define

$$\begin{aligned} D^M(u_1, u, u_2) &= \frac{(u-u_1)(u_2-u)}{(u_2-u_1)^{3/2}} \{\Lambda_1^M(u_1, u) - \Lambda_1^M(u, u_2)\}, \\ V_\delta^M(u_1, u, u_2) &= \int_{u_1+\delta}^{u-\delta} \frac{(r-u_1)^2(u-r)^2}{(u_2-u_1)^2(u-u_1)^2} \{\Lambda_1^M(u_1, r) - \Lambda_1^M(r, u)\}^{\otimes 2} \, dr \\ &\quad + \int_{u+\delta}^{u_2-\delta} \frac{(r-u)^2(u_2-r)^2}{(u_2-u_1)^2(u_2-u)^2} \{\Lambda_1^M(r, u_2) - \Lambda_1^M(u, r)\}^{\otimes 2} \, dr. \end{aligned}$$

Here, $D(u_1, u, u_2)$, $V_\delta(u_1, u, u_2)$ and $D^M(u_1, u, u_2)$, $V_\delta^M(u_1, u, u_2)$ are used to quantify the asymptotic limit of the contrast statistic and the self-normalizer under the no change-point scenario with $m_0 = 0$.

Theorem 1 *(Consistency)Suppose Assumptions 1, 2, 3(i)–(ii) hold at τ for GOALS and Assumptions 1, 2, 3(i)–(ii)′ hold at all $\tau \in \boldsymbol{\tau}^M$ for M-GOALS. Furthermore, suppose* $\sup_{1\le t\le n} \|\boldsymbol{\gamma}_t\| < C_0 < \infty$.

Case 1 [No change-point scenario, $m_0 = 0$]: For GOALS, we have

$$\max_{k=1,\ldots,n} T_{n,\epsilon,\delta}(k) \to_D \mathcal{T}(\epsilon,\delta) = \sup_{u\in(\epsilon,1-\epsilon)} \max_{(u_1,u_2)\in G_\epsilon(u)} D(u_1,u,u_2)^\top V_\delta(u_1,u,u_2)^{-1} D(u_1,u,u_2).$$

For M-GOALS, we have

$$\max_{k=1,\ldots,n} T^M_{n,\epsilon,\delta}(k) \underset{D}{\to} \mathcal{T}^M(\epsilon,\delta) = \sup_{u\in(\epsilon,1-\epsilon)} \max_{(u_1,u_2)\in G_\epsilon(u)} D^M(u_1,u,u_2)^\top V^M_\delta(u_1,u,u_2)^{-1} D^M(u_1,u,u_2).$$

In particular, for any ζ_n, $\zeta_n^M \to \infty$, we have for both GOALS and M-GOALS,

$$\lim_{n\to\infty} P(\widehat{m} = 0) = 1.$$

Case 2 [Change-point scenario, $\mathbf{q} = (q_1, \ldots, q_{m_0})$, $m_0 \ge 1$]: Assume $\kappa \to 0$ and $\log(n)^{-2}(n\kappa^2) \to \infty$ as $n \to \infty$. Suppose additionally Assumption 4(i)–(ii) hold for GOALS and Assumption 4(i)–(ii)′ hold for M-GOALS. For any ζ_n, $\zeta_n^M = (n\kappa^2)^\iota$ with $\iota \in (0, 1)$, we have for both GOALS and M-GOALS,

$$\lim_{n\to\infty} P(\widehat{m} = m_0 \text{ and } \min_{1\le i\le m_0} |q_i - \widehat{q}_i| < \eta) = 1, \quad \text{for any } \eta > 0,$$

where $\widehat{\mathbf{q}} = (\widehat{q}_1, \ldots, \widehat{q}_{\widehat{m}}) = (\widehat{k}_1, \ldots, \widehat{k}_{\widehat{m}})/n$ denotes the estimated change-points.

Theorem 1 establishes the consistency result for GOALS and M-GOALS. In particular, Theorem 1 indicates that under the no change-point scenario, the global SN test statistic $\max_{1\le k\le n} T_{n,\epsilon,\delta}(k)$ of GOALS and $\max_{1\le k\le n} T^M_{n,\epsilon,\delta}(k)$ of M-GOALS converge to non-degenerate limiting distributions $\mathcal{T}(\epsilon, \delta)$ and $\mathcal{T}^M(\epsilon, \delta)$, respectively. Due to the use of SN, $\mathcal{T}(\epsilon, \delta)$ and $\mathcal{T}^M(\epsilon, \delta)$ do not depend on the temporal dependence and the density function of the error process $\{\varepsilon_t\}$. On the other hand, SN cannot fully remove the heteroscedasticity effect caused by $\boldsymbol{\gamma}$ due to the presence of the deterministic trend, and thus in general $\mathcal{T}(\epsilon, \delta)$ and $\mathcal{T}^M(\epsilon, \delta)$ are not pivotal. However, elementary algebra shows that for $\boldsymbol{\gamma}_1^{(1)} = 0$ (i.e. the error is homogeneous), $\mathcal{T}(\epsilon, \delta)$ and $\mathcal{T}^M(\epsilon, \delta)$ only depend on (ε, δ) and are indeed pivotal, thus providing a viable solution to the choice of ζ_n and ζ_n^M in practice.

Specifically, given (ε, δ) and a significance level α, we set ζ_n of GOALS to be the $(1 - \alpha) \times 100\%$ quantile of the pivotal $\mathcal{T}(\epsilon, \delta)$ and set ζ_n^M of M-GOALS to be the $(1 - \alpha) \times 100\%$ quantile of the pivotal $\mathcal{T}^M(\epsilon, \delta)$. Thus the threshold ζ_n and ζ_n^M is adaptive to the choices of (ε, δ) and M.

Throughout the paper, we set $\alpha = 0.1$ and $(\varepsilon, \delta) = (0.1, 0.02)$, which yields ζ_n=65.41 for GOALS and $\zeta_n^M = 49.89$ for M-GOALS with $M = 3$. Simulation studies and real data analysis indicate the satisfactory performance of ζ_n and ζ_n^M in practice. We refer to Section S5 of the supplement for a sensitivity analysis w.r.t. the significance level α and trimming parameters (ε, δ).

4 | REAL DATA ANALYSIS

In this section, we analyse the COVID-19 pandemic via the piecewise linear quantile trend model and the proposed M-GOALS algorithm. Section 4.1 provides in-sample analysis of the coronavirus infection curves in 35 major countries. Section 4.2 further designs an accurate short-term forecasting scheme based on M-GOALS and demonstrates its promising utility for public health decision-making.

4.1 | In-sample analysis of daily new cases

We focus our analysis on 35 major countries, which are the union of G20 nations and top 30 countries leading the total coronavirus cases as of Nov-07, 2020.[1] For each country, denote the observed trajectory of daily new cases as $\{(Y_t, d_t)\}_{t=1}^n$, where $Y_t = \log(R_t + 1)$ is the logarithm of new cases recorded on date d_t. We set d_1 as the date when the total cases of the country exceeded 1000 and set d_n as Nov-07. The average starting date d_1 is mid-late March and the average observation length n is 224 days across the 35 countries. Note that $\varepsilon = 0.1$ indicates that the spread pattern of the coronavirus remains stable for at least 22 days (around 3 weeks), which is a reasonable assumption. More information, including country names (and its abbreviation) and data length, can be found in Table S6 of the supplement.

For each country, we analyse the trajectory of its (log-scale) daily new cases $\{Y_t\}_{t=1}^n$ via the piecewise linear quantile trend model (1) estimated by M-GOALS with $\boldsymbol{\tau}^M = (0.1, 0.5, 0.9)$ and $(\varepsilon, \delta) = (0.1, 0.02)$. For the ith estimated segment, based on its linear trend parameter $(\widehat{\beta}_0(\tau)^{(i)}, \widehat{\beta}_1(\tau)^{(i)})$, we define $S^{(i)}(\tau) = \widehat{\beta}_1(\tau)^{(i)}/n$ as its normalized slope at $\tau \in \boldsymbol{\tau}^M$. Note that $S^{(i)}(\tau)$ measures $Q_\tau(Y_{t+1}) - Q_\tau(Y_t) = \log[Q_\tau(R_{t+1} + 1)/Q_\tau(R_t + 1)]$ and therefore can be nicely interpreted as the (approximate) growth rate of the daily new cases R_t at the τth quantile. Thus, the piecewise linear quantile trend model (1) allows us to assess the growth rate of the coronavirus at any given time, which helps better understand the trajectory of the pandemic and further facilitates short-term forecast across all quantile levels in $\boldsymbol{\tau}^M$.

Case studies for United States and United Kingdom: We first present a detailed case study for United States (d_1 = Mar-11, $n = 242$) and United Kingdom (d_1 = Mar-13, $n = 240$). Detailed results for seven other representative countries (India, Brazil, France, Russia, Spain, South Africa and Australia) can be found in the supplement.

Figure 1 visualizes the estimated piecewise linear quantile trend models, where we plot the recovered change-points and the normalized slope (i.e. growth rate) of each segment. One notable feature is the multiple peaks exhibited by the two trajectories, which is further manifested by the alternating signs of the estimated growth rate. Specifically, the pandemic exhibited multiple waves in both United States and United Kingdom and for both countries, the recent peak is more

[1]We obtain the data from https://ourworldindata.org/coronavirus-source-data maintained by 'Our World in Data'.

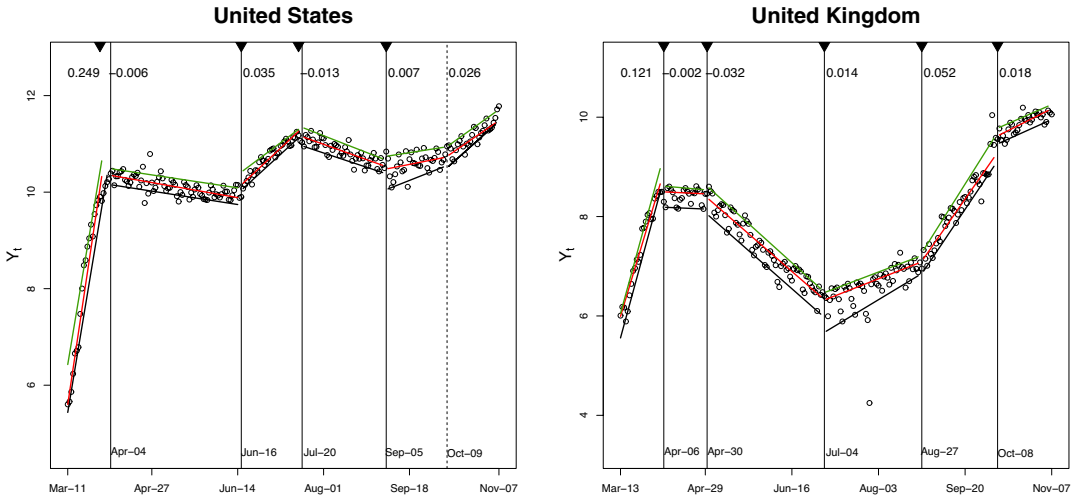

FIGURE 1 Estimated piecewise linear quantile trend models by M-GOALS. Vertical lines (solid) mark the estimated change-points (with exact dates given by the texts). The vertical dashed line marks the h-local maximizer below the threshold ζ_n^M. The broken lines mark the estimated piecewise linear quantile trend models at $\tau = 0.1$ (black), $\tau = 0.5$ (red) and $\tau = 0.9$ (green). Numbers in the upper part of the plots give the estimated normalized slopes (i.e. growth rate) for each segment at $\tau = 0.5$. Black triangles mark the estimated change-points by multi-scanning M-GOALS [Colour figure can be viewed at wileyonlinelibrary.com]

severe than previous ones. The first change-points of the two countries are close, suggesting the initial public health interventions taken by United States and United Kingdom have similar effectiveness. However, the new cases in United States remain at a high-level after the first peak while United Kingdom is more effective at flattening the first wave of the pandemic. The most recent wave in United States and United Kingdom both initiated around the beginning of fall, which is consistent with the timing of reopening policies in the two countries. While the growth rate of new cases in United States is accelerating, United Kingdom seems to be levelling off and approaching the peak of its second wave. Figure 1 also plots the estimated change-points by multi-scanning M-GOALS (see the supplement for its detailed implementation), which almost perfectly match the estimation by M-GOALS for both countries, indicating the robustness of our finding.

Figure S4 of the supplement gives the estimation result by the L_1 quantile trend filtering (TF) in Brantley et al. (2020), where the result seems less intuitive and notably worse than M-GOALS. Specifically, the quantile TF seems to miss several evident change-points, and although the algorithm guarantees quantile non-crossing, due to its lack of power for detecting change-points, the estimated quantiles across $\tau \in \boldsymbol{\tau}^M$ coincide on a large portion of the data. This suggests that a change-point-based approach may be more suitable for the modelling of COVID-19 infection curves compared to the non-parametric TF. See Fearnhead et al. (2019) for comparisons between TF and the change-point approach in linear trend models for change in mean.

Clustering trajectories via growth rates: To gather a relatively complete assessment of the pandemic trajectories around the world, we further conduct a clustering analysis based on the recovered normalized slopes of the 35 countries. Specifically, for a country A, based on the estimation by M-GOALS, we can recover its growth rate path $G_A = \{\widehat{\beta}_t(\tau)/n,\, d_t\}_{t=1}^n$ at any quantile level $\tau \in \boldsymbol{\tau}^M$. Denote $G_A^\beta = \{\widehat{\beta}_t(\tau)/n\}_{t=1}^n$ and $G_A^d = \{d_t\}_{t=1}^n$. For two countries (A, B), we then measure the (dis)-similarity of their coronavirus trajectories by $d(A, B) = 1 - \rho(A,B)$, where

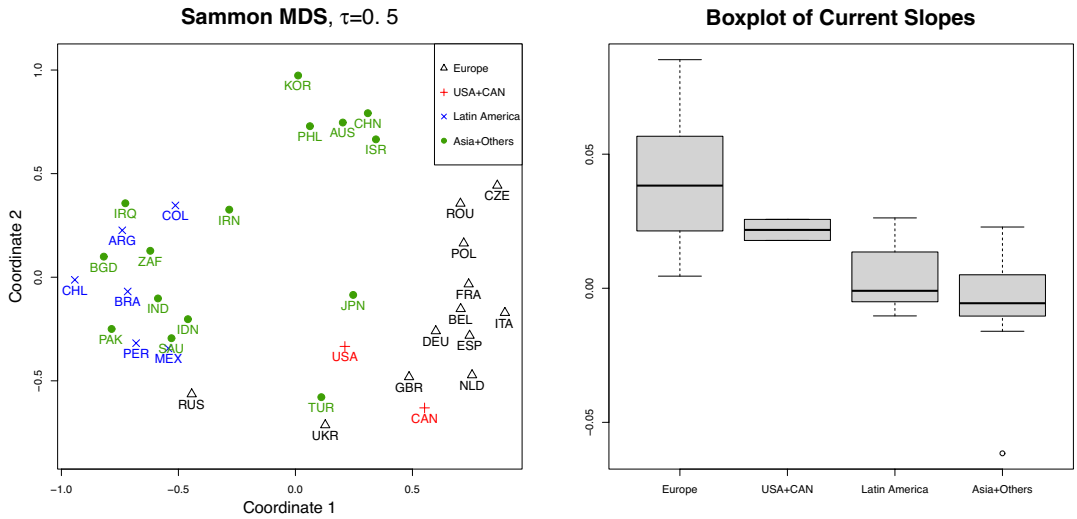

FIGURE 2 Left panel: Visualization of the dissimilarity matrix D for $\tau = 0.5$ via the Sammon MDS. Right panel: Boxplot of the current growth rate at $\tau = 0.5$ across four groups of countries [Colour figure can be viewed at wileyonlinelibrary.com]

we define $\rho(A, B)$ as the sample correlation of the two growth rate paths G_A^β and G_B^β calculated on common dates $G_A^d \cap G_B^d$. Thus, a dissimilarity matrix D of the 35 countries can be readily attained. The left panel of Figure 2 visualizes D for $\tau = 0.5$ by multidimensional scaling (MDS) in Sammon (1969). One notable feature of the plot is the closeness among continental European countries, which is not surprising considering the geographical proximity and economic ties among European Union. On the other hand, developing countries in Asia and Latin America seem to cluster together and exhibit similar growth patterns. Another distinct group is China, South Korea, Australia, Israel and Philippines, which may be regarded as countries that control the pandemic relatively well (among the selected 35 nations). The clustering result is consistent and robust across the quantile level τ, as indicated by the MDS plots for $\tau = 0.1, 0.9$ in the supplement. The right panel of Figure 2 gives the boxplot of the normalized slope $\widehat{S^{m+1}}(\tau)$ in the last segment (i.e. the current growth rate) at $\tau = 0.5$ across four groups of countries. It can be seen that European and North American countries have notably higher growth rates, indicating the severe situations due to the second wave. Figure S6 of the supplement further plots the clustering results based on multi-scanning M-GOALS, which closely matches patterns in Figure 2 and confirms the robustness of the analysis.

4.2 | Short-term forecast of daily new cases

As stated by the Centers for Disease Control and Prevention (CDC),[2] accurate forecast of new cases is critical for public health decision-making, as it projects the likely impact of coronavirus to health systems in coming weeks and thus provides invaluable information for developing data-driven public health policies to contain the pandemic. In this section, we propose a simple

[2]https://www.cdc.gov/coronavirus/2019-ncov/covid-data/forecasting- us.html/why-forecasting-critical.

and intuitive M-GOALS based short-term forecasting scheme and demonstrate its effectiveness in predicting new cases of COVID-19.

As suggested by the analysis in Section 4.1, one notable characteristic of the coronavirus pandemic is its multiple epidemic phases evidenced by the non-stationary and alternating growth rates of the new cases, suggesting that any prediction model built on stationarity will inevitably lead to erroneous forecasts. Thus, a natural (and simple) solution from the change-point perspective is to first segment the time series into periods with relatively stable behavior and then generate forecast based on observations in the last segment, see for example Pesaran and Timmermann (2002) and Bauwens et al. (2015).

Method: Following this idea, we propose an M-GOALS based two-stage forecasting scheme for new cases prediction. Specifically, in the first stage, given the observed trajectory of daily new cases $\{Y_t\}_{t=1}^n$, a piecewise linear quantile trend model is estimated via M-GOALS with $\boldsymbol{\tau}^M$ to obtain the potential change-points. In the second stage, for each quantile level $\tau \in \boldsymbol{\tau}^M$, a flexible function $f_\tau(t)$ is fitted on the last segment $\{Y_t\}_{t=\hat{k}_{\hat{m}}+1}^n$ with the assumption $Q_\tau(Y_t) = f_\tau(t)$. The k-day ahead forecast for new cases at the τth quantile is thus defined as $\widehat{f}_\tau(n+k)$ via extrapolation. Importantly, the proposed procedure naturally generates robust prediction intervals thanks to the multiple quantile levels within $\boldsymbol{\tau}^M$.

For flexible out-of-sample extrapolation, in the second stage, we fit both a linear trend model $f_\tau(t) = \beta_0(\tau) + \beta_1(\tau)(t/n)$ and a quadratic trend model $f_\tau(t) = \beta_0(\tau) + \beta_1(\tau)(t/n) + \beta_2(\tau)(t/n)^2$ to the last segment, and generate forecast based on the model selected by quantile regression BIC (Lee et al., 2014). The quadratic model may potentially improve forecast accuracy (if selected by BIC) for the scenario where the trajectory undergoes a very recent change that may be detected by M-GOALS with delay.

Data and forecast results: We set $\boldsymbol{\tau}^M = (0.1, 0.5, 0.9)$ and backtest the M-GOALS based prediction method for generating short-term forecast of cumulative new cases in the United States and benchmark its performance with the CDC Ensemble, which is an ensemble model built on forecasts generated by around 70 modeling groups across the world.[3] Specifically, following the CDC website, the forecast is generated on every Monday starting from Aug-03 and the forecast target is 5-day (one-week) ahead and 12-day (two-week) ahead *cumulative* new cases in the United States.

Based on the estimated quantile regression $\widehat{f}_\tau(t)$, we forecast the τth quantile of the k-day ahead cumulative new cases via $C_\tau = \exp(\sum_{i=1}^k \widehat{f}_\tau(n+i))$. We use $C_{0.5}$ as a point forecast for the median of the cumulative new cases and $[C_{0.1}, C_{0.9}]$ naturally forms a 80% prediction interval.[4] For illustration, Figure S7 of the supplement visualizes the forecasting scheme at four representative dates.

Figure 3 visualizes the forecast results given by M-GOALS and CDC Ensemble from Aug-03 to Nov-09 (see tabulated results in Tables S7 and S8 of the supplement). In general, M-GOALS provides reasonable short-term forecasts with accuracy comparable to CDC Ensemble. Moreover, the coverage rate of the 80% prediction interval $(C_{0.1}, C_{0.9})$ seems more satisfactory than the 95% prediction interval given by CDC Ensemble. On the other hand, both models seem to suffer noticeable downward forecast bias (though less severe for M-GOALS) starting from mid-October, highlighting the unexpected severity of the third wave of the US coronavirus pandemic. Figure S8 of the supplement further plots the forecast results based on multi-scanning M-GOALS, which

[3]See more details of CDC Ensemble at https://github.com/reichlab/covid19-forecast-hub#ensemble-model.

[4]In general sum of quantiles differs from quantile of sum. However, Proposition 11 in the supplement shows that the prediction interval given by $[C_{0.1}, C_{0.9}]$ remains valid albeit conservative if the daily new cases follow an elliptical distribution.

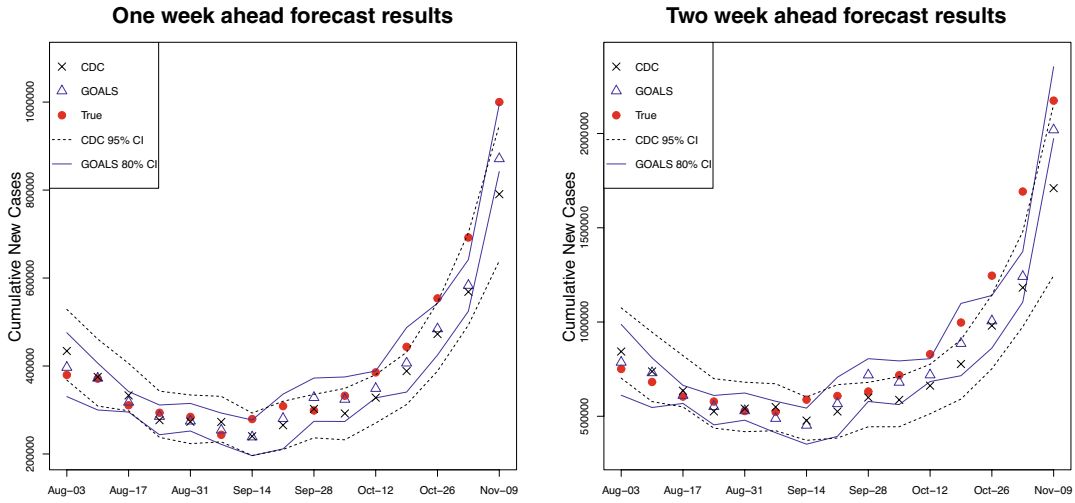

FIGURE 3 Forecast results for one-week and two-week ahead cumulative new cases from Aug-03 to Nov-09 [Colour figure can be viewed at wileyonlinelibrary.com]

are almost identical as the ones reported in Figure 3 and can be seen as further support for the robustness of our forecast analysis.

5 | CONCLUSION

In this paper, we propose a piecewise linear quantile trend model to study the COVID-19 infection curves simultaneously across multiple quantiles. A novel segmentation algorithm, M-GOALS, is proposed for multiple change-point estimation, which integrates SN-based change-point tests conducted at global scales with a local-scanning procedure. The consistency of M-GOALS is established under a location-scale model that incorporates heteroscedasticity and temporal dependence. Simulation studies and real data applications confirm the favourable performance of our method and further demonstrate its promising ability to provide crucial information for public health decision-making to combat the COVID-19 pandemic.

Same as most works in statistical modelling, our model and analysis come with assumptions and thus limitations. Our analysis relies on the time series of COVID-19 daily new cases R_t reported by each country. The value R_t is likely an underestimate of the actual daily new cases R_t^{true}, especially at the very early stage of the pandemic, due to reasons such as lack of testing capacity, delays in diagnosis/notification, and viral latency. The degree of such underestimation may also vary from country to country. We note, however, if R_t can reflect an approximately constant fraction $\theta \in (0, 1]$ of R_t^{true} with $R_t \approx \theta R_t^{true}$, our segmentation result and infection growth rate estimation (via normalized slopes of the estimated quantile regression) will stay valid for R_t^{true}, as our analysis is performed at the log scale of R_t. On the other hand, our current model is unable to account for the case where the fraction θ is significantly time-varying, and our analysis result obtained based on R_t needs to be interpreted with caution when generalized to the actual cases R_t^{true}.

In our current work, the proposed M-GOALS conducts change-point detection and case forecasting for the spread of COVID-19 pandemic solely based on information contained in $\{R_t\}_{t=1}^n$. One

potential extension is to further incorporate other available information in the analysis, such as strictness of social distancing rules, testing positivity rates, and death cases, which could be useful from an epidemiology perspective. In addition, our current analysis treats the infection trajectories from different countries separately and thereby does not exploit the potential dependency/similarity of change-point locations and growth rates among different nations, such as neighbouring countries in Europe. One natural and promising research direction is to further extend the piecewise linear quantile trend model and M-GOALS to a panel data setting, where multiple time series share similarity in change-point locations and parameters of the quantile regression. Another research direction of theoretical interest is to establish rigorous theoretical guarantees for the multi-scanning M-GOALS, which is seen to offer promising numerical performance.

We note that our model is purely statistical in that it directly models the observed time series of new cases, instead of modelling the dynamics of the coronavirus transmission based on mechanistic models such as SIR and its variants, see Anastassopoulou et al. (2020), Bai et al. (2020), Chen et al. (2020), Lin et al. (2020) and Wu et al. (2020) among others. As evidenced by the meaningful in-sample analysis and accurate out-of-sample forecast, we believe our model can serve as a good complement to the large literature of COVID-19 modelling via mechanistic models based on epidemiology principles.

Though primarily motivated by the study for COVID-19 infection curves, the proposed piecewise linear quantile trend model and the segmentation method M-GOALS can be useful in other applications as well, such as detecting structural breaks of Value-at-Risk for financial and macroeconomics time series, where outliers and heteroscedasticity are often encountered.

ACKNOWLEDGEMENTS

We thank Professor Zhongjun Qu for providing the code used in Oka and Qu (2011). We are also grateful to the reviewers for helpful comments, which led to substantial improvements. Jiang acknowledges that part of the work was carried out during a visit to University of Illinois at Urbana-Champaign and at Tsinghua University.

ORCID

Feiyu Jiang http://orcid.org/0000-0002-6243-6200
Zifeng Zhao http://orcid.org/0000-0001-8466-1321
Xiaofeng Shao http://orcid.org/0000-0001-5822-8209

REFERENCES

Anastassopoulou, C., Russo, L., Tsakris, A. & Siettos, C. (2020) Data-based analysis, modelling and forecasting of the COVID-19 outbreak. *PLoS One*, 15, e0230405.

Andrews, D.W. (1993) Tests for parameter instability and structural change with unknown change point. *Econometrica*, 61, 821–856.

Aue, A. and Horváth, L. (2013) Structural breaks in time series. *Journal of Time Series Analysis*, 34, 1–16.

Aue, A., Cheung, R.C.Y., Lee, T.C. & Zhong, M. (2014) Segmented model selection in quantile regression using the minimum description length principle. *Journal of the American Statistical Association*, 109, 1241–1256.

Aue, A., Cheung, R.C.Y., Lee, T.C. & Zhong, M. (2017) Piecewise quantile autoregressive modeling for non-stationary time series. *Bernoulli*, 23, 1–22.

Bai, J. & Perron, P. (2003) Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, 18, 1–22.

Bai, Y., Safikhani, A. & Michailidis, G. (2020) Non-stationary spatio-temporal modeling of COVID-19 progression in the U.S. *medRxiv*, available at https://doi.org/10.1101/2020.09.14.20194548.

Baranowski, R., Chen, Y. & Fryzlewicz, P. (2019) Narrowest-over-threshold detection of multiple change points and changepoint-like features. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81, 649–672.

Bauwens, L., Koop, G., Korobilis, D. & Rombouts, J.V. (2015) The contribution of structural break models to forecasting macroeconomic series. *Journal of Applied Econometrics*, 30, 596–620.

Beran, R. (1987) Prepivoting to reduce level error of confidence sets. *Biometrika*, 74, 457–468.

Bondell, H., Reich, B. & Wang, H. (2010) Noncrossing quantile regression curve estimation. *Biometrika*, 97, 825–838.

Brantley, H.L., Guinness, J. & Chi, E.C. (2020) Baseline drift estimation for air quality data using quantile trend filtering. *Annals of Applied Statistics*, 14, 585–604.

Chen, Y.-C., Lu, P.-E. & Chang, C.-S. (2020) A time-dependent SIR model for COVID-19. Available at: *Arxiv*, https://arxiv.org/pdf/2003.00122.pdf.

Dong, C., Gao, J., Linton, O. & Peng, B. (2020) On time trend of COVID-19: a panel data study. Available at: http://www.econ.cam.ac.uk/research-files/repec/cam/pdf/cwpe2065.pdf.

Fearnhead, P.,Maidstone, R. & Letchford, A. (2019) Detecting changes in slope with an L_0 penalty. *Journal of Computational and Graphical Statistics*, 28, 265–275.

Gu, J., Yan, H., Huang, Y., Zhu, Y., Sun, H., Zhang, X. et al. (2020) Better strategies for containing COVID-19 epidemics – a study of 25 countries via an extended varying coefficient SEIR model. *medRxiv*, available at: https://doi.org/10.1101/2020.04.27.20081232.

Hao, N., Niu, S.Y. & Zhang, H. (2013) Multiple change-point detection via a screening and ranking algorithm. *Statistica Sinica*, 23, 1553–1572.

Harvey, A. & Kattuman, P. (2020) Time series models based on growth curves with applications to forecasting coronavirus. *Harvard Data Science Review*, to appear.

Jiang, F., Zhao, Z. & Shao, X. (2020) Time series analysis of COVID-19 infection curve: a change-point perspective. *Journal of Econometrics*. In press, available at: https://www.sciencedirect.com/science/article/pii/S0304407620302633.

Koenker, R. (2005) *Quantile regression*. Cambridge: Cambridge University Press.

Lee, E.R., Noh, H. & Park, B.U. (2014) Model selection via Bayesian information criterion for quantile regression models. *Journal of the American Statistical Association*, 109, 216–229.

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

Lin, Q.-S., Hu, T.-J. & Zhou, X.-H. (2020) Estimating the daily trend in the size of the COVID-19 infected population in Wuhan. *Infectious Diseases of Poverty*, 9, 1–8.

Liu, L., Moon, H. R. and Schorfheide, F. (2021) Panel forecasts of country-level COVID-19 infections. *Journal of Econometrics*, 220, 2–22.

Niu, S.Y. & Zhang, H. (2012) The screening and ranking algorithm to detect DNA copy number variations. *Annals of Applied Statistics*, 6, 1306–1326.

Oka, T. & Qu, Z. (2011) Estimating structural changes in regression quantiles. *Journal of Econometrics*, 162, 248–267.

Perron, P. (2006) Dealing with structural breaks. *Palgrave Handbook of Econometrics*, 1, 278–352.

Pesaran, M.H. & Timmermann, A. (2002) Market timing and return prediction under model instability. *Journal of Empirical Finance*, 9, 495–510.

Rho, Y. & Shao, X. (2015) Inference for time series regression models with weakly dependent and heteroscedastic errors. *Journal of Business & Economic Statistics*, 33, 444–457.

Sammon, J.W. (1969) A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, 100, 401–409.

Scott, A.J. & Knott, M. (1974) A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, 30, 507–512.

Shao, X. (2010) A self-normalized approach to confidence interval construction in time series. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72, 343–366.

Shao, X. (2015) Self-normalization for time series: a review of recent developments. *Journal of the American Statistical Association*, 110, 1797–1817.

Shao, X. & Zhang, X. (2010) Testing for change points in time series. *Journal of the American Statistical Association*, 105, 1228–1240.

Wu, W.B. (2005) Nonlinear system theory: another look at dependence. *Proceedings of the National Academy of Sciences of the United States of America*, 102(40), 14150–14154.
Wu, W.-C. & Zhou, Z. (2018) Gradient-based structural change detection for nonstationary time series M-estimation. *Annals of Statistics*, 46, 1197–1224.
Wu, J.T., Leung, K. & Leung, G.M. (2020) Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. *The Lancet*, 395, 689–697.
Yau, C.Y. & Zhao, Z. (2016) Inference for multiple change points in time series via likelihood ratio scan statistics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78, 895–916.
Zhang, T. & Lavitas, L. (2018) Unsupervised self-normalized change-point testing for time series. *Journal of the American Statistical Association*, 113, 637–648.
Zhou, Z. & Shao, X. (2013) Inference for linear models with dependent errors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75, 323–343.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Jiang, F., Zhao, Z. & Shao, X. (2022) Modelling the COVID-19 infection trajectory: A piecewise linear quantile trend model. *J R Stat Soc Series B*, 84, 1589–1607. https://doi.org/10.1111/rssb.12453