Statistica Sinica **31** (2021), 361-390
doi:https://doi.org/10.5705/ss.202018.0375

A THRESHOLD AUTOREGRESSIVE MODEL FOR ANALYZING THE INFLUENCE OF MEDIA REPORTS OF SUICIDE ON THE ACTUAL SUICIDES

Huazhen Lin[1], Jiakun Jiang[1,2], Binhuan Wang[3] and Paul S. F. Yip[4]

[1]*Southwestern University of Finance and Economics,* [2]*Tsinghua University,*
[3]*New York University School of Medicine and* [4]*University of Hong Kong*

Abstract: The extensive coverage of suicides in the media has long been considered a trigger for copycat suicides. However, evidence of such an effect is indirect and, thus, inconclusive. Here, we propose a flexible threshold autoregressive model to examine whether suicides reported in the media influence actual suicides and, thus, identify a possible copycat effect. In particular, we employ a penalized smoothing least squares estimator to conveniently estimate the parameters and unknown functions of the proposed model. We evaluate the performance of the proposed method using simulation studies, and examine the asymptotic behavior of the corresponding estimators under mild regularity conditions. Lastly, we apply our model to investigate the relationship between the daily suicide incidence and the number of suicides reported in a top-selling tabloid newspaper in Hong Kong between January 2002 and December 2006. Our results identify a copycat suicide effect due to excessive media reporting, as well as the threshold number of reports that may trigger such an effect.

Key words and phrases: Autoregressive model, copycat suicide effect, penalized smoothing least squares estimator, threshold model, time series data.

1. Introduction

Numerous studies have investigated whether media coverage of suicides results in copycat suicides (Phillips (1974); Pirkis and Blood (2001); Chen, Chen and Yip (2011); Niederkrotenthaler et al. (2012); Niederkrotenthaler and Stack (2017)). Here, suicides of celebrities have been found to have a significant social effect (Yip et al. (2006); Fu and Yip (2007); Chen et al. (2013)). However, evidence of such effects is indirect, inconclusive, and not specific. In addition, no studies have attempted to identify number of reports that triggers a copycat

Corresponding author: Binhuan Wang, 180 Madison Avenue, Suite 450, New York, NY 10016, USA.
E-mail: Binhuan.wang@nyulangone.org.

effect.

To examine whether the number of suicides reported in the media is related to actual suicide incidence, we record the coverage of suicides in a popular Hong Kong-based tabloid newspaper, the Apple Daily (AD). The AD is the mostly widely circulated community newspaper in Hong Kong, with a readership of more than two million (the total population of Hong Kong is about seven million) and, thus, a high household penetration rate. Furthermore, the AD is well known for its reports on celebrities, gossip, and scandals. Capturing a wide readership using sensationalism, exaggerated headlines, and attention-grabbing graphic images, the AD quickly became a top-selling newspaper in Hong Kong soon after its first issue in 1995. Unfortunately, such exaggerated media reporting has serious, adverse implications for the media industry (Chen et al. (2013)).

The Hong Kong suicide rate was particularly high between 2002 and 2006, especially in 2003. The SARS epidemic between March and May of 2003 seriously affected Hong Kong, resulting in a record unemployment rate of 8.6 and a historically highest suicide rate. Furthermore, after the death of a celebrity, Mr. Leslie Cheung, on April 1, 2003, the suicide rate surged by more than 20% in the subsequent four to six weeks, and remained at a high level until the end of that year (Yip et al. (2006)). The media coverage of Mr. Cheung's death was extensive and sensational. In addition, the number of charcoal-burning suicides rose to a historical level, since the first recorded case in 1997, of about 320 people in 2003, contributing significantly to the increase in the suicide rate during that period (Law et al. (2014)). The spread of charcoal-burning suicides has also been linked to media reports and Google searches in Taiwan (Chang et al. (2015)). The World Health Organization, International Association of Suicide Prevention, and many other organizations and press associations have issued guidelines on how to report suicide incidences (World Health Organization (2014)).

Based on a Poisson time series autoregression model, Chen, Chong and Bai (2012) examined whether the widespread media reporting of the suicide of a young female singer by charcoal burning increased suicide rates in Taiwan. Their results confirmed that a detailed description of the method used by a celebrity to commit suicide may incur a strong copycat effect. Cheng et al. (2007) found a mutual causation between suicide reporting and suicide incidence; that is, an increase in the number of reported suicide triggers more actual suicides, and an increase in the number of actual suicides results in more reported suicides. Consequently, the effect of media coverage on actual suicides is multiplicative and interactive, and not linear. However, evidence of copycat suicide effects remains indirect and unclear (Cheng, Chang and Yip (2012); Chen, Chen and Yip (2011)).

Using online search tools, we collected daily data on the number of AD headlines containing key words (in Chinese) related to suicidal behavior (e.g., "suicide," "building jumping," "charcoal burning," or "hanging") for the period January 2002 to December 2006. Data on the daily numbers of suicides are obtained from the Coroner's Court, which is responsible for certifying any unnatural cause of death (including suicide). The main purpose of this study is to explore the relation between media coverage of suicides and suicide incidence.

Let Y_t and X_t denote the number of suicides and the number of reports in AD on day t ($t = 0, \ldots, n$), respectively. Two characteristics are incorporated into the model. The first is whether a copycat suicide effect exists. Researchers believe that the effect of Y_{t-j} on Y_t is amplified if suicides are reported extensively in the media. However, it is not clear how many reports are required to trigger this amplified effect of Y_{t-j} on Y_t. Second, the effect of previous media coverage X_{t-j} and previous suicides Y_{t-j} on Y_t may depend on the time gap j, such that the effects are stronger for more recent media coverage on suicides, and may diminish over time. If so, it is important to know when and how previous media coverage and previous suicides cease to have an effect. To address these issues, we propose the following model:

$$\begin{aligned} E\left\{Y_t | X_s, Y_s, s < t\right\} = \mu &+ \sum_{j=1}^{p} \alpha_1(j) Y_{t-j} + \sum_{j=1}^{q} \alpha_2(j) X_{t-j} I(X_{t-j} \geq c_1) \\ &+ \sum_{j=1}^{w} \alpha_3(j) X_{t-j} Y_{t-j} I(X_{t-j} \geq c_2), \end{aligned} \tag{1.1}$$

where $\alpha_1(j)$, $\alpha_2(j)$, and $\alpha_3(j)$ quantify the correlations between observations over time and, thus, reflect when and how previous media coverage and previous suicides cease to have an effect; c_1 and c_2 are unknown threshold parameters that relate to the occurrence of a copycat suicide effect; p, q, and w are the maximum time gaps for the second to the fourth term, respectively, of the right-hand side of (1.1) to be nonzero. The main goals of this model are to determine the threshold parameters c_k and to estimate the size of the effects $\alpha_2(j)$ and $\alpha_3(j)$, if a copycat effect occurs.

First, the proposed model (1.1) is a threshold model. Existing threshold models can be grouped into two broad categories. In the first, only one threshold variable is included in the model. This threshold variable can be an actual variable (Hanse (1999, 2000); Chan (1993); Caner and Hansen (2001, 2008);Qian (1998); Koop and Potter (1999); Delgadoa and Hidalgo (2000); Li and Ling (2012)) or a combination of multiple variables (Seo and Linton (2007); Chen and So (2006); Tsay (1998)). In the second category, multiple threshold variables are included in the model. Chen, Chong and Bai (2012) proposed a two-threshold variable autoregressive (TTV-AR) model and applied a grid search approach to estimate the threshold values. Ni, Xia and Liu (2018) proposed a Bayesian stochastic search variable selection method to study subsets selected using the TTV-AR model and to estimate the parameters simultaneously. Wu and Chen (2007) proposed a threshold-variable-driven switching AR model, where the threshold variable is a random latent (unobservable) indicator that depends on the covariates through link functions.

Statistical inferences on the threshold model in (1.1) cannot be solved using a traditional regression problem, because they include the unknown threshold parameters c_k. A common practice is to estimate the thresholds using a simple grid search method. Here, the threshold estimates are obtained from the point yielding the least squared error across an arbitrarily finite number of candidate points. The computation time required for a grid search on G grid points is $O(G^2)$, which is computationally costly for large G. The threshold parameters also make it difficult to derive the asymptotic distributions of the resulting estimators, because standard asymptotic methods require a smooth criterion function, which is not the case in our model. In this paper, we propose a smoothing technique to solve this problem. The computation with the proposed method is straightforward, and can be accomplished using a standard Newton–Raphson algorithm. Furthermore, this smoothing technique helps us to establish the asymptotic theory and construct a sandwich formula to estimate the variances of the estimators.

There is an additional issue related to the estimations of $\alpha_k(j)$, for $k = 1, 2, 3$. Because j takes a finite number of values, we can specify each $\alpha_k(j)$ as a separate parameter. We call this a simple parametric method. This simple approach may lose information because $\alpha_k(j)$, in general, varies slowly over j; that is, $\alpha_k(j)$ is smooth, in some sense. A popular method that incorporates this smoothness is the nonparametric smoothing technique. However, because the arguments of $\alpha_k(\cdot)$, for $k = 1, 2, 3$, are discrete and finite, the traditional nonparametric method does not fit. In this paper, we propose a penalized least squares method to incorporate the smoothness of $\alpha_k(\cdot)$ using discrete and finite arguments.

The paper proceeds as follows. In Section 2, we introduce the flexible threshold autoregressive (FTAR) method, and then establish its asymptotic properties in Section 3. Section 4 discusses the bandwidth and tuning parameter selection. Numerical simulations and analyses of Hong Kong suicide data using the FTAR procedure and other methods are provided in Sections 5 and 6, respectively. Section 7 concludes the paper. All technical proofs are deferred to the Appendix.

2. FTAR Estimation

For notational simplicity, we set $p = q = w$ by replacing p, q, and w in model (1.1) with the maximum of p, q, and w and setting some $\alpha_k(j)$ to zero. Let $\boldsymbol{V}_t = (V_{t1}, \ldots, V_{tp})' \equiv (Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p})'$ and $\boldsymbol{X}_t = (X_{t1}, \ldots, X_{tp})' \equiv (X_{t-1}, \ldots, X_{t-p})'$. Denote $\boldsymbol{\alpha}_k = (\alpha_k(1), \ldots, \alpha_k(p))'$, for $k = 1, 2, 3$, and $\mathbf{c} = (c_1, c_2)'$ and $\boldsymbol{\Theta} = (\mu, \boldsymbol{\alpha}_1', \boldsymbol{\alpha}_2', \boldsymbol{\alpha}_3', \mathbf{c}')'$. Here, $\boldsymbol{\Theta}$ represents all parameters defined in model (1.1).

First, we develop estimators for c_1 and c_2. The objective least squares function is not continuous with respect to c_1 or c_2. This discontinuity, which stems from the indicator functions $I(X_{tj} > c_k)$, makes deriving the asymptotic distributions of the estimators computationally difficult (Sherman (1993); Han (1987); Faraggi and Simon (1996)) Here, we solve the discontinuity problem using the kernel smoothing technique (Brown and Wang (2005); Lin, Yip and Huggins (2011)). Let Φ denote the standard normal distribution function. Note that if $X_{tj} > c_k$, $\Phi\left((X_{tj} - c_k)/h\right) \to 1$ as $h \to 0$, whereas if $X_{tj} < c_k$, $\Phi\left((X_{tj} - c_k)/h\right) \to 0$, where the bandwidth h goes to zero as the sample size increases; that is, $\Phi\left((X_{tj} - c_k)/h\right) \to I(X_{tj} > c_k)$. The inequality (B.3) in the proof in Appendix B shows that when h is sufficiently small, the error from the approximation is negligible. Rather than a normal approximation, other approximations for $I(X_{tj} > c_k)$, such as a sigmoid approximation (Ma and Huang (2007)), can also be used. These simplify the computation of $\boldsymbol{\Theta}$ (especially for $\mathbf{c}$), which can be accomplished using a standard Newton–Raphson iterative algorithm. Finally, to incorporate that $\alpha_k(j)$, for $k = 1, 2, 3$, vary slowly over j, we estimate $\boldsymbol{\Theta}$ by minimizing the following penalized least squares function:

$$L_n(\boldsymbol{\Theta}) = l_n(\boldsymbol{\Theta}) + \lambda J(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\alpha}_3) \tag{2.1}$$

with respect to $\boldsymbol{\Theta}$, where

$$l_n(\boldsymbol{\Theta}) = \frac{1}{n} \sum_{t=1}^{n} \Bigg[Y_t - \mu - \sum_{j=1}^{p} \alpha_1(j) V_{tj}.$$

$$-\sum_{k=1}^{2}\sum_{j=1}^{p}\alpha_{k+1}(j)X_{tj}V_{tj}^{k-1}\Phi\left(\frac{X_{tj}-c_k}{h}\right)\Bigg]^2, \tag{2.2}$$

λ is a tuning parameter, and $J(\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\boldsymbol{\alpha}_3)$ is a penalty function that enforces the smoothness on $\alpha_k(\cdot)$, for $k=1,2,3$. The choice of penalty function $J(\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\boldsymbol{\alpha}_3)$ is crucial. Note that $\alpha_k(j)$ varies slowly over j, and the argument of $\alpha_k(\cdot)$ is an ordinal variable. Therefore, we may assume that $\alpha_k(\cdot)$ changes smoothly between any two adjacent levels j and $j+1$. This leads to a quadratic second-order difference penalty,

$$J(\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\boldsymbol{\alpha}_3)=\sum_{k=1}^{3}w_k\sum_{j=2}^{p-1}\left\{\alpha_k(j+1)-2\alpha_k(j)+\alpha_k(j-1)\right\}^2, \tag{2.3}$$

where w_k, for $k=1,2,3$, are weights for each coefficient function. The weights w_k are introduced to make the quadratic second order for $\alpha_k(\cdot)$ comparable by taking variations of corresponding variables into account. Consequently, we can avoid using separate tuning parameters for each $\alpha_k(\cdot)$. In the simulation studies and the real-data analysis, we choose $w_1=\text{median}\{\text{SD}(V_{tj}), j=1,\ldots,p\}, w_2=\text{median}\{\text{SD}(X_{tj}), j=1,\ldots,p\}$, and $w_3=\text{median}\{\text{SD}(X_{tj}V_{tj}), j=1,\ldots,p\}$. The simulation results suggest those choices perform well. This penalty mimics the cubic spline by penalizing the L_2-norm of the discrete version of the second-order derivatives of the coefficients $\alpha_k(\cdot)$, which encourages the smoothness of the coefficients (Guo et al. (2015)). Compared with the fused lasso penalty (Tibshirani et al. (2005)), the above penalty (2.3) is computationally simple and captures smoothly varying features.

It is straightforward to develop a Newton–Raphson algorithm to solve the minimization problem given in (2.1). The following notation is necessary to present the gradient and Hessian matrix of $L_n(\boldsymbol{\Theta})$. Let $\boldsymbol{\alpha}=(\boldsymbol{\alpha}_1',\boldsymbol{\alpha}_2',\boldsymbol{\alpha}_3')'$, and let $\phi_h(x)=\phi(x/h)/h$, where $\phi(\cdot)$ is the standard normal density function and $\dot{\phi}_h(x)=\partial\phi_h(x)/\partial x$ is the derivative of $\phi_h(x)$. Then, $\Upsilon_{tj1}(\boldsymbol{\Theta})=V_{tj}$, $\Upsilon_{tjk}(\boldsymbol{\Theta})=X_{tj}V_{tj}^{k-2}\Phi\left((X_{tj}-c_{k-1})/h\right)$, for $k=2,3$, $\Upsilon_{tk}(\boldsymbol{\Theta})=(\Upsilon_{t1k}(\boldsymbol{\Theta}),\ldots,\Upsilon_{tpk}(\boldsymbol{\Theta}))'$, for $k=1,2,3$, and $\Upsilon_t(\boldsymbol{\Theta})=(\Upsilon_{t1}(\boldsymbol{\Theta})',\Upsilon_{t2}(\boldsymbol{\Theta})',\Upsilon_{t3}(\boldsymbol{\Theta})')'$; $\Upsilon_{tjk}(\boldsymbol{\Theta})=X_{tj}V_{tj}^{k-2}\phi_h(X_{tj}-c_{k-1})$ and $\Upsilon_{tk}(\boldsymbol{\Theta})=(\Upsilon_{t1k}(\boldsymbol{\Theta}),\ldots,\Upsilon_{tpk}(\boldsymbol{\Theta}))'$, for $k=2,3$; $\dot{\Upsilon}_{tjk}(\boldsymbol{\Theta})=X_{tj}V_{tj}^{k-2}\dot{\phi}_h(X_{tj}-c_{k-1})$, and $\dot{\Upsilon}_{tk}(\boldsymbol{\Theta})=(\dot{\Upsilon}_{t1k}(\boldsymbol{\Theta}),\ldots,\dot{\Upsilon}_{tpk}(\boldsymbol{\Theta}))'$, for $k=2,3$; $\Omega=\mathbf{D}'\mathbf{D}$; and

$$\mathbf{D} = \begin{pmatrix} 1 & -2 & 1 & 0 & 0 \cdots 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 \cdots 0 & 0 & 0 \\ 0 & 0 & 1 & -2 & 1 \cdots 0 & 0 & 0 \\ \cdots & & & & & & \\ 0 & 0 & 0 & 0 & 0 \cdots 1 & -2 & 1 \end{pmatrix}_{(p-2)\times p} .$$

First, we obtain

$$\frac{\partial L_n(\boldsymbol{\Theta})}{\partial \mu} = -\frac{2}{n}\sum_{t=1}^{n}\left[Y_t - \mu - \boldsymbol{\alpha}'\Upsilon_t(\boldsymbol{\Theta})\right],$$
$$\frac{\partial L_n(\boldsymbol{\Theta})}{\partial \boldsymbol{\alpha}} = -\frac{2}{n}\sum_{t=1}^{n}\left[Y_t - \mu - \boldsymbol{\alpha}'\Upsilon_t(\boldsymbol{\Theta})\right]\Upsilon_t(\boldsymbol{\Theta}) + 2(\lambda I_3 \otimes \Omega)\boldsymbol{\alpha},$$
$$\frac{\partial L_n(\boldsymbol{\Theta})}{\partial c_1} = \frac{2}{n}\sum_{t=1}^{n}\left[Y_t - \mu - \boldsymbol{\alpha}'\Upsilon_t(\boldsymbol{\Theta})\right]\boldsymbol{\alpha}_2'\Upsilon_{t2}(\boldsymbol{\Theta}),$$
$$\frac{\partial L_n(\boldsymbol{\Theta})}{\partial c_2} = \frac{2}{n}\sum_{t=1}^{n}\left[Y_t - \mu - \boldsymbol{\alpha}'\Upsilon_t(\boldsymbol{\Theta})\right]\boldsymbol{\alpha}_3'\Upsilon_{t3}(\boldsymbol{\Theta}),$$

where I_3 is a three-dimensional identity matrix, and $\otimes$ denotes the Kronecker product. Then, the gradient of $L_n(\boldsymbol{\Theta})$ is given by

$$\mathbf{g}(\boldsymbol{\Theta}) \triangleq \frac{\partial L_n(\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}} = \left(\frac{\partial L_n(\boldsymbol{\Theta})}{\partial \mu}, \frac{\partial L_n(\boldsymbol{\Theta})}{\partial \boldsymbol{\alpha}'}, \frac{\partial L_n(\boldsymbol{\Theta})}{\partial c_1}, \frac{\partial L_n(\boldsymbol{\Theta})}{\partial c_2}\right)'.$$

The elements for the Hessian matrix $\mathbf{H}(\boldsymbol{\Theta}) \triangleq \partial^2 L_n(\boldsymbol{\Theta})/(\partial\boldsymbol{\Theta}\partial\boldsymbol{\Theta}')$ are given by

$$\frac{\partial^2 L_n(\boldsymbol{\Theta})}{\partial \mu^2} = 2, \quad \frac{\partial^2 L_n(\boldsymbol{\Theta})}{\partial \mu \partial \boldsymbol{\alpha}'} = \frac{2}{n}\sum_{t=1}^{n}\Upsilon_t(\boldsymbol{\Theta})',$$
$$\frac{\partial^2 L_n(\boldsymbol{\Theta})}{\partial \mu \partial c_1} = -\frac{2}{n}\sum_{t=1}^{n}\boldsymbol{\alpha}_2'\Upsilon_{t2,c_1}(\boldsymbol{\Theta}), \quad \frac{\partial^2 L_n(\boldsymbol{\Theta})}{\partial \mu \partial c_2} = -\frac{2}{n}\sum_{t=1}^{n}\boldsymbol{\alpha}_3'\Upsilon_{t3,c_2}(\boldsymbol{\Theta}),$$
$$\frac{\partial^2 L_n(\boldsymbol{\Theta})}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}'} = \frac{2}{n}\sum_{t=1}^{n}\Upsilon_t(\boldsymbol{\Theta})\Upsilon_t(\boldsymbol{\Theta})' + 2(\lambda I_3 \otimes \Omega),$$
$$\frac{\partial^2 L_n(\boldsymbol{\Theta})}{\partial \boldsymbol{\alpha}\partial c_1} = -\frac{2}{n}\sum_{t=1}^{n}\boldsymbol{\alpha}_2'\Upsilon_{t2,c_1}(\boldsymbol{\Theta}) + \frac{2}{n}\sum_{t=1}^{n}\left[Y_t - \mu - \boldsymbol{\alpha}'\Upsilon_t(\boldsymbol{\Theta})\right]\begin{pmatrix} 0 \\ \Upsilon_{t2,c_1}(\boldsymbol{\Theta}) \\ 0 \end{pmatrix},$$
$$\frac{\partial^2 L_n(\boldsymbol{\Theta})}{\partial \boldsymbol{\alpha}\partial c_2} = -\frac{2}{n}\sum_{t=1}^{n}\boldsymbol{\alpha}_3'\Upsilon_{t3,c_2}(\boldsymbol{\Theta}) + \frac{2}{n}\sum_{t=1}^{n}\left[Y_t - \mu - \boldsymbol{\alpha}'\Upsilon_t(\boldsymbol{\Theta})\right]\begin{pmatrix} 0 \\ 0 \\ \Upsilon_{t3,c_2}(\boldsymbol{\Theta}) \end{pmatrix},$$

$$\frac{\partial^2 L_n(\boldsymbol{\Theta})}{\partial c_1^2} = \frac{2}{n}\sum_{t=1}^{n}\left[\boldsymbol{\alpha}_2'\Upsilon_{t2,c_1}(\boldsymbol{\Theta})\right]^2 - \frac{2}{n}\sum_{t=1}^{n}\left[Y_t - \mu - \boldsymbol{\alpha}'\Upsilon_t(\boldsymbol{\Theta})\right]\boldsymbol{\alpha}_2'\Upsilon_{t2,c_1c_1}(\boldsymbol{\Theta}),$$

$$\frac{\partial^2 L_n(\boldsymbol{\Theta})}{\partial c_1 \partial c_2} = \frac{2}{n}\sum_{t=1}^{n}\left(\boldsymbol{\alpha}_2'\Upsilon_{t2,c_1}(\boldsymbol{\Theta})\right)\left(\boldsymbol{\alpha}_3'\Upsilon_{t3,c_2}(\boldsymbol{\Theta})\right),$$

$$\frac{\partial^2 L_n(\boldsymbol{\Theta})}{\partial c_2^2} = \frac{2}{n}\sum_{t=1}^{n}\left[\boldsymbol{\alpha}_3'\Upsilon_{t3,c_2}(\boldsymbol{\Theta})\right]^2 - \frac{2}{n}\sum_{t=1}^{n}\left[Y_t - \mu - \boldsymbol{\alpha}'\Upsilon_t(\boldsymbol{\Theta})\right]\boldsymbol{\alpha}_3'\Upsilon_{t3,c_2c_2}(\boldsymbol{\Theta}),$$

where $\Upsilon_{t2,c_1}(\boldsymbol{\Theta}) = \partial\Upsilon_{t2}(\boldsymbol{\Theta})/\partial c_1, \Upsilon_{t3,c_2}(\boldsymbol{\Theta}) = \partial\Upsilon_{t3}(\boldsymbol{\Theta})/\partial c_2, \Upsilon_{t2,c_1c_1}(\boldsymbol{\Theta}) = \partial^2 \Upsilon_{t2}(\boldsymbol{\Theta})/\partial c_1^2$, and $\Upsilon_{t3,c_2c_2}(\boldsymbol{\Theta}) = \partial^2\Upsilon_{t3}(\boldsymbol{\Theta})/\partial c_2^2$. Finally, using the initial value $\boldsymbol{\Theta}^{(0)}$ for $\boldsymbol{\Theta}$, we update the estimate of $\boldsymbol{\Theta}$ at the $(k+1)$th iteration with

$$\boldsymbol{\Theta}^{(k+1)} = \boldsymbol{\Theta}^{(k)} - \left(\mathbf{H}(\boldsymbol{\theta}^{(k)})\right)^{-1}\mathbf{g}(\boldsymbol{\Theta}^{(k)}),$$

until convergence.

To initialize the algorithm, we choose initial values of c_1 and c_2, for example, $c_1 = c_2 = \text{median}_{t,j} X_{tj}$. Given c_1 and c_2, we estimate the parameters μ and $\boldsymbol{\alpha}$ by minimizing the squared errors without a penalty, which is the standard least squares problem.

3. Large-Sample Properties of the Estimators

Now, we establish the consistency and asymptotic normality of the FTAR estimator. Without loss of generality, we assume the support of X_{tj} is $[0, 1]$. Some regularity conditions are stated in Appendix A. Denote $\boldsymbol{\Theta}_1 \equiv (\mu, \boldsymbol{\alpha}')'$, and the true values of $\boldsymbol{\Theta}$, $\boldsymbol{\Theta}_1$, and $\mathbf{c}$ by $\boldsymbol{\Theta}_0$, $\boldsymbol{\Theta}_{10}$, and $\mathbf{c}_0$, respectively. The consistency of $\hat{\boldsymbol{\Theta}}$ is presented in Theorem 1.

Theorem 1. *From Conditions* A.1 *to* A.3 *in Appendix* A, *it follows that*

$$\| \hat{\boldsymbol{\Theta}}_1 - \boldsymbol{\Theta}_{10} \| = O_p(n^{-1/2} + \lambda) \text{and} \| \hat{\mathbf{c}} - \mathbf{c}_0 \| = O_p\left(\sqrt{\frac{h}{n}} + \lambda\sqrt{h}\right).$$

It is expected that there exists a root-n consistent penalized estimator for the common regression coefficients $\boldsymbol{\Theta}_1$, with $\lambda = o(1/\sqrt{n})$. However, the estimator for $\mathbf{c}$ converges to the true values at rate $O(\sqrt{h/n})$ with $\lambda = o(1/\sqrt{n})$, which is faster than root-n. Although a little surprising, this result is not new, and has been observed by Seo and Linton (2007). It occurs because $\mathbf{c}$ is observed using $I(X_{tj} > c_1)$ and $I(X_{tj} > c_2)$, which are indicator functions from zero to one. The jump implies an infinite derivative and yields a large amount of information for $\mathbf{c}$. From simulation experiments, we observe that the mean squared errors for the estimator of $\mathbf{c}$ are smaller than those for other regression coefficients. See Table 1 for details.

Furthermore, under mild conditions, the penalized smoothing estimator is asymptotically normal.

Theorem 2. *From Conditions* A.1 *to* A.3 *in Appendix* A*, it follows that*

$$\sqrt{n}(\hat{\mathbf{\Theta}}_1 - \mathbf{\Theta}_{10} + \lambda V_{11}^{-1}\mathbf{b}) \to N(0, V_{11}^{-1}\Sigma_2 V_{11}^{-1}),$$
$$\sqrt{\frac{n}{h}}(\hat{\mathbf{c}} - \mathbf{c}_0 - h\lambda V_{22}^{-1}V_{12}'V_{11}^{-1}\mathbf{b}) \to N(0, V_{22}^{-1}\Sigma_1 V_{22}^{-1}),$$

where $V_{11}, V_{22}, V_{12}, \Sigma_1, \Sigma_2$*, and* $\mathbf{b}$ *are defined in Appendix* B.

Therefore, both $\hat{\mathbf{\Theta}}_1$ and $\hat{\mathbf{c}}$ can be made asymptotically unbiased by choosing a small tuning parameter $\lambda = o(1/\sqrt{n})$. These proofs of Theorems 1 and 2 are provided in Appendix C.

4. Selection of Bandwidth and Smoothing Parameter

The estimation procedure requires that we select a bandwidth h. The leading terms for the estimators of the regression parameters, $\hat{\mathbf{\Theta}}_1$, are independent of the bandwidth h, indicating that h is not crucial to the asymptotic performance of $\hat{\mathbf{\Theta}}_1$. The asymptotic variance and bias of $\hat{\mathbf{c}}$ are of order $O(h/n)$ and $\sqrt{nh}\lambda$, respectively, which both decrease as h decrease. Thus, a smaller h may lead to a better estimator. However, numerical studies show that for extremely small h, the proposed estimator may be unstable. Our extensive empirical results suggest that h can be the minimum difference between any two values of X_t, such that $\Phi((X_{tj} - c_k)/h)$ well approximates the indicator function around the threshold parameters. In practice, we can generate a sequence of h around this minimum and find an appropriate value that generates stable estimates.

Next, we consider the smoothing parameter λ for $\boldsymbol{\alpha}$. Most existing tuning parameter selection methods are designed for independent data. Cai, Fan and Yao (2000) proposed an analogue to the cross-validation (CV) method for time series data, and we use their method to select λ. Given a sufficiently small m, we first use R subseries of lengths $n - r \times m (r = 1, \ldots, R)$ from the beginning to estimate the unknown coefficient functions and parameters. Then, we compute the one-step forecasting errors for the subsequesnt section of the time series with length m, based on the estimated model. Finally, we choose λ that minimizes the average mean squared (AMS) error, $AMS(\lambda) = \sum_{r=1}^{R} AMS_r(\lambda)$, where

$$AMS_r(\lambda) = \frac{1}{m} \sum_{t=n-rm+1}^{n-rm+m} \Bigg\{ Y_t - \hat{\mu}^r - \sum_{j=1}^{p} \hat{\alpha}_1^r(j) V_{tj} - \sum_{k=1}^{2} \sum_{j=1}^{p} \hat{\alpha}_{k+1}^r(j) X_{tj} V_{tj}^{k-1} I(X_{tj} \geq \hat{c}_k^r) \Bigg\}^2, \tag{4.1}$$

and $\hat{\mu}^r$, $\hat{\alpha}_k^r(j)$, $\hat{c}_1^r$, and $\hat{c}_2^r$ are estimates for μ, $\alpha_k(j)$, c_1, and c_2, respectively, based on the data $\{(Y_t, \boldsymbol{X}_t, \boldsymbol{V}_t), t = 1, \ldots, n - rm\}$, given λ, for $k = 1, 2, 3, j = 1, \ldots, p$. Cai, Fan and Yao (2000) suggested using $m = [0.1n]$ and $R = 4$. Our simulation studies and application show that this method yields reasonable smoothing parameters.

5. Simulation Studies

In this section, we present simulation studies that assess the finite-sample performance of the FTAR method. We evaluate the performance of the method by comparing it to that of the least squares method without penalty (termed LS-UNP). In this way, we can determine the efficiency of the proposed method as a result of incorporating the smoothness of $\boldsymbol{\alpha}_k(\cdot)$. We are also interested in the effects of the bandwidth h and the smoothing parameter λ on the resulting estimators. Finally, we investigate the performance of the FTAR in choosing λ_n using (4.1). The performance of the estimators is assessed using the empirical bias and the standard deviation of the resulting estimators. Specifically, for $\boldsymbol{\alpha}_k = (\alpha_k(1), \ldots, \alpha_k(p))'$, we assess empirical Bias $= [p^{-1} \sum_{j=1}^{p} \{E^* \hat{\alpha}_k(j) - \alpha_k(j)\}^2]^{1/2}$, SD $= [p^{-1} \sum_{j=1}^{p} E^* \{\hat{\alpha}_k(j) - E^* \hat{\alpha}_k(j)\}^2]^{1/2}$, and the root MSE, RMSE $= (\text{Bias}^2 + \text{SD}^2)^{1/2}$, where $E^*(\cdot)$ is the empirical expectation over 200 simulated data sets.

We simulate observations under the following model:

$$Y_t = \sum_{j=1}^{p} \alpha_1(j) Y_{t-j} + \sum_{j=1}^{p} \alpha_2(j) X_{t-j} I(X_{t-j} \geq c_1) + \sum_{j=1}^{p} \alpha_3(j) X_{t-j} Y_{t-j} I(X_{t-j} \geq c_2) + \varepsilon_t, \tag{5.1}$$

where $X_t \sim \text{Unif}(0, 4)$, $\varepsilon_t \sim \mathcal{N}(0, 1)$, $c_1 = 2$, and $c_2 = 3$. We choose $\alpha_k(j)$, for $k = 1, 2, 3$, according to the following three cases:

(1) if $p = 15$,

- $\alpha_1(j) = -0.006(j-(p-10)/2)^2/((p/2-5)^2 + 0.03$,
- $\alpha_2(j) = -4.5(j-(p-4)/2)^2/((p/2-2)^2 + 18$,
- $\alpha_3(j) = -0.004(j-(p-6)/2)^2/((p/2-3)^2 + 0.02$;

(2) if $p = 30$,

- $\alpha_1(j) = -0.006(j-(p-10)/2)^2/((p/2-5)^2 + 0.0165$,
- $\alpha_2(j) = -4.5(j-(p-4)/2)^2/((p/2-2)^2 + 14.4$,
- $\alpha_3(j) = -0.003(j-(p-6)/2)^2/((p/2-3)^2 + 0.01$;

(3) if $p = 60$,

- $\alpha_1(j) = -0.006(j-(p-10)/2)^2/((p/2-5)^2 + 0.0075$,
- $\alpha_2(j) = -4.5(j-(p-4)/2)^2/((p/2-2)^2 + 18$,
- $\alpha_3(j) = -0.003(j-(p-6)/2)^2/((p/2-3)^2 + 0.006$.

The idea behind these settings is that, for each p, we select $\alpha_1(j), \alpha_2(j)$, and $\alpha_3(j)$ to generate similar variances for each term in (5.1). In addition, we consider a fourth setting with $c_1 = 2$, $p = 10$, and $\varepsilon_t \sim \mathcal{N}(0, 0.3^2)$, in which $\alpha_1(j)$ and $\alpha_2(j)$ have similar shapes to those of the real data without the interaction term:

(4)
- $\alpha_1(j) = 0.0008(j-12)^2 - 0.0008$,
- $\alpha_2(j) = 0.004(j-5)^2 + 0.025$.

For each setting, we use a sample size $n = 200$ and sequences of a length $n + p$ to accommodate the auto-regression structure. The summary statistics of the estimated parameters are reported in Table 1 for $p = 15, 30, 60$, and 10, from which we draw the following conclusions:

(1) Both the FTAR and the LS-UNP methods are unbiased. The estimator for $\mathbf{c}$ performs similarly in most of cases because there is no penalty on $\mathbf{c}$. However, the FTAR method for $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2$, and $\boldsymbol{\alpha}_3$ generates much smaller standard deviations and, hence, has much smaller MSEs than those of the LS-UNP method. This suggests that the FTAR method for $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2$, and $\boldsymbol{\alpha}_3$ is better than the LS-UNP method, in terms of the MSE.

Table 1. Simulation results for $p = 15, 30, 60, 10$ using the proposed method with $\lambda = 0.007, 0.60, 4.88, 1.54$, respectively, and the LS-UNP method.

		Proposed			LS-UNP		
		Bias	SD	RMSE	Bias	SD	RMSE
p=15	$\boldsymbol{\alpha}_1$	0.00087	0.00793	0.00797	0.00077	0.01614	0.01616
	$\boldsymbol{\alpha}_2$	0.03005	0.16802	0.17068	0.01875	0.21356	0.21438
	$\boldsymbol{\alpha}_3$	0.00004	0.00036	0.00036	0.00003	0.00039	0.00039
	c_1	0.00001	0.00145	0.00145	0.00003	0.00147	0.00147
	c_2	0.00007	0.00108	0.00108	0.00004	0.00098	0.00098
p=30	$\boldsymbol{\alpha}_1$	0.00105	0.00524	0.00535	0.00184	0.01681	0.01691
	$\boldsymbol{\alpha}_2$	0.08210	0.42211	0.43002	0.08953	0.62154	0.62796
	$\boldsymbol{\alpha}_3$	0.00005	0.00033	0.00034	0.00007	0.00049	0.00050
	c_1	0.00011	0.00178	0.00178	0.00006	0.00165	0.00166
	c_2	0.00004	0.00186	0.00186	0.00022	0.00296	0.00297
p=60	$\boldsymbol{\alpha}_1$	0.000253	0.002078	0.002094	0.001629	0.021288	0.021350
	$\boldsymbol{\alpha}_2$	0.021940	0.216963	0.218070	0.033450	0.344298	0.345919
	$\boldsymbol{\alpha}_3$	0.000008	0.000112	0.000112	0.000014	0.000191	0.000191
	c_1	0.000019	0.000260	0.000261	0.000011	0.000131	0.000131
	c_2	0.000002	0.000056	0.000056	0.000002	0.000031	0.000032
p=10	$\boldsymbol{\alpha}_1$	0.00645	0.02048	0.02147	0.00671	0.06292	0.06328
	$\boldsymbol{\alpha}_2$	0.00284	0.00947	0.00989	0.00125	0.01537	0.01542
	c_1	0.00013	0.00224	0.00224	0.00079	0.01127	0.01129

(2) Comparing the simulation results for $p = 15, 30, 60$, and 10, we see that the differences between MSEs of the proposed method and the LS-UNP method increases as p increases. This may be because the degrees of freedom for the parameter space are controlled in our method as a result of smoothing $\hat{\alpha}_k(\cdot)$, for $k = 1, 2, 3$. In contrast, the dimension of the parameter space for the LS-UNP method increases linearly with p.

(3) The empirical standard deviations of the proposed estimators for c_1 and c_2 are smaller than those for $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ under all settings in Table 1, whereas $\boldsymbol{\alpha}_3$ has a small MSE owing to its small scale. These results confirm the asymptotic result in Theorem 2 that $\hat{\mathbf{c}}$ has a faster convergence rate than $\hat{\boldsymbol{\alpha}}_k$, for $k = 1, 2, 3$.

Figures 1, 2, 3, and 4 present the estimates and 95% point-wise confidence bands of $\alpha_1(j)$, $\alpha_2(j)$, and $\alpha_3(j)$, for $j = 1, \ldots, p$, under four settings ($p = 15, 30, 60$ and 10), using the proposed method with an AMS-tuned λ. The results in Figures 1 to 4 suggest that the performance of the proposed method with AMS-tuned parameters is satisfactory.

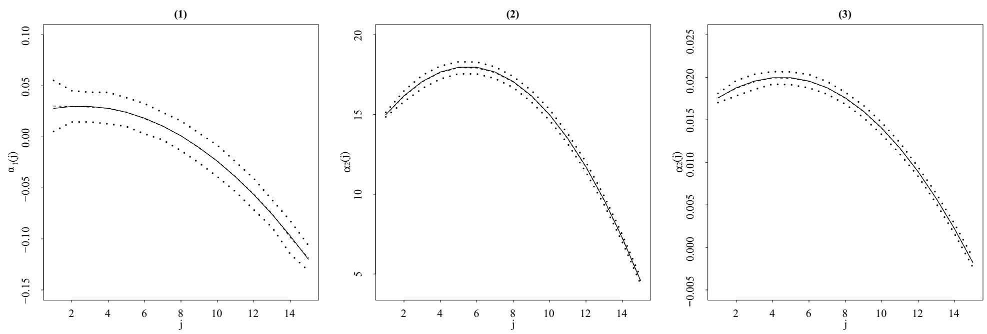

Figure 1. Panels (1), (2), and (3) show the estimates for $\alpha_1(j), \alpha_2(j)$, and $\alpha_3(j)$, respectively, for $j = 1, \ldots, 15$, and the associated 95% point-wise confidence bands for $p = 15$ and AMS-tuned $\lambda = 0.007$.

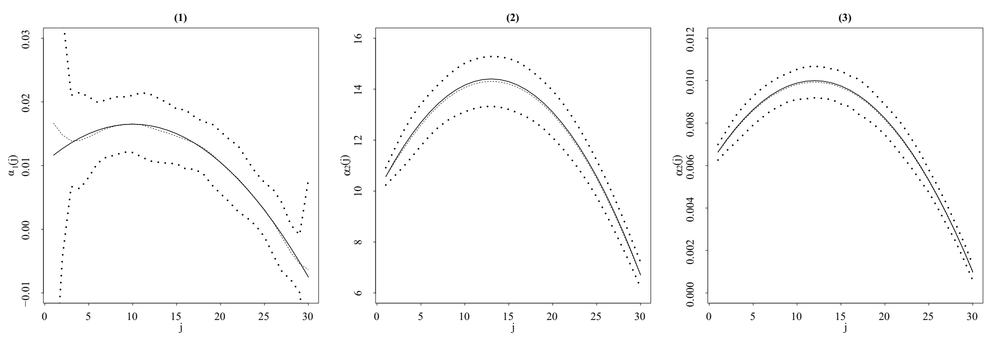

Figure 2. Panels (1), (2), and (3) show the estimates for $\alpha_1(j), \alpha_2(j)$, and $\alpha_3(j)$, respectively, for $j = 1, \ldots, 30$, and the associated 95% point-wise confidence bands for $p = 30$ and AMS-tuned $\lambda = 0.60$.

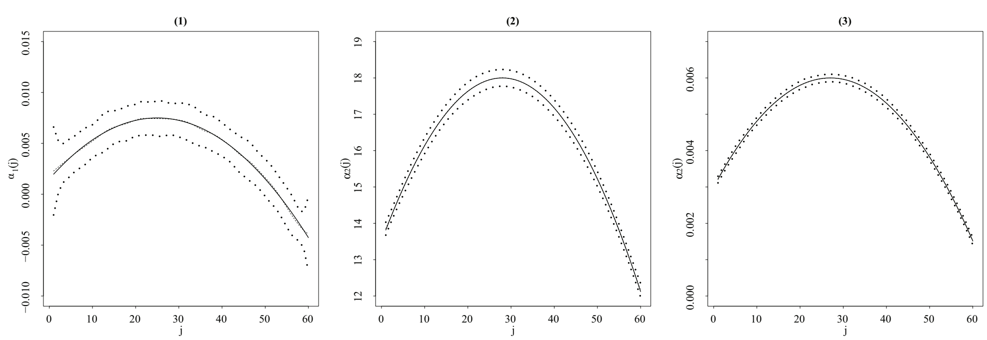

Figure 3. Panels (1), (2), and (3) show the estimates for $\alpha_1(j), \alpha_2(j)$, and $\alpha_3(j)$, respectively, for $j = 1, \ldots, 60$, and the associated 95% point-wise confidence bands for $p = 60$ and AMS-tuned $\lambda = 4.88$.

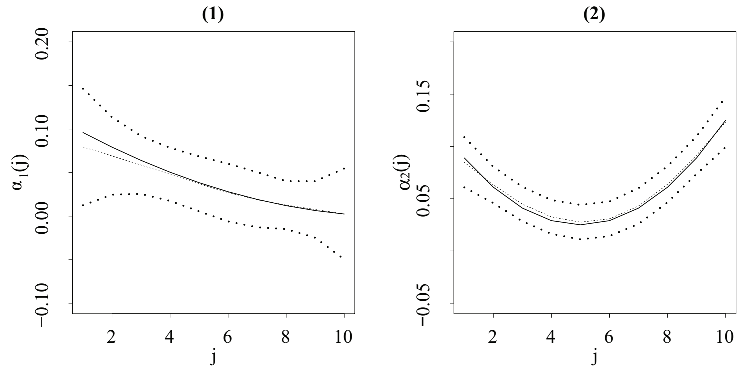

Figure 4. Panels (1) and (2) show the estimates of $\alpha_1(j)$ and $\alpha_2(j)$, respectively, for $j = 1, \ldots, 10$, and the associated 95% point-wise confidence bands for $p = 10$ and AMS-tuned $\lambda = 1.54$.

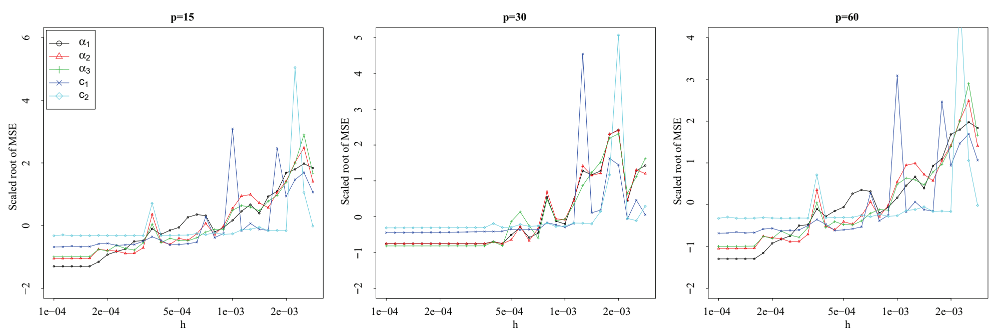

Figure 5. RMSE for $p = 15, 30, 60$ for various values of h, given $\lambda = 0.007, 0.60, 4.88$, respectively.

Finally, we investigate the effect of varying h on the resulting estimates. We fix λ at $0.44, 6.66, 30$ for $p = 15, 30, 60$, respectively. In order to show all RMSEs in the same figure, we scale the sequence of RMSEs over h to a one-unit variance for each coefficient function. The scaled RMSEs against h for each parameter are shown in Figure 5, and suggest that a smaller h is preferred. However, extremely small h causes sensitivity to initial values. Therefore, we fix $h = 0.001$ for $p = 10, 15, 30$, and $h = 0.0001$ for $p = 60$.

6. Analyzing Hong Kong Suicide Data and Media Coverage

In order to examine whether media reports of suicides influence actual suicides, we apply the proposed method to analyze the coverage of reported suicides

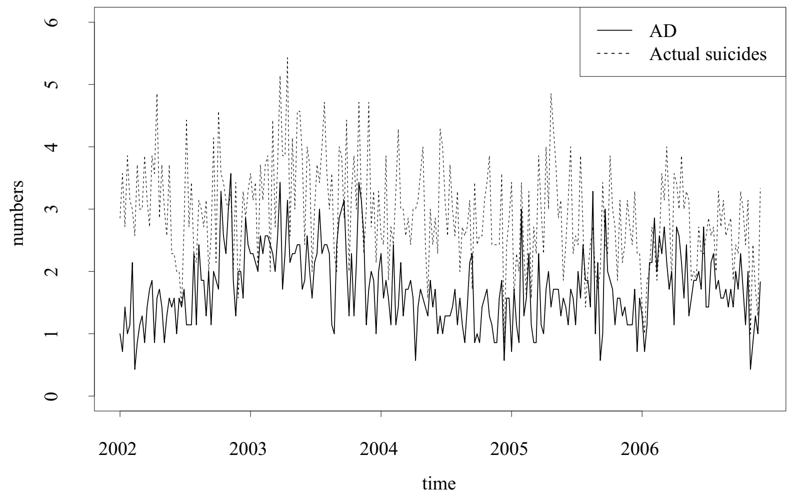

Figure 6. Plots of the averaged daily numbers of suicides reported by AD and actual suicides in Hong Kong for each week from January 2002 to December 2006.

in a Hong Kong-based tabloid newspaper (the AD) during the period January 2002 to December 2006 identified using the WISENEWS search. In total, there were 1,827 such reports. January 2002 to December 2006 is taken as the study period because Hong Kong's suicide rate increased to a historical maximum, with a rate of 18.6 per 100,000 (i.e., 1,264 suicide deaths in 2003, or an average of about three deaths each day).

The daily number of reports are collected. Because many days do not include such reports, we aggregate the daily numbers of suicides reported by the AD and of actual suicides by means of a weekly average. The outcome variable Y_t can be viewed as a continuous variable. A histogram and Q–Q plot of the weekly average show that its distribution is close to normal, which means we can apply the proposed model. Figure 6 displays the aggregated daily numbers of reported and actual suicides in Hong Kong for each week from January 2002 to December 2006. The raw curves in Figure 6 show some lags in the spikes between AD reported and actual suicides.

As described in Section 1, we fit the following model on the data:

$$E\left\{Y_t | Y_s, s < t, X_s, s \leq t\right\} = \mu + \sum_{j=1}^{p} \alpha_1(j) Y_{t-j} + \sum_{j=1}^{p} \alpha_2(j) X_{t-j} I(X_{t-j} \geq c_1)$$

Table 2. Estimate, standard deviation, and 95% CI for thresholds.

		Estimate	SD	95% CI
With interaction	c_1	2	0.07	(1.87, 2.13)
	c_2	3	0.05	(2.90, 3.10)
Without interaction	c_1	2	0.05	(1.90, 2.10)

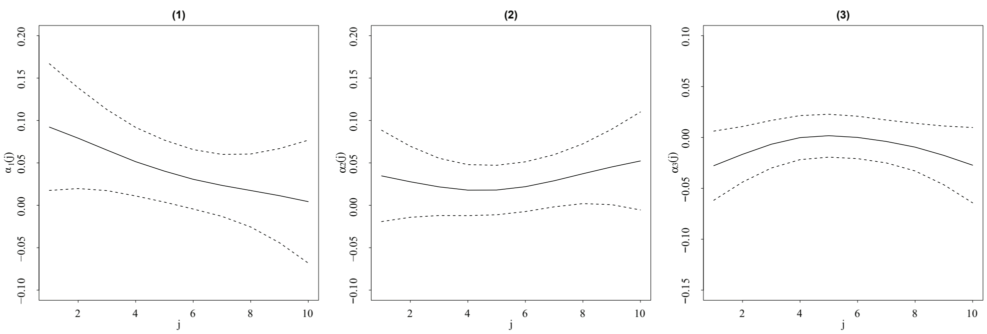

Figure 7. Panels (1), (2), and (3) show the estimates of $\alpha_1(j), \alpha_2(j)$, and $\alpha_3(j)$, respectively, and the associated 95% point-wise confidence bands.

$$+\sum_{j=1}^{p} \alpha_3(j) X_{t-j} Y_{t-j} I(X_{t-j} \geq c_2).$$

First, we consider a relatively large order $p = 10$ to investigate the autoregression property. We take $h = 0.001$, which is close to the minimal difference between any two values of X_t, and can generate stable estimates. The cross-validation method defined in Section 4 yields $\lambda = 18.71$. The estimates of the parameters and their standard deviations are shown in Table 2 and Figure 7. The standard deviation is calculated using the resampling method described in Fan and Yao (2003), with 500 bootstrapping samples. The results in Figure 7 show that $\hat{\alpha}_1(\cdot)$ is significantly different from zero, but $\hat{\alpha}_3(\cdot)$ is not significantly different from zero.

Therefore, we refit the model by removing the interaction term. The estimates for c_1 are also shown in Table 2, and the estimates for $\alpha_1(\cdot)$ and $\alpha_2(\cdot)$ are plotted in Figure 8, which are based on $\lambda = 104.22$. The estimate for $\alpha_1(\cdot)$ implies that the effect of previous suicides decreases over time, as expected. The estimate for $\hat{\alpha}_2(\cdot)$ and its 95% confidence bands suggest that the copycat effect of media coverage is at the borderline of significance, and can last up to eight weeks. Furthermore, $\hat{c}_1 = 2$ implies that a copycat suicide effect occurs when the number of reports is greater than two.

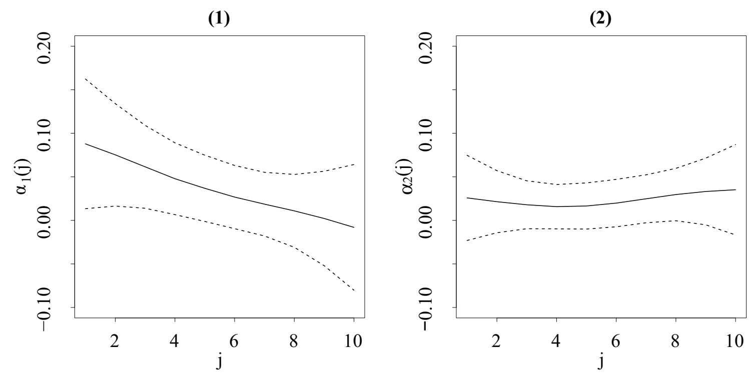

Figure 8. Panels (1) and (2) show the estimates of $\alpha_1(j)$ and $\alpha_2(j)$, respectively, based on the model without the interaction term, and the associated 95% point-wise confidence bands.

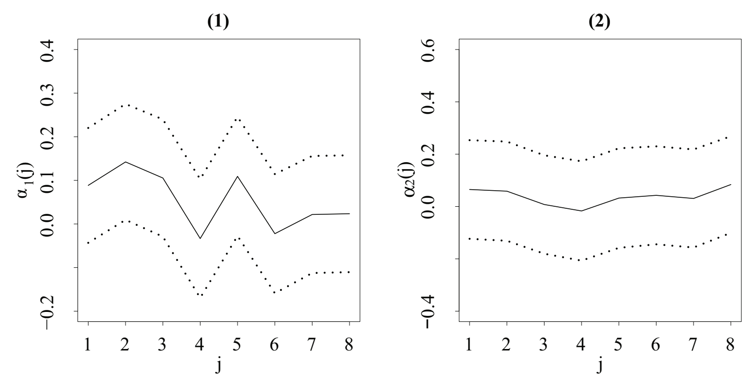

Figure 9. A simple time series regression model with lag 8.

For comparison purposes, we fit the data using a simple time series regression of the form

$$E\{Y_t|Y_s, s < t, X_s, s \leq t\} = \mu + \sum_{j=1}^{p} \alpha_1(j)Y_{t-j} + \sum_{j=1}^{p} \alpha_2(j)X_{t-j},$$

with $p = 8$. The estimates for $\alpha_1(\cdot)$ and $\alpha_2(\cdot)$ are plotted in Figure 9. The results show that $\hat{\alpha}_1(\cdot)$ is marginally different from zero, but that $\hat{\alpha}_2(\cdot)$ is not significantly different from zero. Comparing our results in Figure 8 with those in Figure 9, the proposed method implies a clearer trend and a narrower confidence band and, hence, is more efficient.

7. Conclusion

We have proposed a flexible threshold autoregressive (FTAR) model to explore whether and how media coverage of suicides is related to the incidence of suicides. A penalized smoothing least squares estimator is adopted to estimate the parameters and unknown functions. The proposed FTAR method yields an accurate estimate for the effect of such coverage, which is confirmed by simulation studies. Theoretical properties, including uniform consistency and asymptotic normality, are proved under mild regularity conditions. Our model shows that a copycat suicide effect occurs when the number of reported cases is greater than two.

We also confirm the existence of an association between media coverage and the incidence of suicides. Although the effect initially diminishes, later media coverage can still trigger a copycat effect. Lastly, note that our model threshold parameters identify the occurrence of a copycat suicide effect. It is of further interest to investigate the pattern of such effects; however, this is left to future research.

Acknowledgements

Lin's research was supported by National Natural Science Foundation of China (Nos. 11931014 and 11829101) and Fundamental Research Funds for the Central Universities (No. JBK1806002) of China. Yip's research was supported by a GRF grant (HKU-106140091) and a Strategic Public Policy Research grant (SPPR-HKU-12).

Appendix

A. Appendix

Condition A.1:

1. $f(\boldsymbol{x}_1, \boldsymbol{x}_r|\boldsymbol{v}_1, \boldsymbol{v}_r; r) \leq M \leq \infty$, for all $r \geq 1$, where $f(\boldsymbol{x}_1, \boldsymbol{x}_r|\boldsymbol{v}_1, \boldsymbol{v}_r; r)$ is the conditional density of $(\boldsymbol{X}_1, \boldsymbol{X}_r)$ given $(\boldsymbol{V}_1, \boldsymbol{V}_r)$, and $f(\boldsymbol{v}|\boldsymbol{x}) \leq M < \infty$, where $f(\boldsymbol{v}|\boldsymbol{x})$ is the conditional density of $\boldsymbol{V}_t$ given $\boldsymbol{X}_t = \boldsymbol{x}$.

2. The process $\{\boldsymbol{X}_t, \boldsymbol{V}_t, Y_t\}$ is α-mixing with $\sum_k k^c[\alpha(k)]^{1-2/\delta} < \infty$ for some $\delta > 2$ and $c > 1 - 2/\delta$, where $\alpha(k) = \sup\{|Pr(A \cap B) - Pr(A)Pr(B)|; A \in \mathcal{F}_{-\infty}^0, B \in \mathcal{F}_k^\infty\}$, $\mathcal{F}_a^b$ is the σ-algebra generated by $\{(\boldsymbol{X}_t, \boldsymbol{V}_t, Y_t); a \leq t \leq b\}$.

3. $E|\boldsymbol{V}_t|^{2\delta} < \infty$, where δ is given in condition A.1.2.

4. $\boldsymbol{X}_t$ is bounded with compact support $[0,1]^p$.

Condition A.2:

1. Assume that $E\{Y_1^2 + Y_l^2 | \boldsymbol{X}_1 = \boldsymbol{x}_1, \boldsymbol{X}_l = \boldsymbol{x}_2, \boldsymbol{V}_1 = \boldsymbol{v}_1, \boldsymbol{V}_l = \boldsymbol{v}_2\} \leq M < \infty$ for all $l > 1$.

2. Assume that $h \to 0$ and $nh \to \infty$. Further, assume that there exists a sequence of positive integers s_n such that $s_n \to \infty$, $s_n = o(\sqrt{nh})$, and $(n/h)^{1/2}\alpha(s_n) \to 0$, as $n \to \infty$.

3. There exists $\delta^* > \delta$, where δ is given in condition A.1.3, such that

$$E\{|Y_t|^{\delta^*} | \boldsymbol{V}_t = \boldsymbol{v}, \boldsymbol{X}_t = \boldsymbol{x}\} \leq M < \infty$$

for any $\boldsymbol{v}$ and $\boldsymbol{x}$ in supports of $\boldsymbol{V}_t$ and $\boldsymbol{X}_t$, respectively, and

$$\alpha(n) = O(n^{-\theta^*}),$$

where $\theta^* \geq \delta\delta^*/\{2(\delta^* - \delta)\}$.

4. $E|\boldsymbol{V}_t|^{2\delta^*} < \infty$, and $n^{1/2-\delta/4}h^{\delta/\delta^*-1/2-\delta/4} = O(1)$.

Condition A.3:

1. Let f_j be the density function of X_{tj}. The density function $f_j(\cdot)$ is positive and has continuous second derivatives on $[0,1]$.

2. $\lambda \to 0$, $h^2 log(n) \to 0$ and $nh \to \infty$ as $n \to \infty$.

The above conditions are used for deriving the convergence properties. Conditions A.1 and A.2 are similar to those in Cai, Fan and Yao (2000).

B. Notations and Lemma

Let f_j be the density function of X_{tj}, where $f_{j,s}(x_{tj}, x_{ts})$ is the joint density of X_{tj} and X_{ts}, $\boldsymbol{\Theta}_0$ is the true value of $\boldsymbol{\Theta}$, and $v_k = \int x^k \phi^2(x)dx$.

$$\sigma^2(\boldsymbol{V}_t, \boldsymbol{X}_t) = \mathrm{var}(Y_t | \boldsymbol{V}_t, \boldsymbol{X}_t),$$
$$b_k(\boldsymbol{\Theta}) = \sum_{j=1}^{p} v_0 \alpha_{k+1}^2(j) c_k^2 f_j(c_k) E\big(\sigma^2(\boldsymbol{V}_t, \boldsymbol{X}_t) V_{tj}^{2(k-1)} | X_{tj} = c_k\big),$$

$$
\begin{aligned}
\Upsilon_{tj1,0}(\boldsymbol{\Theta}) &= \Upsilon_{tj1}(\boldsymbol{\Theta}), \Upsilon_{tjk,0}(\boldsymbol{\Theta}) = X_{tj}V_{tj}^{k-1}I\left(\frac{X_{tj} > c_{k-1}}{h}\right), k = 2, 3,\\
C_{tk}(\boldsymbol{\Theta}) &= \sum_{\ell=1}^{p} \alpha_{k+1}(\ell)X_{t\ell}V_{t\ell}^{k-1}\phi\left(\frac{X_{t\ell} - c_k}{h}\right),\\
F_k(\boldsymbol{\Theta}) &= 2\sum_{j=1}^{p} \alpha_{k+1}^2(j)c_k^2 v_0 f_j(c_k)E[V_{tj}^2|X_{tj} = c_k],\\
F(\boldsymbol{\Theta}) &= \sum_{j\neq l}^{p} \alpha_2(j)\alpha_3(l)c_1c_2f_{j,l}(c_1, c_2)E(V_{tl}|X_{tj} = c_1, X_{tl} = c_2),\\
\kappa_{kjl} &= 2\sum_{m=1}^{p} \alpha_{l+1,0}(m)c_{l,0}f_m(c_{l,0})E\big[V_{tm}^{l-1}[V_{tj}I(k = 1)\\
&\quad +X_{tj}V_{tj}^{k-1}I(X_{tj} > c_{k-1,0})I(k \neq 1, j \neq m)\\
&\quad +c_{l,0}V_{tj}I(c_{l,0} > c_{k-1,0})I(k \neq 1)I(j = m)]\big|X_{tm} = c_{l,0}\big],\\
\varphi_l(\boldsymbol{\Theta}) &= 2\sum_{j=1}^{p} \alpha_{l+1}(j)c_l f_j(c_l)E(V_{tj}^{l-1}|X_{tj} = c_l),\\
\delta_{kjmv} &= 4E\sigma^2(\boldsymbol{V}_t, \boldsymbol{X}_t)\Upsilon_{tjk,0}(\boldsymbol{\Theta}_0)\Upsilon_{tvm,0}(\boldsymbol{\Theta}_0),\\
\zeta_1(\Theta) &= \sum_{j\neq s}^{p} \alpha_2(j)\alpha_3(s)c_1c_2f_{j,s}(c_1, c_2)E\big(\sigma^2(\boldsymbol{V}_t, \boldsymbol{X}_t)V_{tj}|X_{tj} = c_1, X_{ts} = c_2\big),\\
\zeta_2 &= 4E\sigma^2(\boldsymbol{V}_t, \boldsymbol{X}_t),\\
\varpi_{kj} &= 4E\sigma^2(\boldsymbol{V}_t, \boldsymbol{X}_t)\Upsilon_{tjk,0}(\boldsymbol{\Theta}_0),\\
\vartheta_{kj} &= 2E\Upsilon_{tjk,0}(\boldsymbol{\Theta}_0),\ \rho_{kjmv} = 2E\Upsilon_{tjk,0}(\Theta_0)\Upsilon_{tvm,0}(\Theta_0),\\
\Sigma_{11} &= \zeta_2,\ \Sigma_{22} = (\delta_{1i1j})_{i\leq p, j\leq p},\ \Sigma_{33} = (\delta_{2i2j})_{i\leq p, j\leq p},\ \Sigma_{44} = (\delta_{3i3j})_{i\leq p, j\leq p},\\
\Sigma_{12} &= (\varpi_{1j})_{j\leq p},\ \Sigma_{13} = (\varpi_{2j})_{j\leq p},\ \Sigma_{14} = (\varpi_{3j})_{j\leq p},\ \Sigma_{23} = (\delta_{1i2j})_{i\leq p, j\leq p},\\
\Sigma_{24} &= (\delta_{1i3j})_{i\leq p, j\leq p},\ \Sigma_{34} = (\delta_{2i3j})_{i\leq p, j\leq p},\\
A_{11} &= 2,\ A_{22} = (\rho_{1j1v})_{1\leq j, v\leq p},\ A_{33} = (\rho_{2j2v})_{1\leq j, v\leq p},\ A_{44} = (\rho_{3j3v})_{1\leq j, v\leq p},\\
A_{12} &= (\vartheta_{1j})_{j\leq p},\ A_{13} = (\vartheta_{2j})_{j\leq p},\ A_{14} = (\vartheta_{3j})_{j\leq p},\ A_{23} = (\rho_{1j2v})_{1\leq j, v\leq p},\\
A_{24} &= (\rho_{1j3v})_{1\leq j, v\leq p},\ A_{34} = (\rho_{2j3v})_{1\leq j, v\leq p},\\
B_1 &= (\varphi_l)_{l=1,2},\ B_2 = (\kappa_{1jl})_{1\leq j\leq p, l=1,2},\ B_3 = (\kappa_{2jl})_{1\leq j\leq p, l=1,2},\\
B_4 &= (\kappa_{3jl})_{1\leq j\leq p, l=1,2}, \Sigma_1 = \mathrm{diag}(b_1(\Theta_0), b_2(\Theta_0)),\ V_{12} = (B_1', B_2', B_3', B_4')',\\
V_{22} &= \mathrm{diag}(F_1, F_2), \mathbf{b} = (0, \boldsymbol{\alpha}_0' A')',
\end{aligned}
$$

$$\mathbf{A} = 2\begin{pmatrix} \Omega & 0 & 0 \\ 0 & \Omega & 0 \\ 0 & 0 & \Omega \end{pmatrix}, \Sigma_2 = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} & \Sigma_{14} \\ \Sigma_{12}' & \Sigma_{22} & \Sigma_{23} & \Sigma_{24} \\ \Sigma_{13}' & \Sigma_{23}' & \Sigma_{33} & \Sigma_{34} \\ \Sigma_{14}' & \Sigma_{24}' & \Sigma_{34}' & \Sigma_{44} \end{pmatrix}, V_{11} = \begin{pmatrix} A_{11} & A_{12} & A_{13} & A_{14} \\ A_{12}' & A_{22} & A_{23} & A_{24} \\ A_{13}' & A_{23}' & A_{33} & A_{34} \\ A_{14}' & A_{24}' & A_{34}' & A_{44} \end{pmatrix}.$$

Denote $U_k(\mathbf{\Theta}) = n^{-1}\sum_{t=1}^{n} R_{t,k}, k = 1, 2,$ where

$$R_{t,k} = \left[Y_t - \mu - \sum_{j=1}^{p}\alpha_1(j)V_{tj} - \sum_{k=1}^{2}\sum_{j=1}^{p}\alpha_{k+1}(j)X_{tj}V_{tj}^{k-1}\Phi\left(\frac{X_{tj}-c_k}{h}\right)\right] \times \left\{\sum_{j=1}^{p}\alpha_{k+1}(j)X_{tj}V_{tj}^{k-1}\phi_h\left(X_{tj}-c_k\right)\right\}.$$

Lemma B.1 *Under Conditions* A.1 *and* A.2*, if $h \to 0$ and $nh \to \infty$ as $n \to \infty$, we have*

(a) $hvar\{R_{t,k}(\mathbf{\Theta}_0)\} = b_k(\mathbf{\Theta}_0) + o(1)$;

(b) $h\sum_{j=1}^{n-1}|cov(R_{1,k}(\mathbf{\Theta}_0), R_{1+j,k}(\mathbf{\Theta}_0))| = o(1)$;

(c) $nhvar\{U_k(\mathbf{\Theta}_0)\} = b_k(\mathbf{\Theta}_0) + o(1)$.

Proof. Denote

$$R_{t,k0} = \left[Y_t - \mu - \sum_{j=1}^{p}\alpha_1(j)V_{tj} - \sum_{k=1}^{2}\sum_{j=1}^{p}\alpha_{k+1}(j)X_{tj}V_{tj}^{k-1}I\left(X_{tj} > c_k\right)\right] \times \left\{\sum_{j=1}^{p}\alpha_{k+1}(j)X_{tj}V_{tj}^{k-1}\phi_h\left(X_{tj}-c_k\right)\right\}, k = 1, 2.$$

Suppose that Z is a standard normal variable. Then we have tail probability,

$$\begin{aligned} 1 - \Phi(t) = P(Z \geq t) &= \frac{1}{\sqrt{2\pi}}\int_t^\infty e^{-z^2/2}dz \leq \frac{1}{\sqrt{2\pi}}\int_t^\infty \frac{z}{t}e^{-z^2/2}dz \\ &\leq \frac{1}{\sqrt{2\pi}t}e^{-t^2/2}, \end{aligned} \tag{B.1}$$

for any $t > 0$. We first consider the case of $X_{tj} > c_k$. To simplify the notation, let $x = (X_{tj} - c_k)/h$ which is positive, since

$$|\Phi(x) - I(x > 0)| = |\Phi(x) - 1|I(X_{tj} \geq c_k + \sqrt{h})$$

$$+|\Phi(x)-1|I(c_k < X_{tj} < c_k+\sqrt{h}),$$

then, $\forall\epsilon > 0$ and $\forall s \geq 1$,

$$\begin{aligned}
&P\left(\frac{|\Phi(x)-I(x>0)|}{h^s} > \epsilon|X_{tj} > c_k\right)\\
&= P\Bigg(\frac{|\Phi(x)-1|I(X_{tj}\geq c_k+\sqrt{h})+|\Phi(x)-1|I(c_k < X_{tj} < c_k+\sqrt{h})}{h^s}\\
&\quad > \epsilon|X_{tj} > c_k\Bigg)\\
&\leq P\left(\frac{|\Phi(x)-1|I(X_{tj}\geq c_k+\sqrt{h})}{h^s} > \frac{\epsilon}{2}|X_{tj} > c_k\right)\\
&\quad +P\left(\frac{|\Phi(x)-1|I(c_k < X_{tj} < c_k+\sqrt{h})}{h^s} > \frac{\epsilon}{2}|X_{tj} > c_k\right)\\
&\leq P\left(\frac{|\Phi(x)-1|I(x\geq 1/\sqrt{h})}{h^s} > \frac{\epsilon}{2}|X_{tj} > c_k\right)\\
&\quad +P(c_k < X_{tj} < c_k+\sqrt{h}|X_{tj} > c_k)\\
&\equiv I_1+I_2.
\end{aligned}$$

By (B.1), we have

$$I_1 \leq P\left(\frac{(1/\sqrt{2\pi}x)e^{-x^2/2}I(x\geq 1/\sqrt{h})}{h^s} > \frac{\epsilon}{2}|X_{tj} > c_k\right). \tag{B.2}$$

Noting that $f(y)=(1/y)e^{-y^2/2}$ is monotonically decreasing function for $y>0$, we have $((1/\sqrt{2\pi}x)e^{-x^2/2})/h^s \leq (1/\sqrt{2\pi h^{2s-1}})e^{-1/2h} \to 0$ when $x \geq 1/\sqrt{h}$. Then

$$I_1 \leq P\left(\frac{1}{\sqrt{2\pi h^{2s-1}}}e^{-1/2h}I\left(x\geq\frac{1}{\sqrt{h}}\right) > \frac{\epsilon}{2}|X_{tj} > c_k\right)\to 0.$$

Moreover, under condition A.3.1, we have

$$I_2 = P(c_k < X_{tj} < c_k+\sqrt{h}|X_{tj} > c_k)) = O(\sqrt{h}).$$

Thus,

$$\lim_{n\to\infty} P\left(\frac{|\Phi(x)-I(x>0)|}{h^s} > \epsilon|X_{tj} > c_k\right) = 0$$

which implies $|\Phi((X_{tj}-c_k)/h)-I(X_{tj}>c_k)|=o_p(h^s)$ for any $s\geq 1$. Similarly, the same conclusion holds for $X_{tj}<c_k$. In summary, we have

$$\left|\Phi\left(\frac{X_{tj}-c_k}{h}\right)-I(X_{tj}>c_k)\right|=o_p(h^s),\ \text{for any } s\geq 1. \tag{B.3}$$

Hence,

$$\begin{aligned}R_{t,k}(\mathbf{\Theta})&=R_{t,k0}(\mathbf{\Theta})+R_{t,k}(\mathbf{\Theta})-R_{t,k0}(\mathbf{\Theta})\\&=R_{t,k0}(\mathbf{\Theta})+o_p(h^s).\end{aligned} \tag{B.4}$$

The rest of the proof is similar to that of Lemma B.1 in Cai, Fan and Yao (2000) and only give the proof of (a).

By conditioning on $(\mathbf{V}_t,\mathbf{X}_t)$, we have

$$\begin{aligned}&Var\big(R_{t,k0}(\Theta)\big)\\&=\sum_{j=1}^{p}E\sigma^2(\mathbf{V}_t,\mathbf{X}_t)\alpha_{k+1}^2(j)X_{tj}^2V_{tj}^{2(k-1)}\phi_h^2\left(X_{tj}-c_k\right)\\&\quad+E\sigma^2(\mathbf{V}_t,\mathbf{X}_t)\sum_{j\neq j'}^{p}\alpha_{k+1}(j)\alpha_{k+1}(j')X_{tj}X_{tj'}V_{tj}^{k-1}V_{tj'}^{k-1}\phi_h\left(X_{tj}-c_k\right)\phi_h\left(X_{tj'}-c_k\right)\\&=I_1+I_2.\end{aligned}$$

Firstly,

$$\begin{aligned}I_1&=\sum_{j=1}^{p}\alpha_{k+1}^2(j)E\int\sigma^2(\mathbf{V}_t,\mathbf{X}_t)X_{tj}^2V_{tj}^{2(k-1)}\phi_h^2\left(X_{tj}-c_k\right)f_j(X_{tj})dX_{tj}\\&=\frac{1}{h^2}\sum_{j=1}^{p}\alpha_{k+1}^2(j)E\int\sigma^2(\mathbf{V}_t,\mathbf{X}_t)X_{tj}^2V_{tj}^{2(k-1)}\phi^2\left(\frac{X_{tj}-c_k}{h}\right)f_j(X_{tj})dX_{tj}\\&=\frac{1}{h}\sum_{j=1}^{p}\alpha_{k+1}^2(j)E\int\sigma^2(\mathbf{V}_t,\mathbf{X}_t)(zh+c_k)^2V_{tj}^{2(k-1)}\phi^2\left(z\right)f_j(zh+c_k)dz\\&=\frac{1}{h}\sum_{j=1}^{p}\alpha_{k+1}^2(j)E\int\sigma^2(\mathbf{V}_t,\mathbf{X}_t)\{c_k^2\}V_{tj}^{2(k-1)}\phi^2\left(z\right)\{f_j(c_k)\}dz\\&\quad+\frac{1}{h}\sum_{j=1}^{p}\alpha_{k+1}^2(j)E\int\sigma^2(\mathbf{V}_t,\mathbf{X}_t)\{c_k^2\}V_{tj}^{2(k-1)}\phi^2\left(z\right)\{f_j'(c_k)zh\}dz\end{aligned}$$

$$+\frac{1}{h}\sum_{j=1}^{p}\alpha_{k+1}^2(j)E\int\sigma^2(\mathbf{V}_t,\mathbf{X}_t)\{2zhc_k\}V_{tj}^{2(k-1)}\phi^2(z)\{f_j(c_k)\}dz$$
$$+\frac{1}{h}\sum_{j=1}^{p}\alpha_{k+1}^2(j)E\int\sigma^2(\mathbf{V}_t,\mathbf{X}_t)\{2zhc_k\}V_{tj}^{2(k-1)}\phi^2(z)\{f_j'(c_k)zh\}dz+\cdots$$
$$=\frac{1}{h}b_k(\Theta)+O(1)+O(1)+O(h)+o(h).$$

Similarly, one can prove $I_2=o(h)$. Thus, we can get

$$hVar\big(R_{t,k0}(\Theta_0)\big)=b_k(\Theta_0)+O(h)=b_k(\Theta_0)+o(1).$$

C. Proofs of Theorems

Proof of Theorem 1.

Let $\alpha_{1n}=n^{-1/2}+\lambda$, $\alpha_{2n}=\sqrt{h/n}+\sqrt{h}\lambda$. Denote $\boldsymbol{\Theta}_0=(\mu_0,\boldsymbol{\alpha}_{10}',\boldsymbol{\alpha}_{20}',\boldsymbol{\alpha}_{30}',\mathbf{c}_0')'$ to be the true value of $\boldsymbol{\Theta}$. We wish to show that for any given $\varepsilon>0$, there exists a large constant τ_1,τ_2 such that

$$\Pr\left[\min_{\|\mathbf{u_1}\|=\tau_1,\|\mathbf{u_2}\|=\tau_2}L_n\{\boldsymbol{\Theta}_0+(\alpha_{1n}\mathbf{u_1}',\alpha_{2n}\mathbf{u_2}')'\}>L_n(\boldsymbol{\Theta}_0)\right]\ge 1-\varepsilon,$$

where $\mathbf{u_1}$ has the same dimension as $\boldsymbol{\Theta}_1$ and $\mathbf{u_2}$ has the same dimension as $\mathbf{c}$. This implies with a probability of at least $1-\varepsilon$, that there exists a local minimum in the ball $\{\boldsymbol{\Theta}_0+(\alpha_{1n}\mathbf{u_1}',\alpha_{2n}\mathbf{u_2}')':\|\mathbf{u_1}\|\le\tau_1,\|\mathbf{u_2}\|\le\tau_2\}$. Hence, there exists a local minimum $(\boldsymbol{\Theta}_1',\mathbf{c}')'$ such that $\|\boldsymbol{\Theta}_1-\boldsymbol{\Theta}_{10}\|=O_p(\alpha_{1n})$, $\|\mathbf{c}-\mathbf{c}_0\|=O_p(\alpha_{2n})$.

With the definition $\boldsymbol{\Theta}^*=\boldsymbol{\Theta}_0+(\alpha_{1n}\mathbf{u_1}',\alpha_{2n}\mathbf{u_2}')'=(\mu^*,\boldsymbol{\alpha}_1^{*\prime},\boldsymbol{\alpha}_2^{*\prime},\boldsymbol{\alpha}_3^{*\prime},\mathbf{c}^{*\prime})',\mathbf{u}=(\mathbf{u}_1',\mathbf{u}_2')'$, we have

$$\begin{aligned}D_n(\mathbf{u})&=L_n(\boldsymbol{\Theta}^*)-L_n(\boldsymbol{\Theta}_0)\\&=l_n(\boldsymbol{\Theta}^*)-l_n(\boldsymbol{\Theta}_0)+\lambda\left\{J(\boldsymbol{\alpha}_1^*,\boldsymbol{\alpha}_2^*,\boldsymbol{\alpha}_3^*)-J(\boldsymbol{\alpha}_{10},\boldsymbol{\alpha}_{20},\boldsymbol{\alpha}_{30})\right\}.\end{aligned}$$

Since $\partial J(\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}=\mathbf{A}\boldsymbol{\alpha}$, by the standard argument on Taylor expansion of the likelihood function, we have

$$\begin{aligned}D_n(\mathbf{u})=&\,\alpha_{1n}\mathbf{u_1}'\frac{\partial l_n(\boldsymbol{\Theta}_0)}{\partial\boldsymbol{\Theta}_1}+\alpha_{2n}\mathbf{u_2}'\frac{\partial l_n(\boldsymbol{\Theta}_0)}{\partial\mathbf{c}}+\frac{1}{2}\alpha_{1n}^2\mathbf{u_1}'\frac{\partial^2 l_n(\boldsymbol{\Theta}_0)}{\partial\boldsymbol{\Theta}_1\partial\boldsymbol{\Theta}_1'}\mathbf{u_1}(1+o_p(1))\\&+\frac{1}{2}\alpha_{2n}^2\mathbf{u_2}'\frac{\partial^2 l_n(\boldsymbol{\Theta}_0)}{\partial\mathbf{c}\partial\mathbf{c}'}\mathbf{u_2}(1+o_p(1))+\alpha_{1n}\alpha_{2n}\mathbf{u_1}'\frac{\partial^2 l_n(\boldsymbol{\Theta}_0)}{\partial\boldsymbol{\Theta}_1\partial\mathbf{c}'}\mathbf{u_2}(1+o_p(1))\\&+\lambda\left\{(\boldsymbol{\alpha}^*-\boldsymbol{\alpha}_0)'\mathbf{A}\boldsymbol{\alpha}_0+(\boldsymbol{\alpha}^*-\boldsymbol{\alpha}_0)'\mathbf{A}(\boldsymbol{\alpha}^*-\boldsymbol{\alpha}_0)\right\}\end{aligned}$$

$$\equiv I_1 + I_2 + I_3 + I_4 + I_5 + I_6,$$

where $\boldsymbol{\alpha}^* = (\boldsymbol{\alpha}_1^{*\prime}, \boldsymbol{\alpha}_2^{*\prime}, \boldsymbol{\alpha}_3^{*\prime})'$. Denote $\Delta_{tjk}(\boldsymbol{\Theta}) \equiv \Phi\{(X_{tj} - c_k)/h\} - I(X_{tj} > c_k)$. By (B.3), we have

$$\begin{aligned}
\left| E\frac{\partial l_n(\boldsymbol{\Theta}_0)}{\partial \mu} \right| &= \left| \frac{2}{n} E \sum_{t=1}^{n} \sum_{r=1}^{2} \sum_{\ell=1}^{p} \alpha_{r+1,0}(\ell) X_{t\ell} V_{t\ell}^{r-1} \Delta_{t\ell r}(\boldsymbol{\Theta}_0) \right| = o(h^s), \\
\left| E\frac{\partial l_n(\boldsymbol{\Theta}_0)}{\partial \alpha_k(j)} \right| &= \left| \frac{2}{n} E \sum_{t=1}^{n} \sum_{r=1}^{2} \sum_{\ell=1}^{p} \alpha_{r+1,0}(\ell) X_{t\ell} V_{t\ell}^{r-1} \Delta_{t\ell r}(\boldsymbol{\Theta}_0) \Upsilon_{tjk}(\boldsymbol{\Theta}_0) \right| = o(h^s), \\
\left| E\frac{\partial l_n(\boldsymbol{\Theta}_0)}{\partial c_k} \right| &= \left| \frac{2}{nh} E \sum_{t=1}^{n} C_{tk}(\boldsymbol{\Theta}_0) \sum_{r=1}^{2} \sum_{\ell=1}^{p} \alpha_{r+1,0}(\ell) X_{t\ell} V_{t\ell}^{r-1} \Delta_{t\ell r}(\boldsymbol{\Theta}_0) \right| = o(h^s),
\end{aligned} \tag{C.1}$$

for $k = 1, 2, 3, j = 1, \ldots, p$, where $\alpha_{r,0}(\ell)$ is the true value of $\alpha_r(\ell)$.

Denote $A_{kj} \equiv E\sigma^2(\boldsymbol{V}_t, \boldsymbol{X}_t)\Upsilon^2_{tjk,0}(\boldsymbol{\Theta}_0)$. Furthermore, similar to Lemma 1, we have

$$\begin{aligned}
\Gamma_1 &\hat{=} \left[\text{var}\left\{ \frac{\partial l_n(\Theta_0)}{\partial \alpha_1(1)} \right\}, \ldots, \text{var}\left\{ \frac{\partial l_n(\Theta_0)}{\partial \alpha_1(p)} \right\}, \text{var}\left\{ \frac{\partial l_n(\Theta_0)}{\partial \alpha_2(1)} \right\}, \ldots, \text{var}\left\{ \frac{\partial l_n(\Theta_0)}{\partial \alpha_3(p)} \right\} \right]' \\
&= \frac{4}{n} \cdot (A_{11}, \ldots, A_{1p}, \ldots, A_{3p})' + o(h^s), \\
\Gamma_2 &\hat{=} \text{var}\left\{ \frac{\partial l_n(\Theta_0)}{\partial \mu} \right\} = \frac{4}{n} \cdot E\sigma^2(\boldsymbol{V}_t, \boldsymbol{X}_t) + o(h^s), \\
\Gamma_3 &\hat{=} \left[\text{var}\left\{ \frac{\partial l_n(\Theta_0)}{\partial c_1} \right\}, \text{var}\left\{ \frac{\partial l_n(\Theta_0)}{\partial c_2} \right\} \right]' = \frac{1}{nh}\{b_1(\Theta_0), b_2(\Theta_0)\}' + O\left(\frac{1}{n}\right).
\end{aligned} \tag{C.2}$$

Combining (C.1) and (C.2), we have

$$\begin{aligned}
\frac{\partial l_n(\Theta_0)}{\partial \boldsymbol{\alpha}} &= E\left\{ \frac{\partial l_n(\Theta_0)}{\partial \boldsymbol{\alpha}} \right\} + O_p(\Gamma_1^{1/2}) = O_p\left(\frac{1}{\sqrt{n}} \right), \\
\frac{\partial l_n(\Theta_0)}{\partial \mu} &= E\left\{ \frac{\partial l_n(\Theta_0)}{\partial \mu} \right\} + O_p(\Gamma_2^{1/2}) = O_p\left(\frac{1}{\sqrt{n}} \right), \\
\frac{\partial l_n(\Theta_0)}{\partial \mathbf{c}} &= E\left\{ \frac{\partial l_n(\Theta_0)}{\partial \mathbf{c}} \right\} + O_p(\Gamma_3^{1/2}) = O_p\left(\frac{1}{\sqrt{nh}} \right).
\end{aligned}$$

then

$$I_1 = O_p\left(\frac{\alpha_{1n}\tau_1}{\sqrt{n}} \right), I_2 = O_p\left(\frac{\alpha_{2n}\tau_2}{\sqrt{nh}} \right). \tag{C.3}$$

Similar to Lemma 1, we also obtain

$$\begin{aligned}
E\frac{\partial^2 l_n(\boldsymbol{\Theta}_0)}{\partial\mu^2} &= Var\bigg(\frac{\partial^2 l_n(\boldsymbol{\Theta}_0)}{\partial\mu^2}\bigg) = 2,\\
E\frac{\partial^2 l_n(\boldsymbol{\Theta}_0)}{\partial\mu\partial\alpha_k(j)} &= 2E\Upsilon_{tjk,0}(\boldsymbol{\Theta}_0) + o(h^s),\\
Var\frac{\partial^2 l_n(\boldsymbol{\Theta}_0)}{\partial\mu\partial\alpha_k(j)} &= \frac{4}{n}\big[E\Upsilon^2_{tjk,0}(\boldsymbol{\Theta}_0) - E^2\Upsilon_{tjk,0}(\boldsymbol{\Theta}_0)\big] + o\bigg(\frac{1}{n}\bigg),\\
E\bigg(\frac{\partial^2 l_n(\boldsymbol{\Theta}_0)}{\partial\alpha_k(j)\partial\alpha_m(v)}\bigg) &= 2E\Upsilon_{tjk,0}(\Theta_0)\Upsilon_{tvm,0}(\Theta_0) + o(h^s),\\
Var\bigg(\frac{\partial^2 l_n(\boldsymbol{\Theta}_0)}{\partial\alpha_k(j)\partial\alpha_m(v)}\bigg) &= \frac{4}{n}\big[E\Upsilon^2_{tjk,0}(\boldsymbol{\Theta}_0)\Upsilon^2_{tvm,0}(\boldsymbol{\Theta}_0)\\
&\quad -\{E\Upsilon_{tjk,0}(\boldsymbol{\Theta}_0)\Upsilon_{tvm,0}(\boldsymbol{\Theta}_0)\}^2\big] + o\bigg(\frac{1}{n}\bigg).
\end{aligned}$$

Hence,

$$I_3 = O_p(\alpha^2_{1n}\tau^2_1). \tag{C.4}$$

Similarly, we have

$$\begin{aligned}
E\frac{\partial^2 l_n(\boldsymbol{\Theta}_0)}{\partial c_k^2} &= \frac{F_k(\boldsymbol{\Theta}_0)}{h} + O(1),\ E\bigg\{\frac{\partial^2 l_n(\boldsymbol{\Theta}_0)}{\partial c_k^2}\bigg\}^2 = \frac{F_k^2(\boldsymbol{\Theta}_0)}{h^2} + O(1+(nh^2)^{-1}),\\
E\frac{\partial^2 l_n(\boldsymbol{\Theta}_0)}{\partial c_1\partial c_2} &= F(\boldsymbol{\Theta}_0) + O\bigg(\frac{1}{h}\exp\bigg\{-\frac{(c_{1,0}-c_{2,0})^2}{4h^2}\bigg\}\bigg),\\
E\bigg\{\frac{\partial^2 l_n(\boldsymbol{\Theta}_0)}{\partial c_1\partial c_2}\bigg\}^2 &= F^2(\boldsymbol{\Theta}_0) + O\bigg(\exp\bigg\{-\frac{(c_{1,0}-c_{2,0})^2}{2h^2}\bigg\}\bigg),
\end{aligned}$$

then we obtain,

$$I_4 = O_p\bigg(\frac{\alpha^2_{2n}\tau^2_2}{h}\bigg). \tag{C.5}$$

Finally, by

$$\begin{aligned}
E\frac{\partial^2 l_n(\boldsymbol{\Theta}_0)}{\partial\alpha_k(j)\partial c_l} &= \kappa_{kjl} + O(h),\ E\bigg\{\frac{\partial^2 l_n(\boldsymbol{\Theta}_0)}{\partial\alpha_k(j)\partial c_l}\bigg\}^2 = \kappa^2_{kjl} + O(h^2),\\
E\frac{\partial^2 l_n(\boldsymbol{\Theta}_0)}{\partial\mu\partial c_l} &= \varphi_l + O(h),\ E\bigg\{\frac{\partial^2 l_n(\boldsymbol{\Theta}_0)}{\partial\mu\partial c_l}\bigg\}^2 = \varphi_l^2 + O(h^2),
\end{aligned}$$

we get

$$I_5 = O_p(\alpha_{1n}\alpha_{2n}\tau_1\tau_2). \tag{C.6}$$

By (C.3),(C.4),(C.5), (C.6) and coupling with $I_6 = O_p(\lambda\alpha_{1n}\tau_1 + \lambda\alpha_{1n}^2\tau_1^2)$, choosing large τ_1, τ_2, then I_1,I_2, I_5 are dominated by I_3, I_4. This completes the proof of Theorem 1.

Proof of Theorem 2.

According to Theorem 1, when $\lambda = o(1/\sqrt{n})$, it can easily be shown that there exists a $\sqrt{n}$-consistent estimator $\hat{\boldsymbol{\Theta}}_1 = (\hat{\mu}, \hat{\boldsymbol{\alpha}}')'$ and $\sqrt{n/h}$-consistent estimator $\hat{\mathbf{c}}$, satisfying the following equations

$$\partial l_n(\hat{\boldsymbol{\Theta}})\partial\boldsymbol{\Theta}_1 + \lambda\hat{\mathbf{b}} = 0,$$
$$\frac{\partial l_n(\hat{\boldsymbol{\Theta}})}{\partial\mathbf{c}} = 0,$$

where $\hat{\mathbf{b}} = (0, \widehat{\boldsymbol{\alpha}}'A')'$ is a vector of $3p+1$ dimension. By Taylor expansion,

$$\begin{aligned}
-\sqrt{n}\frac{\partial l_n(\boldsymbol{\Theta}_0)}{\partial\boldsymbol{\Theta}_1} &= \frac{\partial^2 l_n(\boldsymbol{\Theta}_0)}{\partial\boldsymbol{\Theta}_1\partial\boldsymbol{\Theta}_1'}\sqrt{n}(\hat{\boldsymbol{\Theta}}_1 - \boldsymbol{\Theta}_{10})(1+o_p(1)) \\
&\quad + \sqrt{h}\frac{\partial^2 l_n(\boldsymbol{\Theta}_0)}{\partial\boldsymbol{\Theta}_1\partial\mathbf{c}'}\sqrt{\frac{n}{h}}(\hat{\mathbf{c}} - \mathbf{c}_0)(1+o_p(1)) + \sqrt{n}\lambda\hat{\mathbf{b}}, \\
-\sqrt{nh}\frac{\partial l_n(\boldsymbol{\Theta}_0)}{\partial\mathbf{c}} &= \sqrt{h}\frac{\partial^2 l_n(\boldsymbol{\Theta}_0)}{\partial\mathbf{c}\partial\boldsymbol{\Theta}_1'}\sqrt{n}(\hat{\boldsymbol{\Theta}}_1 - \boldsymbol{\Theta}_{10})(1+o_p(1)) \\
&\quad + h\frac{\partial^2 l_n(\boldsymbol{\Theta}_0)}{\partial\mathbf{c}\partial\mathbf{c}'}\sqrt{\frac{n}{h}}(\hat{\mathbf{c}} - \mathbf{c}_0)(1+o_p(1)). \qquad (C.7)
\end{aligned}$$

Since

$$\begin{aligned}
E\left(\frac{\partial l_n(\boldsymbol{\Theta}_0)}{\partial c_k}\right)^2 &= \frac{1}{nh}b_k(\boldsymbol{\Theta}_0) + O\left(\frac{1}{n}\right), \\
E\frac{\partial l_n(\boldsymbol{\Theta}_0)}{\partial c_1}\frac{\partial l_n(\boldsymbol{\Theta}_0)}{\partial c_2} &= \frac{1}{n}\zeta_1 + O(h) \\
E\frac{\partial l_n(\boldsymbol{\Theta}_0)}{\partial \alpha_k(j)}\frac{\partial l_n(\boldsymbol{\Theta}_0)}{\partial \alpha_m(v)} &= \frac{1}{n}\delta_{kjmv} + o\left(\frac{h^s}{n}\right), \\
E\left(\frac{\partial l_n(\boldsymbol{\Theta}_0)}{\partial \mu}\right)^2 &= \frac{1}{n}\zeta_2 + o\left(\frac{h^s}{n}\right), \\
E\frac{\partial l_n(\boldsymbol{\Theta}_0)}{\partial \mu}\frac{\partial l_n(\boldsymbol{\Theta}_0)}{\partial \alpha_k(j)} &= \frac{1}{n}\varpi_{kj} + o\left(\frac{h^s}{n}\right),
\end{aligned}$$

we get $nh(\partial l_n(\mathbf{\Theta}_0)/\partial \mathbf{c})^{\otimes 2} \to_p \Sigma_1$ and $n(\partial l_n(\mathbf{\Theta}_0)/\partial \mathbf{\Theta}_1)^{\otimes 2} \to_p \Sigma_2$. Then by CLT we have

$$\sqrt{nh}\frac{\partial l_n(\mathbf{\Theta}_0)}{\partial \mathbf{c}} \to N(0,\Sigma_1),\ \sqrt{n}\frac{\partial l_n(\mathbf{\Theta}_0)}{\partial \mathbf{\Theta}_1} \to N(0,\Sigma_2). \tag{C.8}$$

Furthermore, from the proof of Theorem 1, we know that $\partial^2 l_n(\mathbf{\Theta}_0)/(\partial \mathbf{\Theta}_1 \partial \mathbf{\Theta}_1') \to_p V_{11}$, $\partial^2 l_n(\mathbf{\Theta}_0)/\partial \mathbf{\Theta}_1 \partial \mathbf{c}' \to_p V_{12}$, $h(\partial^2 l_n(\mathbf{\Theta}_0))/(\partial \mathbf{c} \partial \mathbf{c}') \to_p V_{22}$. Based on all these results coupled with (C.7),(C.8) and Slutsky's theorem, we obtain

$$\begin{aligned}\sqrt{n}(\hat{\mathbf{\Theta}}_1 - \mathbf{\Theta}_{10} + \lambda V_{11}^{-1}\mathbf{b}) &\to N(0, V_{11}^{-1}\Sigma_2 V_{11}^{-1}),\\ \sqrt{\frac{n}{h}}(\hat{\mathbf{c}} - \mathbf{c}_0 - h\lambda V_{22}^{-1}V_{12}'V_{11}^{-1}\mathbf{b}) &\to N(0, V_{22}^{-1}\Sigma_1 V_{22}^{-1}),\end{aligned}$$

where $\mathbf{b}$ is defined in Appendix B. We complete the proof of Theorem 2.

References

Brown, B. M. and Wang, Y. G. (2005). Standard errors and covariance matrices for smoothed rank estimators. *Biometrika* **92**, 149–158.

Cai, Z., Fan, J. and Yao, Q. (2000). Functional-coefficient regression models for nonlinear time series models. *Journal of the American Statistical Association* **95**, 941–956.

Caner, M. and Hansen, B. E. (2001). Threshold autoregression with a unit root. *Econometrica* **69**, 1555–1596.

Caner, M. and Hansen, B. E. (2008). Instrumental variable estimation of a threshold model. *Econometric Theory* **20**, 813–843.

Chan, K. S. (1993). Consistency and limiting distribution of the least squares estimator of a threshold autoregressive model. *The Annals of Statistics* **21**, 520–533.

Chang, S. S., Kwok, S. S. M., Cheng, Q., Yip, P. S. F. and Chen, Y. Y. (2015). The association of trends in charcoal-burning suicide with Google search and newspaper reporting in Taiwan: a time series analysis. *Social Psychiatry and Psychiatric Epidemiology* **50**, 1451-1461.

Chen, Y.-Y., Chen, F., Gunnell, D. and Yip, P. S. F., (2013). The impact of media reporting on the emergence of charcoal burning suicide in Taiwan. *PloS One* **8**.

Chen, Y. Y., Chen, F. and Yip, P. S. F. (2011). The impact of media reporting of suicide on actual suicides in Taiwan, 2002-05. *Journal of Epidemiology and Community Health* **65**, 934–940.

Chen, H., Chong, T. L. and Bai, J. (2012). Theory and applications of tar model with two threshold varialbles. *Econometric Reviews* **31**, 142–170.

Cheng, A. T., Hawton, K., Chen, T. H., Yen, A. M., Chen, C. Y., Chen, L. C. and Teng, P. R. (2007). The influence of media coverage of a celebrity suicide on subsequent suicide attempts. *The Journal of Clinical Psychiatry* **68**, 862–866.

Cheng, Q., Chang, S. S. and Yip, P. S. F. (2012). The opportunity and challenges of online data collection for suicide prevention. *Lancet* **379**, 53–54.

Chen, C. W. S. and So, M. K. P. (2006). On a threshold teteroscedastic model. *International*

Journal of Forecasting **22**, 73–89.

Delgadoa, M. A. and Hidalgo, J. (2000). Nonparametric inference on structural breaks. *Journal of Econometrics* **96**, 113–144.

Fan, J. and Yao, Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods.* Springer-Verlag, New York.

Faraggi, D. and Simon, R. (1996). A simulation study of cross-validation for selecting an optimal cutpoint in univariate survival analysis. *Statistics in Medicine* **15**, 2203–2213.

Fu, K. W. and Yip, P. S. F. (2007). Long-term impact of celebrity suicide on suicidal ideation: Results from a population-based study. *Journal of Epidemiology and Community Health* **61**, 540–546.

Koop, G. and Potter, S. M. (1999). Dynamic ssymmetries in U.S. unemployment. *Journal of Business and Economic Statistics* **17**, 298–312.

Guo, J., Hu, J., Jing, B. and Zhang, Z. (2015). Spline-Lasso in high-dimensional linear regression. *Journal of the American Statistical Association* **111**, 288–297.

Han, A. K. (1987). Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator. *Journal of Econometrics* **35**, 303–316.

Hanse, B. E. (1999). Threshold effects in non-dynamic panels: estimation, testing, and inference. *Journal of Econometrics* **93**, 345–368.

Hanse, B. E. (2000). Sample splitting and threshold estimation. *Econometrica* **68**, 575–603.

Law, Y. W., Yip, P. S., Zhang, Y. and Caine, E. D. (2014). The chronic impact of work on suicides and under-utilization of psychiatric and psychosocial services. *Journal of Affective Disorders* **168**, 254–261.

Li, D. and Ling, S. (2012). On the least squares estimation of multiple-regime threshold autoregressive models. *Journal of Econometrics* **167**, 240–253.

Lin, H. Z., Yip, P. S. F. and Huggins, R. M. (2011). A nonparametric estimation of the infection curve. *Science China Mathematics* **54**, 1815.

Ma, S. and Huang, J. (2007). Combining multiple markers for classification using ROC. *Biometrics* **63**, 751–757.

Ni, S., Xia, Q. and Liu, J. (2018). Bayesian subset selection for two-threshold variable autoregressive models. *Studies in Nonlinear Dynamics and Econometrics.*

Niederkrotenthaler, T., Fu, K. W., Yip, P. S., Fong, D. Y., Stack, S., Cheng, Q. and Pirkis, J. (2012). Changes in suicide rates following media reports on celebrity suicide: A meta-analysis. *Journal of Epidemiology and Community Health* **66**, 1037–1042.

Niederkrotenthaler, T. and Stack, S. (2017). *Media and Suicide: International Perspectives on Research, Theory, and Policy.* Routledge, New York.

Phillips, D. (1974). The influence of suggestion on suicide: Substantive and theoretical implications of the werther effect. *American Sociological Review* **39**, 340–354.

Pirkis, J. and Blood, R. W. (2001). Suicide and the media: Part 1. Reportage in nonfictional media. *Crisis-The Journal of Crisis Intervention and Suicide Prevention* **22**, 146–154.

Qian, L. (1998). On maximum likelihood estimators for a threshold autoregression. *Journal of Statistical Planning and Inference* **75**, 21–46.

Seo, M. H. and Linton, O. (2007). A smoothed least squares estimator for threshold regression models. *Journal of Econometrics* **141**, 704–735.

Sherman, R. P. (1993). The limiting distribution of the maximum rank correlation estimator.

Econometrica **61**, 123–137.

Tsay, R. (1998). Testing and modeling multivariate threshold models. *Journal of the American Statistical Association* **93**, 1188–1202.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005). Sparsity and smoothness via the fused Lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**, 91–108.

World Health Organization (2014). *World Suicide Report*. World Health Organization, Geneva.

Wu, S. and Chen, R. (2007). Threshold variable determination and threshold variable driven switching autoregressive models. *Statistica Sinica* **17**, 241–264.

Yip, P. S., Fu, K. W., Yang, K. C., Ip, B. Y., Chan, C. L., Chen, E. Y., Lee, D. T., Law, Y. W. and Hawton, K. (2006). The effects of a celebrity suicide on suicide rates in Hong Kong. *Journal of Affective Disorders* **93**, 245–252.

Huazhen Lin

Center of Statistical Research, 555 Liutai Avenue, Wenjiang District, Chengdu, Sichuan 611130, P.R. China.

E-mail: linhz@swufe.edu.cn

Jiakun Jiang

Weiqing Building, Room 114, Tsinghua University, Beijing 100091, P.R. China.

E-mail: jiakunj@mail.tsinghua.edu.cn

Binhuan Wang

180 Madison Avenue, Suite 450, New York, NY 10016, USA.

E-mail: Binhuan.wang@nyulangone.org

Paul S. F. Yip

CJT-534, 5/F, The Jockey Club Tower, The Centennial Campus, HKU, HK.

E-mail: sfpyip@hku.hk

(Received March 2018; accepted April 2019)