Electronic Journal of Statistics
Vol. 13 (2019) 2043–2092
ISSN: 1935-7524
https://doi.org/10.1214/19-EJS1571

Generalized threshold latent variable model

Yuanbo Li

University of International Business and Economics,
e-mail: jacoblyb@gmail.com

Xunze Zheng and Chun Yip Yau

Chinese University of Hong Kong,
e-mail: s1155002000@sta.cuhk.edu.hk; cyyau@sta.cuhk.edu.hk

Abstract: This article proposes a generalized threshold latent variable model for flexible threshold modeling of time series. The proposed model encompasses several existing models, and allows a discrete valued threshold variable. Sufficient conditions for stationarity and ergodicity are investigated. The minimum description length principle is applied to formulate a criterion function for parameter estimation and model selection. A computationally efficient procedure for optimizing the criterion function is developed based on a genetic algorithm. Consistency and weak convergence of the parameter estimates are established. Moreover, simulation studies and an application for initial public offering data are presented to illustrate the proposed methodology.

MSC 2010 subject classifications: 62M10.
Keywords and phrases: Compound Poisson process, ergodicity, genetic algorithm, minimum description length principle, multiple-threshold, piecewise modeling.

Received April 2018.

Contents

1. Introduction

The threshold autoregressive (TAR) model, proposed by [58], has enormous popularity in a wide range of applications. It allows the modeling of diverse behaviors under different regimes, which provides flexible descriptions of many real-world scenarios. The backgrounds, theory, applications and extensions of TAR models can be found in the excellent surveys of [59, 61].

Recently, generalizations to the TAR model have been considered by introducing threshold structures to non-linear models instead of autoregressive models. One important direction is on generalized linear models with thresholds, such as the generalized threshold mixed model (GTMM) ([53]) and the generalized threshold stochastic regression model (GTSRM) ([54]). In addition, threshold modeling has been extended to heteroscedasticity of time series, for example, the double-threshold autoregressive moving average conditional heteroskedastic (DTARMACH) model ([38]), the threshold stochastic volatility (THSV) model ([56]), the multiple-threshold double autoregressive (MTDAR) model ([33]), and the threshold double autoregressive model (TDAR) ([34]). Furthermore, threshold models with more elastic regime switching mechanisms are considered in the endogenous delay threshold model (EDTAR) of [27] and the hysteretic autoregressive (HAR) model of [36].

For the asymptotic theory of TAR models, the strong consistency and asymptotic distributions of the parameter estimates are studied by [10] and [13]. On the other hand, tests for a linear series against its threshold extension are considered in [12], [63] and [35]. For many sophisticated threshold models, asymptotic theories are developed by assuming stationarity and ergodicity of the process (for example, [54]). However, conditions for strict stationarity and ergodicity are investigated only for the self-excited threshold autoregressive (SETAR) model ([11]), the TAR model with order p ([3]), the DTARMACH model in [38], the HAR model in [36], the MTDAR model in [33], and the TDAR model in [34]. Corresponding results for the generalized linear-type threshold models remain unexplored.

More importantly, despite the well developed theoretical background of estimation theory, the estimation procedure of threshold models demands a high computational cost. Due to the irregular nature of the threshold parameters, locating the global minimum of the likelihood requires a multi-parameter grid search over all possible values of the threshold parameters, which is typically computationally infeasible; see [32] and [65]. Consequently, many threshold models are implemented assuming one threshold a priori; see [56], [53], and [54]. When the number of thresholds and the parametric model of each regime are unknown, no practical estimation method appears available except for the simplest TAR model; see [66], [15] and [16].

To tackle the above problems and further extend the flexibility, in this article we first propose the generalized threshold latent variable model (GTLVM) which covers most of the aforementioned models as special cases. In particular, the threshold variables may be continuous or discrete valued. As far as we know, asymptotic theory for threshold models with discrete-valued thresholds has not

been investigated in the literature. Second, we establish sufficient conditions for the stationarity and ergodicity of the proposed model. Third, and most importantly, we develop a computationally efficient estimation procedure using an information criterion derived from the minimal description length (MDL) principle, which substantially generalizes the procedure of [66] for TAR model. The procedure allows not only fast estimation of the number and the location of thresholds, but also model selection in each regime.

This paper is organized as follows. Section 2 defines the GTLVM. Section 3 establishes sufficient conditions for strict stationarity and ergodicity of GTLVM. Section 4 proposes a genetic algorithm using MDL principle for parameter estimation. Section 5 establishes asymptotic consistency and convergence rates of parameter estimations. Section 6 provides two simulation studies to illustrate the effectiveness of the proposed estimation method. Section 7 presents an application of modeling IPO volumes in U.S. stock market. The proof of Theorems are provided in the appendix.

2. Generalized threshold latent variable model

Consider a time series $\{y_t\}_{t=1,\ldots,n}$ in which y_t depends on its past observations and a latent variable λ_t, and follows different models when a threshold variable z_{t-d} belongs to different regimes. The latent variable λ_t is not only associated with past observations of y_t but also exogenous covariates X_t, and has a regime switching structure. Specifically, the conditional density function of y_t satisfies

$$f(y_t|Y_{t-1}, X_t, z_{t-d}) = \sum_{i=1}^{r+1} \left[\int f_i(y_t|Y_{t-1}, \lambda_t) h_i(\lambda_t|Y_{t-1}, X_t) d\lambda_t \right] \times I(z_{t-d} \in (\theta_{i-1}, \theta_i]), \tag{2.1}$$

where $-\infty = \theta_0 < \theta_1 < \cdots < \theta_r < \theta_{r+1} = \infty$ are the thresholds that classify y_t into $r+1$ regimes based on the threshold variable z_{t-d} and *threshold delay parameter* d, $Y_{t-1} = \{y_{t-1}, \ldots, y_{t-p_Y}\}$ are the past observations, and $X_t = \{x_{t,1}, \ldots, x_{t,p_X}\}$ are the covariates. Moreover, f_i and h_i are conditional densities of y_t and λ_t of regime i, respectively. We assume that z_{t-d} is measurable with respect to the sigma-field generated by $\{X_{t-d}, y_{t-d}, X_{t-d-1}, y_{t-d-1}, \ldots\}$. Typical examples include $z_{t-d} = y_{t-d}$ for self-excited threshold models, and $z_{k-d} = x_{k-d,k}$ for some exogenous covariates $x_{t-d,k}$.

To explicitly describe the dependence of y_t on λ_t, Y_{t-1}, X_t and z_{t-d}, the GTLVM can be specified as

$$\begin{cases} y_t = \sum_{i=1}^{r+1} \left[\omega_i + \sum_{l=1}^{p_{Y,1}} \psi_{y,1,i}^{(l)} g_{y,1}(y_{t-l}) + g_\lambda(\lambda_t, u_t) + \sum_{m=0}^{p_e} \psi_{e,i}^{(m)} e_{t-m} \right] \\ \qquad \times I(z_{t-d} \in (\theta_{i-1}, \theta_i]), \\ \phi(\lambda_t) = \sum_{i=1}^{r+1} \left[\alpha_i + \sum_{j=1}^{p_{Y,2}} \psi_{y,2,i}^{(j)} g_{y,2}(y_{t-j}) + \sum_{k=1}^{p_X} \psi_{x,i}^{(k)} x_{t,k} + \sum_{q=0}^{p_\epsilon} \psi_{\epsilon,i}^{(q)} \epsilon_{t-q} \right] \\ \qquad \times I(z_{t-d} \in (\theta_{i-1}, \theta_i]), \end{cases} \tag{2.2}$$

where $\phi(\cdot)$ is the link function which is smooth and monotonic increasing; $\{\epsilon_t\}$ and $\{e_t\}$ are independent and identical (i.i.d.) mean-zero errors; $g_{y,1}(\cdot)$ and $g_{y,2}(\cdot)$ are known continuous functions; $g_\lambda(\lambda_t, u_t)$ is an inverse cumulative distribution function with parameter λ_t, and $\{u_t\}$ are i.i.d. uniform random variables on $[0,1]$. Without loss of generality, we may assume $g_{y,1}(0) = g_{y,2}(0) = 0$ by adjusting ω_i and α_i accordingly. Note that (2.2) is a special case of (2.1): In (2.2), for any regime i, the conditional density of y_t given λ_t, denoted as $f_i(y_t|Y_{t-1}, \lambda_t)$, can be determined by the distributions of u_t and e_t. Also, the conditional density of λ_t given Y_{t-1}, X_t, denoted as $h_i(\lambda_t|Y_{t-1}, X_t)$, can be found from the second equation of (2.2). Integrating out the effect of λ_t in $f_i(y_t|Y_{t-1}, \lambda_t)$ with respect to h_i, (2.1) follows. Although (2.1) is slightly more general than (2.2), we focus our attention on (2.2) for convenience in parametric modeling.

By properly choosing λ_t, $g_\lambda(\lambda_t, u_t)$, $\phi(\lambda_t)$ and z_{t-d}, model (2.2) covers a number of the aforementioned threshold models in the literature. For instance, if $g_\lambda(\lambda_t, u_t) = 0$ and $g_{y,1}$ is the identity function, then (2.2) reduces to threshold autoregressive and moving average (TARMA) model

$$y_t = \sum_{i=1}^{r+1} \left[\omega_i + \sum_{l=1}^{p_{Y,1}} \psi_{y,1,i}^{(l)} y_{t-l} + \sum_{m=0}^{p_e} \psi_{e,i}^{(m)} e_{t-m} \right] I(z_{t-d} \in (\theta_{i-1}, \theta_i]), \tag{2.3}$$

as in [31]. If, in addition, $p_e = 0$, then (2.3) reduces to threshold autoregressive (TAR) model as in [58]. If $g_\lambda(\lambda_t, u_t) = g_\lambda(u_t)$ is a quantile function with $p_e = 0$, then (2.2) reduces to the quantile self-excited threshold autoregressive (QSETAR) model in [8]. Next, denote $\lambda_t = \mathrm{E}(y_t)$ and let $g_\lambda(\lambda_t, u_t) = F^{-1}(\lambda_t, u_t)$ be the inverse of the cumulative distribution function of some exponential family distribution with probability density

$$f(y_t; \lambda_t, a_t, \nu) = \exp\left[\frac{1}{\nu a_t} \{ y_t \eta(\lambda_t) - b(\lambda_t) \} + c(y_t; \nu a_t) \right],$$

where $\eta(\lambda_t)$ is the canonical parameter, ν is an overdispersion parameter and a_t is a user-specified weight. Then, the special case of (2.2) given by

$$\begin{cases} y_t = g_\lambda(\lambda_t, u_t), \\ \phi(\lambda_t) = \sum_{i=1}^{r+1} \left[\alpha_i + \sum_{j=1}^{p_Y} \psi_{y,i}^{(j)} g_y(y_{t-j}) + \sum_{k=1}^{p_X} \psi_{x,i}^{(k)} x_{t,k} + \psi_{\epsilon,i} \epsilon_t \right] \\ \qquad \times I(z_{t-d} \in (\theta_{i-1}, \theta_i]), \end{cases} \tag{2.4}$$

reduces to the GTMM in [53] when $\phi(\lambda_t) = 0$ in the first or last regime, and reduces to the GTSRM in [54] when $\psi_{\epsilon,j} = 0$ for all j. In addition, (2.2) covers models with double threshold structure and conditional heteroskedasticity. For example, with $\lambda_t = \sigma_t$ and $\phi(\sigma_t) = \sigma_t^2$,

$$\begin{cases} y_t = \sum_{i=1}^{r+1} \left[\omega_i + \sum_{l=1}^{p} \psi_{y,1,i}^{(l)} y_{t-l} + \sigma_t \epsilon_t \right] I(y_{t-d} \in (\theta_{i-1}, \theta_i]), \\ \sigma_t^2 = \sum_{i=1}^{r+1} \left(\alpha_i + \sum_{j=1}^{p} \psi_{y,2,i}^{(j)} y_{t-j}^2 \right) I(y_{t-d} \in (\theta_{i-1}, \theta_i]), \end{cases} \tag{2.5}$$

is the MTDAR model ([33]). Similar arguments apply to the THSV model in [56].

Therefore, the proposed model (2.2) allows a unified treatment to the open problem of establishing the stationarity and ergodicity of many existing threshold-type models. Moreover, the algorithm developed in Section 4 provides an efficient solution to the computationally challenging problem of estimation and model selection for these models.

Denote the threshold parameters by $\Theta = (\theta_1, \ldots, \theta_r)$ and the model parameters for the ith regime by $\Psi_i = (\omega_i, \Psi_{Y,1,i}, \Psi_{e,i}, \alpha_i, \Psi_{Y,2,i}, \Psi_{X,i}, \Psi_{\epsilon,i})$, where $\Psi_{Y,1,i} = (\psi_{y,1,i}^{(1)}, \ldots, \psi_{y,1,i}^{(p_{Y,1})})$, $\Psi_{e,i} = (\psi_{e,i}^{(0)}, \ldots, \psi_{e,i}^{(p_e)})$, $\Psi_{Y,2,i} = (\psi_{y,2,i}^{(1)}, \ldots, \psi_{y,2,i}^{(p_{Y,2})})$, $\Psi_{X,i} = (\psi_{x,i}^{(1)}, \ldots, \psi_{x,i}^{(p_X)})$, $\Psi_{\epsilon,i} = (\psi_{\epsilon,i}^{(0)}, \ldots, \psi_{\epsilon,i}^{(p_\epsilon)})$. Combining the parameters in all regimes, we define $\Psi = (\Psi_1, ..., \Psi_{r+1})$. Note that we allow some of $\psi_{y,1,i}^{(l)}$, $\psi_{e,i}^{(m)}$, $\psi_{y,2,i}^{(j)}$, $\psi_{x,i}^{(k)}$ and $\psi_{\epsilon,i}^{(q)}$ equal to zero so that different autoregressive orders and covariates can be included in different regimes. Denote the model order parameter as $p = (p_1, \ldots, p_{r+1})$, where $p_i = (p_{Y,1,i}, p_{e,i}, p_{Y,2,i}, p_{X,i}, p_{\epsilon,i})$, $p_{Y,1,i}, p_{e,i}$, $p_{Y,2,i}$ and $p_{\epsilon,i}$ are integers, and $p_{X,i} = (p_{X,i}^{(1)}, \ldots, p_{X,i}^{(p_X)})$ is a binary vector indicating the nonzero entries of $\Psi_{X,i}$. Thus, in (2.2), $p_{Y,1} = \max_{i=1,\ldots,r+1} p_{Y,1,i}$, and $p_{Y,2}$, p_e and p_ϵ are defined similarly.

3. Stationarity and ergodicity

We state a set of sufficient conditions for the strict stationarity and ergodicity of the GTLVM (2.2) as follows:

Condition 1.

a) The covariate X_t is independent of $\{y_s\}_{s<t}$ and $\mathrm{E}|X_t| < \infty$. In addition, there exists a positive integer $\tilde{p}$ such that $\{(X_t, \ldots, X_{t-\tilde{p}+1})\}_{t=1,\ldots}$ is Markovian, strictly stationary and ergodic. Moreover, there exists a non-negative integer q such that z_t is measurable with respect to the sigma-field generated by $\{X_t, y_t, \ldots, X_{t-q}, y_{t-q}\}$.

b) The link function $\phi(\cdot)$ is smooth and strictly increasing, and is either concave or a polynomial of order $\gamma \geq 1$;

c) There exist constants $b_1 > 0$ and $\tilde{y} > 0$ such that $|g_{y,1}(y)| \leq b_1|y|$ for all $|y| > \tilde{y}$, and $g_{y,1}(y) \leq G_1$ for all $|y| \leq \tilde{y}$. In addition, $\phi(\cdot)$ and $g_{y,2}(\cdot)$ satisfy $\phi^{-1}(g_{y,2}(y)) \leq b_2|y|$ for some $b_2 > 0$ and all $|y| > \tilde{y}$, and $\phi^{-1}(g_{y,2}(y)) \leq G_2$ for some constant G_2 and all $|y| \leq \tilde{y}$;

d) $\mathrm{E}[|g_\lambda(\lambda, u)| \mid \lambda]$ is increasing in λ. Moreover, there exist some positive constants b_λ, $\tilde{\lambda}$ and H such that $\mathrm{E}[|g_\lambda(\lambda, u)| \mid \lambda] \leq b_\lambda|\lambda|$ if $|\lambda| > \tilde{\lambda}$, and $\mathrm{E}[|g_\lambda(\lambda, u)| \mid \lambda] \leq H$ if $|\lambda| \leq \tilde{\lambda}$;

e) $\mathrm{E}(|e_t|) < \infty$, $\mathrm{E}[\phi^{-1}(\epsilon_t)] < \infty$, and $\mathrm{E}[\phi^{-1}(x_{t,k})] < \infty$ for all $x_{t,k}$;

f) For the case $p_e > 0$, $\{y_t\}$ is an irreducible process.

Remark 1. Denote $\xi_t = (y_t, X_t, e_t, \epsilon_t)$ and $p^* = \max\{p_{Y,1}, p_{Y,2}, \tilde{p}, p_e, p_\epsilon, q + d, 1\}$. Condition 1a) ensures the Markovian property of $\Xi_t = (\xi_t, \ldots, \xi_{t-p^*+1})$, which helps to prove the strict stationarity and ergodicity of $\{y_t\}$. Condition

1b) is satisfied for the concave link from the exponential family distribution (for example, log-link in [54]) and the square-link from modeling conditional heteroskedasticity ([33]). Conditions 1c) and 1d) are regularity conditions that control the increasing rates to avoid explosive behaviors. For example, taking $\phi(\cdot)$ and $g_{y,2}(\cdot)$ as the log-transform in [54], 1c) and 1d) hold with $b_2 = b_\lambda = 1$. Condition 1e) regulates the tails of the noise and covariates. For example, if ϕ is the log link, then ϵ_t must have tails lighter than Laplace distribution with intensity 1. Similar conditions are found in [11], [38] and [33]. Finally, the irreducibility required in Condition 1f) is used to derive the geometric ergodicity of the process when $p_e > 0$, see [37] and [38]. For $p_e = 0$, the irreducibility is guaranteed by the following lemma.

Lemma 3.1. *Under Condition 1, if $p_e = 0$ in (2.2), then $\{y_t\}$ is irreducible.*

To establish stationarity and ergodicity of a threshold model, a condition for preventing the series from exploding is required. For example, $\max_i \sum_{l=1}^{p_{Y,1}} |\psi_{y,1,i}^{(j)}| < 1$ for TAR model in [3]. Under Condition 1, we have the following result for GTLVM, where the proof is provided in the Appendix.

Theorem 3.1. *Suppose that a process $\{y_t\}$ satisfies (2.2) and Condition 1. Let*

$$\rho_i(\gamma) = \left(b_1 \sum_{l=1}^{p_{Y,1}} |\psi_{y,1,i}^{(l)}| + b_\lambda b_2 \sum_{j=1}^{p_{Y,2}} |\psi_{y,2,i}^{(j)}|^{1/\gamma} \right) .$$

If either one of the following conditions holds:

1) ϕ is concave, $\max_{i=1,\ldots,r+1} \rho_i(1) < 1$, and $\max_{i=1,\ldots,r+1} \sum_{j=1}^{p_{Y,2}} |\psi_{y,2,i}^{(j)}| < 1$;
2) ϕ is a polynomial of order $\gamma \geq 1$ and $\max_{i=1,\ldots,r+1} \rho_i(\gamma) < 1$,

then $\{y_t\}$ is strictly stationary and ergodic.

Remark 2. The HAR model in [36] has a two-regime structure

$$\begin{aligned} y_t &= \begin{cases} \omega_1 + \sum_{l=1}^{p} \psi_{y,1}^{(l)} y_{t-l} + \sigma_1 e_t \,, & z_t^* = 1 \,, \\ \omega_2 + \sum_{l=1}^{p} \psi_{y,2}^{(l)} y_{t-l} + \sigma_2 e_t \,, & z_t^* = 0 \,, \end{cases} \\ z_t^* &= \begin{cases} 1 \,, & \text{if} \quad y_{t-d} \leq \theta_1 \,, \\ 0 \,, & \text{if} \quad y_{t-d} > \theta_1 + a \,, \\ z_{t-1}^* \,, & \text{otherwise} \,, \end{cases} \end{aligned} \tag{3.1}$$

where $a > 0$ and $(\theta_1, \theta_1 + a]$ is called the hysteresis region, is covered by (2.2) by defining $g_\lambda(\lambda_t, u_t) = 0$, $p_e = 0$, and

$$\begin{aligned} z_{t-d} \quad = \quad & y_{t-d} I(y_{t-d} \notin (\theta_1, \theta_1 + a]) \\ & + \sum_{l=1}^{\infty} y_{t-d-l} I(y_{t-d-l} \notin (\theta_1, \theta_1 + a]) \prod_{j=0}^{l-1} I(y_{t-d-j} \in (\theta_1, \theta_1 + a]) \,. \end{aligned} \tag{3.2}$$

Note that under (3.2), the threshold variable z_{t-d} is measurable with respect to the sigma field generated by $\{y_{t-d}, y_{t-d-1}, \ldots\}$ and thus does not satisfy Condition 1a). Nevertheless, Theorem 3.1 can be established if we replace Condition 1a) by:

Condition 1a′) The covariate X_t is independent of $\{y_s\}_{s<t}$ and $\mathrm{E}|X_t| < \infty$. In addition, there exists a positive integer $\tilde{p}$ such that $\{(X_t, \ldots, X_{t-\tilde{p}+1})\}_{t=1,\ldots}$ is Markovian, strictly stationary and ergodic. Moreover, there exists some integer q such that the vector $z_t^* = (I(z_{t-d} \in (\theta_0, \theta_1]), \ldots, I(z_{t-d} \in (\theta_{r-1}, \theta_r]))$ is measurable with respect to the sigma-field generated by $\{y_t, z_{t-1}^*, y_{t-1}, \ldots, y_{t-q}, z_{t-q}^*\}$. Also, z_t^* is irreducible and aperiodic.

Clearly, with $q = 1$ and $r = 1$, Condition 1a′) covers HAR model (3.1). In other words, Theorem 3.1 guarantees the stationarity and ergodicity of the multiple-regime extension of HAR model.

When specific knowledge is available on the threshold variable z_{k-d}, the conditions 1) and 2) in Theorem 3.1 can be relaxed as follows.

Corollary 1. *If $z_{t-d} = x_{t-d,k}$, then $\{y_t\}$ is strictly stationary and ergodic under either one of the following conditions:*

1′) ϕ is concave, $\sum_{i=1}^{r+1} \rho_i(1)\mathrm{pr}(z_{t-d} \in (\theta_{i-1}, \theta_i]) < 1$ and $\sum_{i=1}^{r+1} \sum_{j=1}^{p_{Y,2}} |\psi_{y,2,i}^{(j)}| \times \mathrm{pr}(z_{t-d} \in (\theta_{i-1}, \theta_i]) < 1$;

2′) ϕ is polynomial of order $\gamma \geq 1$ and $\sum_{i=1}^{r+1} \rho_i(\gamma)\mathrm{pr}(z_{t-d} \in (\theta_{i-1}, \theta_i]) < 1$.

Corollary 2. *If $z_{t-d} = y_{t-d}$, then the conditions of ρ_i in Theorem 3.1 can be relaxed to $\rho_1 < 1$, $\rho_{r+1} < 1$ and $\max\{\rho_1, \rho_{r+1}\}(1 - \pi_y) + \max_{i=2,\ldots,r} \rho_i \pi_y < 1$, where $\pi_y = \sup_u \mathrm{pr}(y_t \in (\theta_1, \theta_r] \mid Y_{t-1} = u)$.*

We have the following corollary for the classical TAR and TARMA models defined in (2.3).

Corollary 3. *Suppose that $\mathrm{E}(|e_t|) < \infty$ and $\max_{i=1,\ldots,r+1} \sum_{l=1}^{p_{Y,1}} |\psi_{y,1,i}^{(l)}| < 1$. If $p_e = 0$ (TAR model), then (2.3) is stationary and ergodic. If $p_e > 0$ (TARMA model), then (2.3) is stationary and ergodic provided that the irreducibility Condition 1f) holds.*

For TAR model, the condition in Corollary 3 is the same as [11]. For TARMA model, [38] also established the stationarity and ergodicity under the irreducibility condition. However, it requires $\sum_{l=1}^{p_{Y,1}} \max_i |\psi_{y,1,i}^{(l)}| < 1$, which is slightly stronger than the condition $\max_i \sum_{l=1}^{p_{Y,1}} |\psi_{y,1,i}^{(l)}| < 1$ in Corollary 3. We remark that recently [14] proves the irreducibility for TARMA(1,1) model under some parametric conditions (see Condition (C3) in [14]). This justifies the potential validity of Condition 1f).

4. Estimation and model selection criterion

Given the delay parameter d, the number of thresholds r, and the model order parameter p, the GTLVM is completely specified. Estimation of model parameters can be performed by maximum likelihood. Specifically, the log-likelihood

of the time series is

$$L(\Psi, \Theta, d) = \sum_{t=1}^{n} \left[\sum_{i=1}^{r+1} l(\Psi_i; y_t, Y_{t-1}, X_t) I(z_{t-d} \in (\theta_{i-1}, \theta_i]) \right], \tag{4.1}$$

where $l(\Psi_i; y_t, Y_{t-1}, X_t) = \log f(y_t \mid Y_{t-1}, X_t; z_{t-d} \in (\theta_{i-1}, \theta_i])$ is the conditional log-likelihood of y_t given $\{Y_{t-1}, X_t\}$ in regime i. Let r^0, d^0, p^0, Θ^0 and Ψ^0 be the true parameter values, and $\hat{\Theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_r)$, $\hat{\Psi} = (\hat{\Psi}_i, \ldots, \hat{\Psi}_{r+1})$ be the corresponding maximum likelihood estimators that maximize (4.1) given d, r and p.

Based on the likelihood function, traditional methods such as sequential chi-square likelihood ratio test from [12] can be derived to determine the number of thresholds. However, different autoregressive orders and combination of covariates in different regimes contribute to complication in implementing the traditional methods. To overcome the computational burden, we adopt a model selection approach by developing a criterion function based on the minimal description length (MDL) principle ([47, 48]). Given a model $\mathcal{M}$, the criterion is defined by

$$\mathrm{MDL}(\mathcal{M}) = \mathrm{CL}(\mathcal{M}) + \mathrm{CL}(\mathcal{E}_n \mid \mathcal{M}), \tag{4.2}$$

where the right hand side of (4.2) are the code lengths in bits for encoding the model and the fitted residuals $\mathcal{E}_n = (\hat{e}_1, \ldots, \hat{e}_n)$ given the model, respectively. Encoding the model $\mathcal{M}$ requires the specification of r, d, p, $\hat{\theta}_i$s and $\hat{\Psi}_i$s. Thus, $\mathrm{CL}(\mathcal{M})$ can be expressed as

$$\mathrm{CL}(\mathcal{M}) = \mathrm{CL}(r) + \mathrm{CL}(d) + \mathrm{CL}(\hat{\theta}_1, \ldots, \hat{\theta}_n) + \mathrm{CL}(\hat{\Psi}_1) + \cdots + \mathrm{CL}(\hat{\Psi}_{r+1}).$$

From [47, 48] and [30], it requires approximately $\log_2(n)$ bits to encode an integer and $(\log_2 n)/2$ bits to encode a maximum likelihood estimator with n data points. From [66], the thresholds can be associated with the order statistics of the threshold variables $\{z_1, \ldots, z_n\}$ and require $\sum_{i=1}^{r} \log_2(n_i)/2$ bits, where n_i is the number of observations in the ith regime. Recall that $\psi_{x,i}^{(k)} = 0$ when the kth covariate is not included in the model. As the integer 0 requires 1 bit and $\log_2 1 = 0$, by denoting $p'_{X,i} = \sum_{w=1}^{p_X} p_{X,i}^{(w)}$ as the number of nonzero entries in $p_{X,i}$, the maximum likelihood estimator $\hat{\Psi}_{X,i}$ can be encoded with $(p'_{X,i}/2)\log_2(n_i)$ bits. Similar arguments suggest that encoding $\hat{\Psi}_{Y,1,i}$, $\hat{\Psi}_{e,i}$, $\hat{\Psi}_{Y,2,i}$ and $\hat{\Psi}_{\epsilon,i}$ require $(p_{Y,1,i}/2)\log_2(n_i)$, $(p_{e,i}/2)\log_2(n_i)$, $(p_{Y,2,i}/2)\log_2(n_i)$ and $(p_{e,i}/2)\log_2(n_i)$ bits, respectively. Thus, we have

$$\begin{aligned}\mathrm{CL}(\mathcal{M}) = {} & \log_2(r) + \log_2(d) + \sum_{i=1}^{r} \frac{\log_2(n_i)}{2} + \sum_{i=1}^{r+1} \log_2(p'_i + 4) \\ & + \sum_{i=1}^{r+1} \frac{p'_i + 4}{2} \log_2(n_i),\end{aligned}$$

where $p'_i = p_{Y,1,i} + p_{e,i} + p_{Y,2,i} + p'_{X,i} + p_{\epsilon,i}$ is the total number of nonzero coefficients in regime i. From [47], $\mathrm{CL}(\mathcal{E}_t|\mathcal{M})$ can be approximated by the negative

of $\log_2$ of the likelihood. Hence,

$$\begin{aligned}\text{MDL}(\mathcal{M}) &= \log_2(r) + \log_2(d) + \sum_{i=1}^{r} \frac{\log_2(n_i)}{2} + \sum_{i=1}^{r+1} \log_2(p_i' + 4) \\ &\quad + \sum_{i=1}^{r+1} \frac{p_i' + 4}{2} \log_2(n_i) \\ &\quad - \log_2(e) L(\hat{\Psi}, \hat{\Theta}, d)\,. \end{aligned} \tag{4.3}$$

The optimal model can then be selected as the values $\hat{r}$, $\hat{d}$ and $\hat{p}$ that minimize the MDL($\mathcal{M}$).

Since the likelihood function remains constant when $\theta_i \in (R_j, R_{j+1}]$, $j = 1, \ldots, n-1$, where $\{R_1, \ldots, R_n\}$ is the ordered observations of the threshold variable, the estimator $\hat{\theta}_i$ may take any value on $(R_j, R_{j+1}]$ for some j. Without loss of generality we take $\hat{\theta}_i = R_{j+1}$. Then, to obtain an approximate solution to the optimization problem, we developed a genetic algorithm which is found to achieve promising performance for related optimization problems in change-point analysis ([18, 20], [41]) and estimation of the TAR model ([64] and [66]). Inspired by [66], we develop the methodology with modifications for a simultaneous detection of both autoregressive and covariate structures.

The genetic algorithm is an imitation of the biological evolution for optimization. It involves inheritance, crossover, mutation and filtering. Specifically, the algorithm begins with a population of *chromosomes*, where each chromosome stores the information of a candidate solution to the optimization problem. For our application, we first fix a d and define each candidate solution as a model $\mathcal{M}$ specified by some r and p. Then, the performance of each chromosome is measured by its information criterion MDL($\mathcal{M}$). Chromosomes with better performance have higher probabilities to conduct crossover and produce *offspring*, and so their good model features are more likely to be inherited. Meanwhile, mutation occurs with a small probability, which brings in new models to seek the global optimum. After several generations of crossover and mutation, the best performing model is selected as the optimal one. Finally, we repeat the procedure with different d and select the best performing model that attains the smallest MDL($\mathcal{M}$).

Specifically, each step of the genetic algorithm is described as follows.

Chromosome Formation: First, generate the initial *population*, which is a set of chromosomes in vector form. Each chromosome is expressed as $c = [r, p_1, (\theta_1, p_2), \ldots, (\theta_r, p_{r+1})]$; the parameter estimate $\hat{\Psi}$ is obtained once c is specified. Similar to [66], the initial population is created as follows:

1) The number of thresholds r is generated from a Poisson distribution with mean 2.
2) Sample θ_is uniformly from $\{z_t\}$. Reject the sample and sample again if any regime has fewer than $\tau_{n,0}$ observations. This *minimum span condition* ensures the estimation accuracy of Ψ.

3) Select $p_{Y,1,i}$, $p_{e,i}$, $p_{Y,2,i}$ and $p_{\epsilon,i}$ uniformly from $\{0,\ldots,Q_{Y,1}\}$, $\{0,\ldots,Q_e\}$, $\{0,\ldots,Q_{Y,2}\}$ and $\{0,\ldots,Q_\epsilon\}$, respectively, where $Q_{Y,1}$, Q_e, $Q_{Y,2}$ and Q_ϵ are pre-specified upper bounds of model orders. Generate a binary vector $p_{X,i}$ of length p_X with each element following an independent Bernoulli distribution with mean 0.5. Set $p_i = (p_{Y,1,i}, p_{e,i}, p_{Y,2,i}, p_{X,i}, p_{\epsilon,i})$.

The MDL($\mathcal{M}$) of each chromosome is then computed by (4.3).

Crossover and Mutation: Crossover and mutation are two methods for generating offspring. In crossover, two chromosomes are selected from the population as "parents" with probabilities proportional to the inverse of their ranks of MDL($\mathcal{M}$). Next, a p_1^o is drawn from one of the parent's p_1 with equal probability. Then, for both parents, each of their $\{\theta_j, p_{j+1}\}$ is selected with probability 0.5. Sort all selected $\{\theta_j, p_{j+1}\}$s in ascending order of θ_j to produce an offspring $[r^o, p_1^o, (\theta_1^o, p_2^o), \ldots, (\theta_r^o, p_{r^o+1}^o)]$. If some thresholds θ_j^os violate the minimum span condition, randomly delete the pair $\{\theta_j^o, p_{j+1}^o\}$ until the condition is satisfied.

In mutation, one parent chromosome is selected from the population with probabilities proportional to the inverse of the ranks of MDL($\mathcal{M}$). Then, a new chromosome is generated to crossover with the parent chromosome to produce an offspring. To achieve a higher degree of exploration, the features from the generated parent are selected with probability 0.7.

To balance between retaining good features with crossover and bringing new solutions with mutation, the probabilities of performing crossover and mutation are 0.9 and 0.1, respectively. To explore more possibilities in the model order, every offspring will have its order parameter in one randomly selected regime replaced by a newly generated order, with probability 0.3.

After conducting crossover and mutation, the group of offspring become a new generation of chromosomes. To ensure monotonicity of optimization, an *elitist* step is conducted to replace the worst performing 20 chromosomes in the new generation by the best performing chromosome in the previous generation.

Migration: With the advance of parallel computing, the *island model* is introduced to accelerate the computation and alleviate trapping in suboptimal solutions. In particular, we perform genetic algorithm on N_I groups of subpopulations with size N_p. These subpopulations conduct their own reproduction steps and thus are treated as distinct islands. To share good features between islands, for every M_i generations, the M_N worst performing chromosomes in the jth island are replaced by the best M_N chromosomes from the $(j-1)$th island, for $j = 1,\ldots,N_I$, where the 0th island is conventionally defined as the N_I-th island. In this article we used $N_I = 50$, $N_p = 200$, $M_i = 4$, and $M_N = 2$. The full mechanism and improvement in performance of the parallelized genetic algorithm can be found in [1, 2].

Claim of Convergence: When the best chromosome remains unchanged over 20 generations, we claim that convergence is achieved and the optimal model is obtained from the parameters of the best chromosome. Alternatively, in consideration of computational efficiency, the algorithm may be stopped after a fixed number of generations.

5. Asymptotic inferences of GTLVM

5.1. Assumptions for asymptotic inferences

Apart from Condition 1, we state the following assumptions for asymptotic inferences. First, Assumptions 1–3 are proposed for the consistency:

Assumption 1. The parameters (Ψ, Θ, d) are in the space $\Omega_{\mathcal{M}} \times \{1, \ldots, D\}$, where $\Omega_{\mathcal{M}} = \Omega_{\Psi} \times \Omega_{\Theta}$ is a compact subset of $\mathbb{R}^{(p_{Y,1}+p_e+p_{Y,2}+p_X+p_\epsilon+4)\times(r+1)} \times \mathbb{R}^r$, and D is some positive integer. In addition, (Ψ^0, Θ^0) is an interior point in $\Omega_{\mathcal{M}}$.

Assumption 2. The conditional density $f(y_t \mid Y_{t-1}, X_t, z_{t-d})$ is regular in the sense that the maximum likelihood estimator $\hat{\Psi}_i$ is asymptotically normal ([17] and [44]). We assume $\mathrm{E}[\{\partial \log(f(y_t \mid Y_{t-1}, X_t, z_{t-d}))/\partial \Psi_i\}^2] < \infty$. Moreover, $l(\Psi_i; y_t, Y_{t-1}, X_t))$ is concave in Ψ_i, and $\partial^3 l(\Psi_i; y_t, Y_{t-1}, X_t)/\partial \Psi_i^3$ exists for $i = 1, \ldots, r+1$ and is bounded by an integrable function in the neighborhood of Ψ_i^0. Furthermore, $f(y_t \mid Y_{t-1}, X_t, z_{t-d} \in (\theta_{i-1}^0, \theta_i^0])$ and $f(y_t \mid Y_{t-1}, X_t, z_{t-d} \in (\theta_i^0, \theta_{i+1}^0])$ are not equal almost everywhere for all i and all $\{Y_{t-1}, X_t\}$.

Assumption 3. The joint probability density of $\{z_{t-i}, z_{t-j}\}$, $\pi_{z,i-j}(\cdot,\cdot)$, is uniformly bounded. In addition, for any vector Φ that has the same dimension as $(1, X_t^{\mathrm{T}})$ and satisfies $|\Phi| = 1$, there exists an $\epsilon > 0$ such that

$$\mathrm{pr}(|\Phi(1, X_t^{\mathrm{T}})^{\mathrm{T}}| > \epsilon | z_{t-i}, z_{t-j}) > 0 \qquad \text{almost surely}, \tag{5.1}$$

with respect to the joint distribution of (z_{t-i}, z_{t-j}), where $i, j = 1, \ldots, D$.

Assumptions 1–2 are standard regulatory conditions for statistical inference in parametric models. The assumption on the third-order derivatives of $L(\Psi, \Theta, d)$ is essential in deriving the asymptotic distribution of $\hat{\Theta}$; see Theorem 5.41 of [62]. Assumption 3, which is in similar spirit as Assumption 3 in [54], assumes the linear independence of X_t to eliminate redundancy in the covariates. It holds if the joint conditional density of the exogenous covariates X_t is non-degenerate given z_{t-i} and z_{t-j}.

Additionally, the following two assumptions are required for the convergence rates of estimators:

Assumption 4. There exists an integrable function $\Gamma(Y_{t-1}, X_t, y_t)$ satisfying

$$\left| l(\Psi^{(1)}; y_t, Y_{t-1}, X_t) - l(\Psi^{(2)}; y_t, Y_{t-1}, X_t) \right| < \Gamma(Y_{t-1}, X_t, y_t) |\Psi^{(1)} - \Psi^{(2)}| \quad \text{a.s.},$$

for any $\Psi^{(1)}, \Psi^{(2)} \in \Omega_{\Psi}$. Furthermore, we assume that either one of the following condition holds:

a) The marginal density of z_t, $\pi_z(\cdot)$, is continuous at $\{\theta_1^0, \ldots, \theta_r^0\}$, and the joint probability density of $\{z_{t-i}, z_{t-j}\}$, $\pi_{z,i-j}(\cdot,\cdot)$, is positive everywhere. Moreover, there exists an $\epsilon > 0$ such that, for all $z_{t-d^0} \in [\theta_i^0 - \epsilon, \theta_i^0 + \epsilon]$, $i = 1, \ldots, r$, $\mathrm{E}[\Gamma^2(Y_{t-1}, X_t, y_t)] < \infty$. The joint conditional distribution function of $\{Y_{t-1}, X_t\}$ given z_{t-d^0} is continuous at $z_{t-d^0} = \theta_1^0, \ldots, \theta_r^0$.

b) The threshold variable z_t is discrete, and y_t is in the exponential family satisfying

$$\begin{cases} f(y_t;\lambda_t,a_t) = \sum_{i=1}^{r+1} \exp\left[\frac{1}{\nu_i a_t}\{T(y_t)\gamma(\lambda_t) - b(\lambda_t)\} + c(y_t;\nu_i a_t)\right] \\ \qquad\qquad \times I(z_{t-d} \in (\theta_{i-1}^0, \theta_i^0])\,, \\ \phi(\lambda_t) = \sum_{i=1}^{r+1}\left[\alpha_i + \sum_{j=1}^{p_Y} \psi_{y,i}^{(j)} g_y(y_{t-j}) + \sum_{k=1}^{p_X} \psi_{x,i}^{(k)} x_{t,k}\right] \\ \qquad\qquad \times I(z_{t-d} \in (\theta_{i-1}, \theta_i])\,, \end{cases} \tag{5.2}$$

where $T(\cdot)$ is a measurable function such that $\mathrm{var}(T(y_t) \mid \lambda_t) \in (0,\infty)$ almost surely, $\gamma(\cdot), b(\cdot)$ are continuous transformation functions, and ν_i is the overdispersion parameter for regime i. Moreover, we assume that $\mathrm{E}[\Gamma^2(Y_{t-1}, X_t, y_t)] < \infty$, $\limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^n c_t^{3/2} < \infty$ and $\log[\mathrm{E}(e^{uT(y_t)} \mid \lambda_t)] \le c_t$ for some constant u and sequence $\{c_t\}$.

Assumption 5. When z_t is continuous, there exists some $\Delta > 0$ such that the process $\{\Gamma(Y_{t-1}, X_t, y_t) I(z_{t-d^0} \in [\theta_i^0 - \Delta, \theta_i^0 + \Delta]), z_{t-d^0} I(z_{t-d^0} \in [\theta_i^0 - \Delta, \theta_i^0 + \Delta])\}$ is ρ-mixing with summable mixing coefficients.

Assumption 4 are the conditions for the likelihood function with respect to continuous and discrete threshold variables. Asymptotic theory for threshold models with discrete threshold variables does not seem to have been studied in the literature. Similar to Assumption 6 in [54], we impose a square-integrable bound function for the difference of log-likelihood in both Assumption 4a) and 4b). While an example of verifying Assumption 4a) has been shown in supplementary materials of [54], an example of verifying Assumption 4b) is given in the Appendix. Furthermore, Assumption 4b) requires that the conditional density of y_t given $\{Y_{t-1}, X_t, z_{t-d}\}$ is in the exponential family. Thus, an explicit form of the difference in log-likelihood is available for applying results in large deviation theory; see [46]. Note that although the assumptions and asymptotic properties of $\hat{\Theta}$ for continuous and discrete z_t are different, the same estimation procedure proposed in Section 4 is applied. Assumption 5 is analogous to Assumption 7 in [54] for proving the convergence of $n(\hat{\Theta} - \Theta^0)$. For any integer j, let $\mathcal{A}$ and $\mathcal{A}^*$ be the σ-algebras generated by $\{w_t\}_{t\le j}$ and $\{w_t\}_{t\ge j+k}$, respectively. If $\{w_t\}$ is ρ-mixing, then there exists a sequence $\{\rho(k)\}_{k=1,2,\ldots}$ with $\lim_{k\to\infty}\rho(k) \to 0$ such that, for all square-integrable random variables g and h that are respectively $\mathcal{A}$ and $\mathcal{A}^*$measurable, $|\mathrm{corr}(g,h)| \le \rho(k)$ holds; see [7] and [21]. See also Examples 1 and 2 in the supplementary materials of [54] for verification of ρ-mixing and selections of the function Γ. We will illustrate the verification of the above assumptions in some explicit examples in Section 6.

5.2. *Asymptotic theorems*

Under Condition 1, we develop the consistency and asymptotic properties of the parameter estimates. The proofs are provided in the appendix.

Theorem 5.1. *If Assumptions 1–3 hold, then* $\hat{d} \to d^0$ *and* $\hat{r} \to r^0$ *almost surely. In addition, on* $\{\hat{d} = d^0, \hat{r} = r^0\}$, $\hat{\Theta} \to \Theta^0$, $\hat{p} \to p^0$ *and* $\hat{\Psi} \to \Psi^0$ *almost surely.*

Next, we derive the convergence rate of $\hat{\Theta}$ for continuous or discrete z_t, respectively:

Theorem 5.2. *Under Assumptions 1–3,*

1) *if Assumptions 4a) and 5 are satisfied, then* $|\hat{\Theta} - \Theta^0| = O_p(n^{-1})$;
2) *if Assumption 4b) is satisfied, then* $\text{pr}(\hat{\Theta} = \Theta^0) = 1 - O(n^{1/2}e^{-an})$ *for some* $a > 0$.

Denote the difference between the log-likelihood of y_s under parameters $\Psi^{(1)}$ and $\Psi^{(2)}$ by

$$\zeta_s(y_s; Y_{s-1}, X_s, \Psi^{(1)}, \Psi^{(2)}) = l(\Psi^{(1)}; y_s, Y_{s-1}, X_s) - l(\Psi^{(2)}; y_s, Y_{s-1}, X_s)\,.$$

Define a double-sided compound Poisson process $\tilde{\ell}^*_i(\kappa_i) = \tilde{\ell}^*_{1,i}(\kappa_i)I(\kappa_i \geq 0) + \tilde{\ell}^*_{2,i}(\kappa_i)I(\kappa_i < 0)$, where

$$\tilde{\ell}^*_{1,i}(\kappa_i) = \sum_{s=1}^{N_{1,i}(\kappa_i)} \zeta_s(y^*_s; Y^*_{s-1}, X^*_s, \Psi^0_{i+1}, \Psi^0_i)\,, \tag{5.3}$$

is a compound Poisson process such that $N_{1,i}(\kappa_i)$ is a Poisson processes with intensity $\pi_z(\theta^0_i)$, and $(y^*_s,\, Y^*_{s-1},\, X^*_s)$ is an independent copy of $(y_s,\, Y_{s-1},\, X_s)$ given $z_{s-d} = (\theta^0_i)^+$. Moreover,

$$\tilde{\ell}^*_{2,i}(\kappa_i) = \sum_{s=1}^{N_{2,i}(-\kappa_i)} \zeta_s(y^*_s; Y^*_{s-1}, X^*_s, \Psi^0_i, \Psi^0_{i+1})\,, \tag{5.4}$$

is a compound Poisson process, independent of $\tilde{\ell}^*_{1,i}(\kappa_i)$, defined by where $N_{2,i}(-\kappa_i)$ is a Poisson processes with the same intensity $\pi_z(\theta^0_i)$, and $(y^*_s,\, Y^*_{s-1},\, X^*_s)$ is an independent copy of $(y_s,\, Y_{s-1},\, X_s)$ given $z_{s-d} = (\theta^0_i)^-$. In addition, define an aggregated compound Poisson process $\tilde{\ell}^*(\kappa) = \sum_{i=1}^{r+1} \tilde{\ell}^*_i(\kappa_i)$. The following theorems derive the asymptotic distribution of the threshold parameters $\hat{\Theta}$ and model parameters $\hat{\Psi}$.

Theorem 5.3. *Under Assumptions 1–3, 4a) and 5, then* $n(\widehat{\Theta} - \Theta^0)$ *weakly converges to* M^-, *where the random* r*-dimension cube* $[M^-, M^+) = [M^-_1, M^+_1) \times [M^-_2, M^+_2) \times \cdots \times [M^-_r, M^+_r)$ *is an almost surely minimizer of the compound Poisson process* $\tilde{\ell}^*(\kappa)$. *In addition, if* $f(y \mid Y_{t-1}, X_t, z_{t-d})$ *is continuous in* y, *then* $[M^-, M^+)$ *is unique.*

Remark 3. Note that $[M^-, M^+)$ depends on Θ^0 and Ψ^0. For constructing confidence intervals for $\hat{\Theta}$, algorithms for estimating M^- can be derived by similar methods in [32] and [67].

Theorem 5.4. *Define*

$$L'(\Psi, \Theta, d) = \frac{\partial}{\partial \Psi}\left[\frac{1}{n}L(\Psi, \Theta, d)\right], \quad L''(\Psi, \Theta, d) = \frac{\partial^2}{\partial \Psi^2}\left[\frac{1}{n}L(\Psi, \Theta, d)\right].$$

Under Assumptions 1–5, $n^{1/2}(\hat{\Psi} - \Psi^0) \to_{d.} N(0, \Sigma^*)$*, where*

$$\Sigma^* = -\left[\mathrm{E}\{L''(\Psi^0, \Theta^0, d^0)\}\right]^{-1}. \tag{5.5}$$

Moreover, Σ^* *is a diagonal block matrix, and thus* $\{\hat{\Psi}_i\}_{i=1,\ldots,r+1}$ *are asymptotically independent. In addition,* $n(\widehat{\Theta} - \Theta^0)$ *and* $n^{1/2}(\widehat{\Psi} - \Psi^0)$ *are asymptotically independent.*

6. Simulation

In this section, two simulation experiments are performed to illustrate model fitting with respect to discrete and continuous thresholds. In each simulation experiment, 300 replications are conducted for every scenario. For simplicity, d is assumed to be known.

6.1. Example 1

Consider a three-regime self-excited GTLVM with Poisson distribution and log-link:

$$\begin{cases} f(y_t; \lambda_t) & = {\lambda_t}^{y_t} e^{-\lambda_t}/y_t!\,, \\ \log(\lambda_t) & = \sum_{i=1}^{r+1} \left[\alpha_i + \sum_{j=1}^{p_{Y,2}} \psi_{y,i}^{(j)} \log(y_{t-j} + 1)\right] I(y_{t-4} \in (\theta_{i-1}, \theta_i])\,, \end{cases} \tag{6.1}$$

with $\{\theta_1, \theta_2\} = \{24, 42\}$, $\{\alpha_1, \alpha_2, \alpha_3\} = \{0.45, 3.4, 6.3\}$, $\{p_{Y,2,1}, p_{Y,2,2}, p_{Y,2,3}\} = \{2, 2, 1\}$, and $\{(\psi_{y,1}^{(1)}, \psi_{y,1}^{(2)}), (\psi_{y,2}^{(1)}, \psi_{y,2}^{(2)}), \psi_{y,3}^{(1)}\} = \{(0.65, 0.25), (0.4, -0.35), -0.95\}$. A time series plot of realization is shown in Figure 1.

Here we verify Condition 1, and the assumptions in Section 5 for model (6.1). Note that (6.1) can be expressed as (2.2) with $\phi(x) = \log(x)$, $g_\lambda(\lambda_t, \cdot)$ being the inverse c.d.f. of Poisson distribution with parameter λ_t, $g_{y,1}(x) \equiv 0$, $g_{y,2}(x) = \log(x + 1)$, $p_X = p_e = p_\epsilon = 0$, and $\psi_{y,1,i}^l = 0$ for all i and l. As no covariate is used and $p_e = p_\epsilon = 0$, it suffices to check Condition 1 b), c), d). From the forms of ϕ, $g_{y,1}$ and $g_{y,2}$, Conditions 1b) and c) clearly hold with $b_1 = b_2 = 1.1$. Given λ and the uniform random variable u on $[0, 1]$, $g_\lambda(\lambda, u)$ is a Poisson random variable with parameter λ. Thus, $\mathrm{E}[|g_\lambda(\lambda, u)| \mid \lambda] = \lambda$ and Condition 1d) holds with $b_\lambda = 1$. Therefore, Condition 1 holds and Theorem 3.1 implies that (6.1) is stationary and ergodic.

As the threshold variable is discrete and no covariate is used, it suffices to verify Assumptions 1, 2 and 4b). As is common in the literature, Assumption 1 can be achieved by focusing attention on a sufficiently large compact subset of the parameter space. For Assumption 2, note that the conditional density $f(y_t \mid Y_{t-1}, y_{t-4})$ is the density of Poisson distribution, and is thus regular. Also, $\mathrm{E}[\{\partial \log(f(y_t \mid Y_{t-1}, y_{t-4}))/\partial \psi_{y,i}^{(j)}\}^2] = E((y_t - \lambda_t)(\log(y_{t-j} + 1))^2 < \infty$. Moreover, the third-order derivative of $l(\Psi, \Theta, d)$ with respect to $\psi_{y,i}^{(j)}$ exists and is equal to $-\lambda_t(\log(y_{t-j} + 1))^3$. Also, $f(y_t \mid Y_{t-1}, y_{t-4} \in (\theta_{i-1}^0, \theta_i^0])$ and

$f(y_t \mid Y_{t-1}, y_{t-4} \in (\theta_i^0, \theta_{i+1}^0])$ are not equal almost everywhere for all i since the latent variable λ_t takes different values in the two regimes. Verification of Assumption 4b) is more technical and is given in the Appendix.

TABLE 1

Example 1: Percentage of correct number of estimated thresholds and correct model structure specification. MDL*: minimum description length;* BIC*: Bayesian information criterion.*

Size	Regime structure classified(%), MDL	Model structure classified(%), MDL	Regime structure classified(%), BIC	Model structure classified(%), BIC
200	91.7	65.3	85.6	60.0
500	96.7	94.0	94.3	87.3
1000	99.3	96.3	98.3	94.3
2000	100.0	97.0	99.3	93.7

TABLE 2

Example 1: Mean and standard derivations of parameter estimates. Parentheses: empirical standard derivations. The threshold estimates are based on replications with correct number of regimes. The coefficient estimates are based on replications with correct model structure.

Size	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\alpha}_1$	$\hat{\psi}^{(1)}_{Y,2,1}$	$\hat{\psi}^{(2)}_{Y,2,1}$	$\hat{\alpha}_2$	$\hat{\psi}^{(1)}_{Y,2,2}$	$\hat{\psi}^{(2)}_{Y,2,2}$	$\hat{\alpha}_3$	$\hat{\psi}^{(1)}_{Y,2,3}$
200	24.036	42.007	0.430	0.631	0.276	3.411	0.395	-0.348	6.260	-0.938
	(0.875)	(0.171)	(0.324)	(0.091)	(0.068)	(0.234)	(0.050)	(0.052)	(0.267)	(0.083)
500	24.003	42.000	0.463	0.647	0.250	3.396	0.397	-0.346	6.285	-0.946
	(0.228)	(0.000)	(0.214)	(0.050)	(0.043)	(0.126)	(0.033)	(0.032)	(0.131)	(0.041)
1000	24.006	42.000	0.452	0.649	0.250	3.400	0.399	-0.349	6.293	-0.948
	(0.082)	(0.000)	(0.174)	(0.039)	(0.034)	(0.094)	(0.023)	(0.023)	(0.097)	(0.030)
2000	24.000	42.000	0.447	0.652	0.249	3.402	0.401	-0.351	6.293	-0.948
	(0.000)	(0.000)	(0.102)	(0.025)	(0.021)	(0.070)	(0.016)	(0.015)	(0.063)	(0.020)
True	24	42	0.45	0.65	0.25	3.4	0.4	-0.35	6.3	-0.95

Table 1 reports the model selection performance. Even for a small sample of size of 200, the percentage of correct number of estimated threshold is over 80%, and the percentage of correct identification of model order in all regimes is over 60%. For comparison, we repeated the experiment with the Bayesian Information Criterion, which is defined as $-2L(\Psi, \Theta, d) + \sum_{i=1}^{r+1} \log(n_i) p_i'$. It is found that minimal description length gives better performance. Other information criteria such as NAIC in [59] might also be adopted; however, NAIC and other AIC-type criteria are not consistent in estimating the true order of the model. See [26] for details about consistency of information criterion.

Furthermore, Table 2 summarizes the performance of thresholds and model parameters estimates within the replications that correctly specify the regime and model structure using minimal description length. A fast convergence speed for discrete thresholds is observed. Furthermore, $O_p(n^{-1/2})$ convergence rate of $\hat{\Psi}$ is realized based on the rapid convergence of $\hat{\Theta}$.

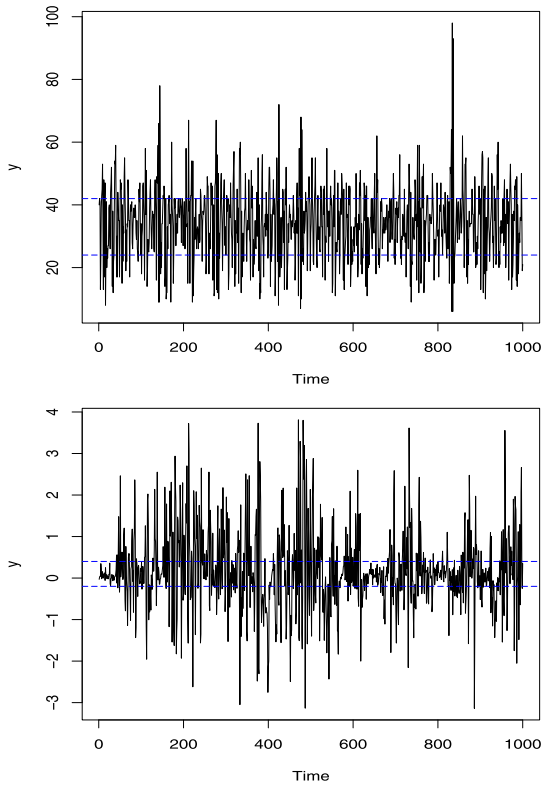

FIG 1. *Sample plots of simulated series. Left: Example 1, with thresholds at* 24 *and* 42 *(horizontal dashed line). Right: Example 2, with thresholds at* -0.2 *and* 0.4 *(horizontal dashed line)*

6.2. Example 2

We consider a double threshold autoregressive model with conditional heteroskedasticity by a log-link on σ_t^2:

$$\begin{cases} y_t & = \sum_{i=1}^{r+1} \left[\omega_i + \sum_{l=1}^{p_{Y,1}} \psi_{y,1,i}^{(l)} y_{t-l} + \sigma_t \epsilon_t\right] I(y_{t-4} \in (\theta_{i-1}, \theta_i]) \,, \\ \log(\sigma_t^2) & = \sum_{i=1}^{r+1} \left[\alpha_i + \sum_{j=1}^{p_{Y,2}} \psi_{y,2,i}^{(j)} \log(y_{t-j}^2)\right] I(y_{t-4} \in (\theta_{i-1}, \theta_i]) \,. \end{cases} \tag{6.2}$$

Here $\{\theta_1, \theta_2\} = \{-0.2, 0.4\}$, $\{\omega_1, \omega_2, \omega_3\} = \{-0.15, 0.1, 0.3\}$, $\{p_{Y,1,1}, p_{Y,1,2}, p_{Y,1,3}\} = \{1, 1, 2\}$, $\{(\psi_{y,1,1}^{(1)}), (\psi_{y,1,2}^{(1)}), (\psi_{y,1,3}^{(1)}, \psi_{y,1,3}^{(2)})\} = \{(0.6), (0.25), (0.25, -0.7)\}$, $\{\alpha_1, \alpha_2, \alpha_3\} = \{-0.4, -0.2, 0.15\}$, $\{p_{Y,2,1}, p_{Y,2,2}, p_{Y,2,3}\} = \{1, 2, 0\}$, $\{(\psi_{y,2,1}^{(1)}), (\psi_{y,2,2}^{(1)}, \psi_{y,2,2}^{(2)}), (\psi_{y,2,3}^{(1)})\} = \{(0.35), (0.45, 0.25), (0)\}$, with $\epsilon_t \overset{iid.}{\sim} \mathcal{N}(0, 1)$. In the third regime, we select $p_{Y,2,3} = 0$ so $\sigma_t^2 = e^{0.15}$ is a constant. A sample time series plot is depicted in Figure 1.

Note that (6.2) can be expressed as (2.2) with $\phi(x) = \log(x)$, $g_\lambda(\lambda_t, \cdot) \equiv 0$, $g_{y,1}(x) = x$, $g_{y,2}(x) = 2\log(x)$ and $p_X = p_e = p_\epsilon = 0$. Condition 1 can be verified readily using similar arguments in the verification for model (6.1). As the threshold variable is continuous and no covariate is used, it suffices to verify Assumptions 1, 2 and 4a). While Assumptions 1, 2 can be verified similarly as in Example 1, the verification of Assumption 4a) is similar to [54].

TABLE 3

Example 2: Percentage of correct number of estimated threshold, and correct model structure specification. MDL, *minimum description length;* BIC, *Bayesian information criterion.*

Size	Regime structure classified(%), MDL	Model structure classified(%), MDL	Regime structure classified(%), BIC	Model structure classified(%), BIC
200	83.3	5.00	56.0	3.00
500	97.3	47.7	78.0	36.7
1000	99.3	79.3	97.7	78.7
2000	100.0	90.0	100.0	88.0

TABLE 4

Example 2: Mean and standard derivations of parameter estimates, thresholds and autoregressive coefficients. Parentheses: empirical standard derivations. The threshold estimates are based on replications with correct number of regimes. The coefficient estimates are based on replications with correct model structure.

Size	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\omega}_1$	$\hat{\psi}^{(1)}_{y,1,1}$	$\hat{\omega}_2$	$\hat{\psi}^{(1)}_{y,1,2}$	$\hat{\omega}_3$	$\hat{\psi}^{(1)}_{y,1,3}$	$\hat{\psi}^{(2)}_{y,1,3}$
200	-0.213	0.402	-0.143	0.565	0.107	0.369	0.244	0.196	-0.708
	(0.122)	(0.063)	(0.057)	(0.096)	(0.028)	(0.086)	(0.177)	(0.110)	(0.093)
500	-0.207	0.402	-0.154	0.594	0.101	0.267	0.294	0.253	-0.696
	(0.030)	(0.018)	(0.036)	(0.067)	(0.018)	(0.061)	(0.096)	(0.078)	(0.079)
1000	-0.202	0.400	-0.148	0.599	0.101	0.250	0.300	0.250	-0.703
	(0.016)	(0.008)	(0.026)	(0.049)	(0.009)	(0.047)	(0.064)	(0.055)	(0.061)
2000	-0.201	0.399	-0.149	0.600	0.100	0.248	0.297	0.247	-0.697
	(0.008)	(0.004)	(0.020)	(0.037)	(0.007)	(0.033)	(0.040)	(0.040)	(0.040)
True	-0.2	0.4	-0.15	0.6	0.1	0.25	0.3	0.25	-0.7

TABLE 5

Example 2: Mean and standard derivations of parameter estimates, log-link coefficients. Parentheses: empirical standard derivations. The coefficient estimates are based on replications with correct model structure.

Size	$\hat{\alpha}_1$	$\hat{\psi}^{(1)}_{y,2,1}$	$\hat{\alpha}_2$	$\hat{\psi}^{(1)}_{y,2,2}$	$\hat{\psi}^{(2)}_{y,2,2}$	$\hat{\alpha}_3$
200	-0.401(0.176)	0.359(0.067)	-0.099(0.242)	0.447(0.077)	0.333(0.096)	0.016(0.138)
500	-0.419(0.135)	0.357(0.044)	-0.213(0.168)	0.455(0.047)	0.253(0.049)	0.114(0.112)
1000	-0.425(0.093)	0.349(0.033)	-0.205(0.123)	0.451(0.032)	0.254(0.033)	0.139(0.074)
2000	-0.404(0.067)	0.350(0.025)	-0.202(0.092)	0.449(0.022)	0.251(0.022)	0.145(0.058)
True	-0.4	0.35	-0.2	0.45	0.25	0.15

The empirical classification rate of regime and model structure for (6.2), and comparisons between criteria, are reported in Table 3. Again, the results are promising for moderately large sample sizes, and minimal description length still achieves superior performance. Tables 4 and 5 summarize estimation results for thresholds and model parameters, where the asymptotic convergences such

as $O_p(n^{-1})$ rate for $\hat{\Theta}$, and $O_p(n^{-1/2})$ for $\hat{\Psi}$, are recognized. In conclusion, the purposed methodology has satisfactory performance.

7. Application

Initial public offerings (IPOs) are one of the most important funding sources in finance. IPO activities are found to be time-varying; see, for example, [50], which discusses the cyclical effect with respect to stock market bull/bear trends and cross-year variation of levels of monthly IPO volumes. This is evidenced by a large variation with clustering of small and large observations across yearly periods. For instance, retreats of IPO activities are observed in 1973-1979 for the oil crisis, in 1982 for the energy crisis, in 2001 for the burst of Internet bubbles, and in 2008-2009 for the financial crisis; meanwhile, fervent IPO activity occurred in 1983-1987 during the stable global market with economic expansion, and in 1992-2000 during the high-tech boom.

However, few studies of IPO activities have considered quantitative modeling. [39] and [40] purposed autoregressive models which incorporates past monthly initial returns and market participation proxies. Nevertheless, linear autoregressive modeling is theoretically questionable for integer IPO volumes. Moreover, as indicated in [49] and [68], different market scenarios exist in the IPO market. These facts suggest the necessity of regime classifications for proper modeling.

Thus, for theoretical soundness and modeling flexibility, the GTLVM is applied to IPO volumes modeling. We model the U.S. monthly net IPO volumes from January 1976 to March 2014, where the dataset is available at [51]) (https://www.quandl.com/data/RITTER/US_IPO_STATS-Historical-US-IPO-Statistics). By the definition in [51], net IPO volumes exclude issuance of penny stocks, units and close-end funds.

To flexibly model counting data, we use a negative binomial distribution with a canonical log-link function in order to capture different dispersions across regimes. Here

$$f_i(y_t;\lambda_t)=\sum_{i=1}^{r+1}\left[\frac{\Gamma(y_t+k_i)}{\Gamma(k_i)\Gamma(y_t+1)}\left(\frac{k_i}{k_i+\lambda_t}\right)^{k_i}\left(\frac{\lambda_t}{k_i+\lambda_t}\right)^{y_t}\right]I(y_{t-d}\in(\theta_{i-1},\theta_i])\,,$$
$$y_t\in\mathbb{N}\,,$$

where λ_t satisfies $\lambda_t=\mathrm{E}(y_t)$, and k_i is the dispersion parameter. With $y^*_{t-j}=\log[\max(y_{t-j},0.01)]=g_{y,2}(y_{t-j})$, the link function is expressed as

$$\log(\lambda_t)=\sum_{i=1}^{r+1}\left[\alpha_i+\sum_{j=1}^{p_{Y,2}}\psi^{(j)}_{y,i}y^*_{t-j}+\sum_{k=1}^{p_X}\psi^{(k)}_{x,i}x_k\right]I(y_{t-d}\in(\theta_{i-1},\theta_i])\,.$$

To capture the dependence of the variables in the previous 12 months, we set the maximum delay as $D=12$ and autoregressive order as $Q_{Y,2}=12$. And as indicated by [39], IPO activities are affected by new information arriving during

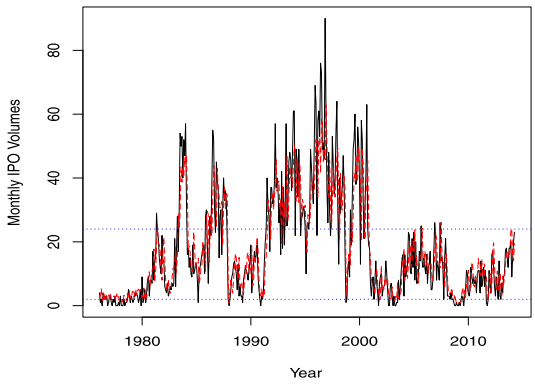

FIG 2. *Plot of U.S. monthly net IPO volumes. Solid: original series; dashed: fitted series; dotted: threshold estimates*

the book-building period which lasts for approximately two to four months. Hence, we check the averaged historical two-, three- or four-month return of the S&P 500 Index, denoted as $x^{(2)}$, $x^{(3)}$, and $x^{(4)}$, respectively, as possible covariates that represent recent market performance. In addition, the past observations y^*_{t-2}, y^*_{t-3} and y^*_{t-4} are included in the model. In each regime, the parameter Ψ_i is estimated by quasi-maximum likelihood with iteratively reweighted least square method, see [42].

For the net IPO series, two thresholds are estimated as $\hat{\theta}_1 = 2$ and $\hat{\theta}_2 = 24$. The estimated dispersion k_i in the three regimes are 1.5962, 4.1817 and 9.9228, with theoretical standard errors of 0.4593, 0.5342 and 1.6211, respectively. The link function estimate is

$$\log(\lambda_t) = \begin{cases} \underset{(0.1178)}{1.0242} + \underset{(0.0484)}{0.1487} y^*_{t-4} + \underset{(2.9743)}{11.9900} x^{(2)}_t\,, & y_{t-1} \le 2\,, \\ \underset{(0.1508)}{0.4552} + \underset{(0.0746)}{0.5595} y^*_{t-1} + \underset{(0.0360)}{0.1549} y^*_{t-2} + \underset{(0.0290)}{0.1300} y^*_{t-4} + \underset{(1.3092)}{7.6926} x^{(2)}_t\,, & 2 < y_{t-1} \le 24\,, \\ \underset{(0.4239)}{1.0459} + \underset{(0.1222)}{0.5608} y^*_{t-1} + \underset{(0.0705)}{0.1468} y^*_{t-4}\,, & y_{t-1} > 24\,. \end{cases} \tag{7.1}$$

The plot of fitted values is displayed in Figure 2. For model diagnostics, the standardized deviance residuals are plotted in Figure 3. The fluctuations of deviance residuals around zero indicates that the fitting is adequate.

Some discussions about the estimation results are as follows. First, [49] asserts that IPO activities can be classified in to two regimes: "cold" and "hot" markets, depending on the volumes of the activities. Later, [68] propose a more sophisticated three-regimes classification in terms of "cold", "normal" and "hot"

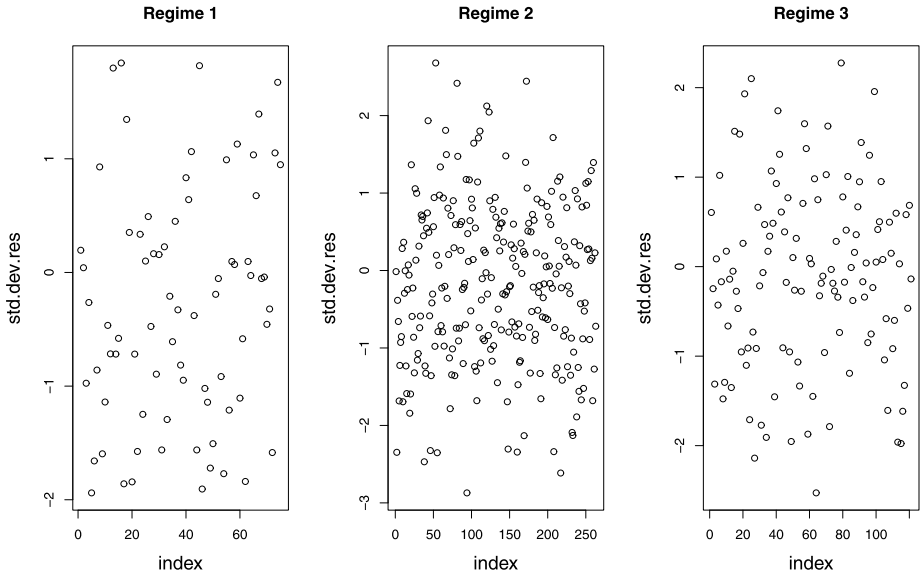

FIG 3. *Model diagnostic plot: standardized deviance residuals of net IPO series for the three estimated regimes.*

markets. The results in (7.1) suggest that the classification in [68] is more appropriate.

Second, (7.1) indicates that stock market return is an effective predictor of IPO volumes. In particular, the positive coefficients of $x^{(2)}$ in regimes 1 and 2 of (7.1) indicate that IPO activities are positively associated with stock market performances. This phenomenon agrees with the theory in [45] that high market returns increase the incentives of IPO issuance, thus contributing to IPO market activity as volumes soar. Moreover, the coefficient of $x^{(2)}$ is decreasing from regime 2 to regime 1, and becomes insignificant in regime 3, indicating that the positive effect of stock market returns diminishes with the increase in recent IPO activities. One possible explanation for this is as follows: as mentioned in [49], when the market is overactive, severe underpricing exists and discourages entrepreneurs. Hence, entrepreneurs choose to issue stocks in other periods, which offsets the market performance influence.

Acknowledgment

We would like to thank the Editor, an Associate Editor and an anonymous reviewer for their helpful comments and useful suggestions, which greatly improve the presentation of this paper. Li's research was supported by the Fundamental Research Funds for the Central Universities in UIBE (CXTD10-09). Yau's research was supported in part by grants from HKSAR-RGC-GRF Nos 405113, 14305517 and 14601015.

Appendix

Proof of Theorem 3.1 (strict stationarity and ergodicity)

First, we state the following definitions from Markov chain theory that provide the background for studying strict stationarity and ergodicity.

Definition 1. *Irreducibility* A Markov process $\{Y_t\}$ on a measurable space $\{\Omega, \mathcal{B}\}$ with transition probability $P^n(Y, A) = \text{pr}(Y_n \in A \mid Y_0 = y)$ is said to be μ-irreducible for a measure μ on $\mathcal{B}$ if $\sum_{n=1}^{\infty} P^n(y, A) > 0$ for all $y \in \Omega$ whenever $\mu(A) > 0$.

Definition 2. *Geometric ergodicity:* a Markov process $\{Y_t\}$ on $\{\Omega, \mathcal{B}\}$ is geometrically ergodic if there exists a probability measure π on $\mathcal{B}$ such that for all $A \in \mathcal{B}$ and $y \in \Omega$, there exist $\rho \in (0, 1)$ and $M_y > 0$ that

$$\|P^n(y, A) - \pi(A)\| \leq \rho^n M_y \,,$$

where $\|\cdot\|$ is the total-variation norm. This implies that $\{y_t\}$ is ergodic, β-mixing and has a unique stationary distribution π; see [4] and [52].

Definition 3. *Small set and petite set:* a set $C \in \mathcal{B}$ is said to be small if there exists an integer $m > 0$ and a non-trivial measure $v_m(\cdot)$ on $\mathcal{B}$ such that for all $y \in C$ and $A \in \mathcal{B}$, $P^m(y, A) \geq v_m(A)$.

Similarly, a set C is said to be petite for $\{y_t\}$ if there exists a probability measure $\gamma^*(\cdot)$ on $\mathbb{N}^+$ and a non-trivial measure $v_{\gamma^*}(\cdot)$ on $\mathcal{B}$ such that for all $y \in C$ and $A \in \mathcal{B}$, $\sum_{n=0}^{\infty} P^n(y, A)\gamma^*(n) \geq v_{\gamma^*}(A)$. Clearly, a small set is a petite set; see [43].

The proof of Theorem 3.1 relies on the following theorem about the ergodicity of Markov chains.

Theorem 7.1. *([43]) Let $n(y) : \Omega \to \mathbb{N}^+$ be an integer valued function. An irreducible chain $\{Y_t\}$ on Ω is geometrically ergodic if it is aperiodic and there exists a non-negative function $V \geq 1$ on Ω which is bounded on a petite set C, and for all $y \in \Omega$, there exist $\rho \in (0, 1)$ and $b \in (0, \infty)$ such that*

$$\begin{aligned} \text{E}[V(Y_{t+n(Y_t)}) \mid Y_t] &= \int P^{n(Y_t)}(Y_t, dY_{t+n(Y_t)})V(Y_{t+n(Y_t)}) \\ &\leq \rho^{n(Y_t)}[V(Y_t) + bI(Y_t \in C)] \,. \end{aligned} \tag{7.2}$$

By Condition 1a), denote

$$\begin{aligned} &p^* = \max\{p_{Y,1}, p_{Y,2}, \tilde{p}, p_e, p_\epsilon, q + d, 1\}, \quad \xi_t = (y_t, X_t, e_t, \epsilon_t), \\ &\Xi_t = \{(\xi_t, \ldots, \xi_{t-p^*+1})\} \,, \end{aligned}$$

and $\mathcal{F}_{t-1}$ as the sigma-field generated by $\{\Xi_{t-1}, \Xi_{t-2}, \ldots\}$. First, we show that $\{\Xi_t\}$ is Markovian under Condition 1a). Note that

$$\begin{aligned} \text{pr}(\{X_t, \ldots, X_{t-p^*+1}\} \mid \mathcal{F}_{t-1}) &= \text{pr}(\{X_t, \ldots, X_{t-p^*+1}\} \mid \{X_{t-1}, \ldots, X_{t-p^*}\}) \\ &= \text{pr}(\{X_t, \ldots, X_{t-p^*+1}\} \mid \Xi_{t-1}) \,. \end{aligned} \tag{7.3}$$

Also, as z_{t-d} is measurable with respect to the sigma-field generated by $\{X_{t-d}, y_{t-d}, \ldots, X_{t-d-q}, y_{t-d-q}\}$, we have that $\mathrm{pr}(z_{t-d} \mid \mathcal{F}_{t-1}) = \mathrm{pr}(z_{t-d} \mid \Xi_{t-1})$. Hence,

$$\begin{aligned}\mathrm{pr}(\Xi_t, z_{t-d} \mid \mathcal{F}_{t-1}) &= \mathrm{pr}(z_{t-d} \mid \mathcal{F}_{t-1})\mathrm{pr}(\Xi_t \mid \mathcal{F}_{t-1}, z_{t-d}) \\ &= \mathrm{pr}(z_{t-d} \mid \Xi_{t-1})\mathrm{pr}(\Xi_t \mid \Xi_{t-1}, z_{t-d}) = \mathrm{pr}(\Xi_t, z_{t-d} \mid \Xi_{t-1})\,.\end{aligned}$$

Integrating both sides with respective to the density of z_{t-d}, we have $\mathrm{pr}(\Xi_t \mid \mathcal{F}_{t-1}) = \mathrm{pr}(\Xi_t \mid \Xi_{t-1})$, and hence $\{\Xi_t\}$ is a Markov process.

Under Condition 1a′), we can analogously verify the Markovian property of $\{\Xi_t, z_t^*\}$. Denote $\mathcal{F}_{t-1}$ as the sigma-field generated by $\{\Xi_{t-1}, z_{t-1}^*, \Xi_{t-2}, z_{t-2}^*, \ldots\}$. As (7.3) and $\mathrm{pr}(z_t^* \mid \mathcal{F}_{t-1}) = \mathrm{pr}(z_t^* \mid \Xi_{t-1}, z_{t-1}^*)$ hold by Condition 1a′), we have

$$\begin{aligned}\mathrm{pr}(\Xi_t, z_t^* \mid \mathcal{F}_{t-1}) &= \mathrm{pr}(z_t^* \mid \mathcal{F}_{t-1})\mathrm{pr}(\Xi_t \mid \mathcal{F}_{t-1}, z_t^*) \\ &= \mathrm{pr}(z_t^* \mid \Xi_{t-1}, z_{t-1}^*)\mathrm{pr}(\Xi_t \mid \Xi_{t-1}, z_t^*) \\ &= \mathrm{pr}(z_t^* \mid \Xi_{t-1}, z_{t-1}^*)\mathrm{pr}(\Xi_t \mid \Xi_{t-1}, z_t^*, z_{t-1}^*) \\ &= \mathrm{pr}(\Xi_t, z_t^* \mid \Xi_{t-1}, z_{t-1}^*)\,.\end{aligned}$$

Thus, $\{\Xi_t, z_t^*\}$ is shown to be Markovian. Next, we illustrate our proof with $p_X = p_e = p_\epsilon = 0$ for simplicity. In this case, it suffices to prove stationarity and ergodicity of $\{Y_t\} = (y_t, \ldots, y_{t-p^*+1})$. We will show the geometric ergodicity of $\{y_t\}$ by using Theorem 7.1. Hence, we need to verify that $\{Y_t\}$ or $\{Y_t, Z_t^*\}$ is an irreducible and aperiodic Markov process and construct the corresponding function V. For the general case (except for the irreducibility when $p_e > 0$), the same verification methodologies could be applied on $\{\Xi_t\}$ or $\{\Xi_t, z_t^*\}$ with mild modifications, and hence the proof is omitted. Therefore, it suffices to show the stationarity and ergodicity of

$$\begin{cases} y_t = \sum_{i=1}^{r+1} \left[\omega_i + \sum_{l=1}^{p_{Y,1}} \psi_{y,1,i}^{(l)} g_{y,1}(y_{t-l}) + g_\lambda(\lambda_t, u_t) + \psi_{e,i}^{(0)} e_t\right] \\ \qquad \times I(z_{t-d} \in (\theta_{i-1}, \theta_i])\,, \\ \phi(\lambda_t) = \sum_{i=1}^{r+1} \left[\alpha_i + \sum_{j=1}^{p_{Y,2}} \psi_{y,2,i}^{(j)} g_{y,2}(y_{t-j}) + \psi_{\epsilon,i}^{(0)} \epsilon_t\right] I(z_{t-d} \in (\theta_{i-1}, \theta_i])\,. \end{cases} \tag{7.4}$$

Denote μ as the Lebesgue measure on $\mathbb{R}$, and μ^{p^*} as the Lebesgue measure on $\mathbb{R}^{p^*}$. Next, we show that $\{Y_t\}$ or $\{Y_t, Z_t^*\}$ is irreducible and aperiodic, and there exists some small set by the following proposition.

Proposition 1. *Under (7.4), we have:*

1. *Under Condition 1a), $\{Y_t\}$ is μ^{p^*}-irreducible and aperiodic. In addition, sets of the form $C = \{Y_t : |Y_t|_\infty \le c\}$ for some $c > 0$ is small for $\{Y_t\}$.*
2. *Under Condition 1a′), $\{Y_t, Z_t^*\}$ is $\mu^{p^*} \times \nu^q$-irreducible with some discrete measure ν on $\{0,1\}^r$ for r-manifolds of set $\{0,1\}$, and $\{Y_t, Z_t^*\}$ is aperiodic. In addition, sets of the form $C = \{(Y_t, Z_t^*) : |Y_t|_\infty \le c, |Z_t^*|_\infty \le q\}$ for some $c > 0$ is small for $\{Y_t, Z_t^*\}$.*

Proof of Proposition 1. We illustrate the proof of $\{Y_t\}$ under Condition 1a), where the proof of $\{Y_t, Z_t^*\}$ under condition 1a′) follows similar arguments. First we assume that e_t and ϵ_t have almost everywhere continuous and positive densities on $\mathbb{R}$. We divide the proof into three parts: irreducibility, existence of small set, and aperiodicity.

1) *Irreducibility.* As e_t has positive density on $\mathbb{R}$, y_t can reach any point on $\mathbb{R}$. Let $A \subset \mathbb{R}$ satisfies $0 < \mu(A) < \infty$, we first show that $\text{pr}(y_t \in A \mid Y_{t-1}, \lambda_t, u_t, z_{t-d})$ is positive for any realization $\{Y_{t-1}, \lambda_t, u_t, z_{t-d}\} = \{(y_{t-1}, \ldots, y_{t-p^*}), \lambda_t, u_t, z_{t-d}\}$ with $z_{t-d} \in (\theta_{i-1}, \theta_i]$. With respect to (7.4), we can construct

$$E_{t,i} = \left\{ e_t : \left[\alpha_i + \sum_{l=1}^{p_{Y,1}} \psi_{y,1,i}^{(l)} g_{y,1}(y_{t-l}) + g_\lambda(\lambda_t, u_t) + \psi_{e,i}^{(0)} e_t \right] \in A \mid Y_{t-1}, \lambda_t, u_t, z_{t-d} \right\}.$$

Hence, an injection exists between $E_{t,i}$ and A and thus $\mu(E_{t,i}) > 0$. The mapping from $E_{t,i}$ to A is surjective and $\mu(E_{t,i}) > 0$ holds. Denote the density of e_t as $f_e(\cdot)$, we have

$$\begin{aligned} \text{pr}(y_t \in A \mid Y_{t-1}, \lambda_t, u_t, z_{t-d}) &= \text{pr}\left(e_t \in E_{t,i} \mid Y_{t-1}, \lambda_t, u_t, z_{t-d}\right) \\ &= \int_{v \in E_{t,i}} f_e(v) dv \geq \inf_{e_t \in E_{t,i}} f_e(e_t) \mu(E_{t,i}) > 0\,. \end{aligned}$$

As e_t and ϵ_t have almost everywhere positive densities, the conditional densities $f_i(y_t \mid Y_{t-1}, \lambda_t, u_t, z_{t-d})$ and $h_i(\lambda_t \mid Y_{t-1}, z_{t-d})$ are thus almost everywhere positive. Therefore, from (2.1), the marginally density f is positive almost everywhere. Denote the marginal density of u_t and ϵ_t as $f_u(\cdot)$ and $f_\epsilon(\cdot)$, respectively, it follows that

$$\begin{aligned} &\text{pr}(y_t \in A \mid Y_{t-1}, z_{t-d}) \\ &= \int \text{pr}(y_t \in A \mid Y_{t-1}, \lambda_t, z_{t-d}) h_i(\lambda_t \mid Y_{t-1}, z_{t-d})\, d\lambda_t \\ &= \iint_0^1 \text{pr}(y_t \in A \mid Y_{t-1}, \lambda_t, u, z_{t-d}) f_u(u)\, du\, h_i(\lambda_t \mid Y_{t-1}, z_{t-d}) d\lambda_t \\ &> 0\,, \end{aligned} \tag{7.5}$$

for some set B with $\mu(B) > 0$, where $f_u(u) = 1$ on $[0, 1]$ and $z_{t-d} \in (\theta_{i-1}, \theta_i]$. Furthermore, denote $f_z(\cdot)$ as the density of z_t, we have, almost surely,

$$\text{pr}(y_t \in A \mid Y_{t-1}) = \int \text{pr}(y_t \in A \mid Y_{t-1}, z_{t-d}) f_z(z_{t-d} \mid Y_{t-1}) dz_{t-d} > 0\,. \tag{7.6}$$

Therefore, for any $\tilde{A} \subset \mathbb{R}^{p^*}$ and $y \in \mathbb{R}^{p^*}$ with Lebesgue measure $\mu_{p^*}(\tilde{A}) > 0$, as there exist some $A_1 \in \mathbb{R}$ and $A_2 \in \mathbb{R}^{p^*-1}$ such that $A_1 \times A_2 \subset \tilde{A}$ and $\mu(A_1) > 0$,

we have $P(Y_t \in \tilde{A}|Y_{t-1} = y) > P(y_t \in A_1|Y_{t-1} = y) > 0$. This completes the proof of irreducibility.

2) *Existence of a small set C.* By Definition 3, we need to construct a small set C such that, for all $Y_0 = y \in C \subset \mathbb{R}^{p^*}$ with $\mu_{p^*}(C) > 0$, there exists a non-trivial measure $v_m(\cdot)$ on the Borel sigma-field on $\mathbb{R}^{p^*}$, $\mathcal{B}$, such that for any $\tilde{A} \in \mathcal{B}$ with $\mu_{p^*}(\tilde{A}) > 0$,

$$P^m(y, \tilde{A}) \geq v_m(\tilde{A}) > 0\,.$$

We construct the set as $C = \{y : y = (y_1, \ldots, y_{p^*}) \in \mathbb{R}^{p^*}, |y|_\infty \leq c\}$ for some constant c, where $|y|_\infty = \max\{|y_1|, \ldots, |y_{p^*}|\}$. First, as the density of e_t is almost everywhere positive, $f(y_t \mid Y_{t-1}, \lambda_t, u_t, z_{t-d}) > 0$. Analogous to (7.5)–(7.6), we have that $f(y_t \mid Y_{t-1}) > 0$, and hence with (7.6),

$$\begin{aligned} \mathrm{pr}(y_{t+1} \in A \mid Y_{t-1}) &= \int \mathrm{pr}(y_{t+1} \in A \mid Y_{t-1}, y_t) f(y_t \mid Y_{t-1}) dy_t \\ &= \int \mathrm{pr}(y_{t+1} \in A \mid Y_t) f(y_t \mid Y_{t-1}) dy_t > 0\,, \end{aligned} \quad (7.7)$$

for any $A \subset \mathbb{R}$ with $\mu(A) > 0$. By using induction on (7.7), $\mathrm{pr}(y_{t+m-1} \in A \mid Y_{t-1}) > 0$ and hence $\mathrm{pr}(Y_{t+m-1} \in \tilde{A} \mid Y_{t-1}) > 0$ for all positive integer m and any $\tilde{A} \in \mathcal{B}$ with $\mu_{p^*}(\tilde{A}) > 0$. Thus, setting

$$v_m(\tilde{A}) = \min_{y \in C} \mathrm{pr}(y_{m-1} \in \tilde{A} \mid Y_0 = y) = \min_{y \in C} P^m(y, \tilde{A})\,,$$

we have $v_m(\tilde{A}) > 0$ by the compactness of C. Thus, the set C is verified as a small set.

3) *Aperiodicity.* From the existence proof of the small set C, it has been shown that $P^1(y, C) > 0$ and $P^2(y, C) > 0$ for all $y \in C$. By Proposition A1.1 of [9], it follows that $\{y_t\}$ is aperiodic.

To relax the assumption that ϵ_t and e_t have almost everywhere continuous and positive densities on $\mathbb{R}$, we extend to the case $e_t = \epsilon_t = 0$ for all t where e_t and ϵ_t do not have almost everywhere positive density. Define a perturbation $\{y_t^m\}$ of $\{y_t\}$ by

$$\begin{cases} y_t^m = \sum_{i=1}^{r+1} \left[\omega_i + \sum_{l=1}^{p_{Y,1}} \psi_{y,1,i}^{(l)} g_{y,1}(y_{t-l}^m) + g_\lambda(\lambda_t, u_t)\right] I(z_{t-d} \in (\theta_{i-1}, \theta_i]) \\ \qquad + \sigma_{1,m} e_{1,t}\,, \\ \phi(\lambda_t) = \sum_{i=1}^{r+1} \left[\alpha_i + \sum_{j=1}^{p_{Y,2}} \psi_{y,2,i}^{(j)} g_{y,2}(y_{t-j}^m)\right] I(z_{t-d} \in (\theta_{i-1}, \theta_i]) + \sigma_{2,m} e_{2,t}\,, \end{cases} \quad (7.8)$$

where $\sigma_{1,m}, \sigma_{2,m} > 0, \sigma_{1,m} \to 0$ and $\sigma_{2,m} \to 0$ as $m \to \infty$, and $e_{1,t}, e_{2,t}$ are i.i.d. zero-mean noises with finite first moment and almost everywhere positive densities. Using the perturbation techniques in [22], $\{y_t^m\}$ is irreducible, aperiodic, and a small set exists. Thus we can derive the strict stationarity and ergodicity of $\{y_t^m\}$ by Theorem 7.1. Since $\{y_t^m\}$ converges almost surely to $\{y_t\}$ as $m \to \infty$, the proof is complete. □

Proof of Lemma 3.1. Lemma 3.1 is established in the irreducibility part of the proof of Proposition 1. □

Proof of Theorem 3.1. First, we give the proof with respect to $\{Y_t\}$ under Condition 1a). Recall that the irreducibility, aperiodicity, and existence of a small set $C = \{y : y = (y_1, \ldots, y_{p^*}) \in \mathbb{R}^{p^*}, |y|_\infty \le c\}$ have been established in Proposition 1. From Theorem 7.1, it suffices to construct a function V that satisfies (7.2). For the model (7.4), let $Y_{t-1} = (y_{t-1}, \ldots, y_{t-p^*})$ and define

$$V(Y_{t-1}) = 1 + |Y_{t-1}|_\infty \,, \tag{7.9}$$

where $|Y_{t-1}|_\infty = \max\{|y_{t-1}|, \ldots, |y_{t-p^*}|\}$. Conditional on Y_{t-1}, we have

$$\begin{aligned}
\mathrm{E}(|y_t| \mid Y_{t-1}) \le \mathrm{E}\Bigg[& \left\{ \sum_{i=1}^{r+1} |\omega_i| + \sum_{l=1}^{p_{Y,1}} \left|\psi_{y,1,i}^{(l)}\right| |g_{y,1}(y_{t-l})| + |g_\lambda(\lambda_t, u_t)| + \left|\psi_{e,i}^{(0)} e_t\right| \right\} \\
& \times I(z_{t-d} \in (\theta_{i-1}, \theta_i]) \mid Y_{t-1} \Bigg] \\
\le & \max_i \left\{ \sum_{l=1}^{p_{Y,1}} \left|\psi_{y,1,i}^{(l)}\right| |g_{y,1}(y_{t-l})| + \mathrm{E}(|g_\lambda(\lambda_t, u_t)| \mid Y_{t-1}) \right\} \\
& + \max_i \left(|\omega_i| + \mathrm{E}\left|\psi_{e,i}^{(0)} e_t\right| \right) \,.
\end{aligned} \tag{7.10}$$

Denote $y^* = \arg\max_{j=1\ldots p_{Y,2}} g_{y,2}(y_{t-j})$. For the case of concave $\phi(\cdot)$, Condition 1b) implies that the inverse function $\phi^{-1}(\cdot)$ exists, and is strictly increasing and convex. Hence, as $\max_i \sum_{j=1}^{p_{Y,2}} \left|\psi_{y,2,i}^{(j)}\right| < 1$,

$$\begin{aligned}
\lambda_t &= \phi^{-1}\left\{ \sum_{i=1}^{r+1} \left(\alpha_i + \sum_{j=1}^{p_{Y,2}} \psi_{y,2,i}^{(j)} g_{y,2}(y_{t-j}) + \psi_{\epsilon,i}^{(0)} \epsilon_t \right) I(z_{t-d} \in (\theta_{i-1}, \theta_i]) \right\} \\
&\le \phi^{-1}\left\{ \max_i \sum_{j=1}^{p_{Y,2}} \left|\psi_{y,2,i}^{(j)}\right| g_{y,2}(y^*) + \max_i \left(\alpha_i + \psi_{\epsilon,i}^{(0)} |\epsilon_t| \right) \right\} \\
&\le \left(1 - \max_i \sum_{j=1}^{p_{Y,2}} \left|\psi_{y,2,i}^{(j)}\right| \right) \phi^{-1}\left\{ \frac{\max_i \left(\alpha_i + \psi_{\epsilon,i}^{(0)} |\epsilon_t| \right)}{1 - \max_i \sum_{j=1}^{p_{Y,2}} \left|\psi_{y,2,i}^{(j)}\right|} \right\} \\
&\quad + \max_i \sum_{j=1}^{p_{Y,2}} \left|\psi_{y,2,i}^{(j)}\right| \phi^{-1}(g_{y,2}(y^*)) \\
&\le \left(1 - \max_i \sum_{j=1}^{p_{Y,2}} \left|\psi_{y,2,i}^{(j)}\right| \right) \phi^{-1}\left\{ \frac{\max_i \left(\alpha_i + \psi_{\epsilon,i}^{(0)} |\epsilon_t| \right)}{1 - \max_i \sum_{j=1}^{p_{Y,2}} \left|\psi_{y,2,i}^{(j)}\right|} \right\} \\
&\quad + \max_i \sum_{j=1}^{p_{Y,2}} \left|\psi_{y,2,i}^{(j)}\right| \max\{b_2 |Y_{t-1}|_\infty, G_2\} \,,
\end{aligned}$$

where the second last inequality follows from Jensen's inequality, and the last inequality follows from Condition 1c). Let $\mathcal{F}_{t-1}$ be the σ-field generated by $\{y_s, \epsilon_s\}_{s\le t-1}$. It follows from $\mathrm{E}\{\phi^{-1}(\epsilon_t)\} < \infty$ that

$$
\begin{aligned}
&\mathrm{E}(|g_\lambda(\lambda_t, u_t)| \mid Y_{t-1}) \\
&\le \mathrm{E}\left[\left| g_\lambda\left\{\phi^{-1}\left(\max_i\left(\alpha_i + \sum_{j=1}^{p_{Y,2}} \psi_{y,2,i}^{(j)} g_{y,2}(y^*) + \psi_{\epsilon,i}^{(0)}\epsilon_t\right)\right), u_t\right\}\right| \mid Y_{t-1}\right] \\
&= \mathrm{E}\left[\mathrm{E}\left[\left| g_\lambda\left\{\phi^{-1}\left(\max_i\left(\alpha_i + \sum_{j=1}^{p_{Y,2}} \psi_{y,2,i}^{(j)} g_{y,2}(y^*) + \psi_{\epsilon,i}^{(0)}\epsilon_t\right)\right), u_t\right\}\right| \right.\right. \\
&\qquad \left.\left. \mid \mathcal{F}_{t-1}\right] \mid Y_{t-1}\right] \\
&\le \mathrm{E}\left[\max\left\{H, b_\lambda\left(1 - \max_i \sum_{j=1}^{p_{Y,2}} \left|\psi_{y,2,i}^{(j)}\right|\right)\phi^{-1}\left(\frac{\max_i\left(\alpha_i + \psi_{\epsilon,i}^{(0)}|\epsilon_t|\right)}{1 - \max_i \sum_{j=1}^{p_{Y,2}}\left|\psi_{y,2,i}^{(j)}\right|}\right)\right.\right. \\
&\qquad \left.\left. + b_\lambda \max_i \sum_{j=1}^{p_{Y,2}} \left|\psi_{y,2,i}^{(j)}\right| \max\{b_2|Y_{t-1}|_\infty, G_2\}\right\} \mid Y_{t-1}\right] \\
&\le b_\lambda b_2 \max_i \sum_{j=1}^{p_{Y,2}} \left|\psi_{y,2,i}^{(j)}\right| |Y_{t-1}|_\infty + \widetilde{H}^*, \qquad (7.11)
\end{aligned}
$$

for some constant $\widetilde{H}^*$. Equipped with (7.11), it can be shown from (7.10) that

$$
\begin{aligned}
\mathrm{E}(|y_t| \mid Y_{t-1}) &\le \max_i\left(\sum_{l=1}^{p_{Y,1}} \left|\psi_{y,1,i}^{(l)}\right| |g_{y,1}(y_{t-l})| + b_\lambda b_2 \sum_{j=1}^{p_{Y,2}} \left|\psi_{y,2,i}^{(j)}\right| |Y_{t-1}|_\infty\right) \\
&\quad + \widetilde{H}^* + \max_i\left(|\omega_i| + \mathrm{E}\left|\psi_{e,i}^{(0)} e_t\right|\right) \\
&\le H^* + \max_i \rho_i(1)|Y_{t-1}|_\infty = H^* + \tilde{\rho}|Y_{t-1}|_\infty, \qquad (7.12)
\end{aligned}
$$

where $\tilde{\rho} = \max_i \rho_i(1) \in (0,1)$ and H^* is some constant. Using (7.12), we have

$$
\begin{aligned}
\mathrm{E}(|y_{t+1}| \mid Y_{t-1}) &= \mathrm{E}\{\mathrm{E}(|y_{t+1}| \mid y_t, Y_{t-1}) \mid Y_{t-1}\} \\
&\le H^* + \tilde{\rho}\max\{|Y_{t-1}|_\infty, \mathrm{E}(|y_t| \mid Y_{t-1})\} \\
&\le (1+\tilde{\rho})H^* + \tilde{\rho}|Y_{t-1}|_\infty .
\end{aligned}
$$

Arguing inductively, for $Y_{t+p^*-1} = (y_{t+p^*-1}, \ldots, y_t)$, we have

$$
\mathrm{E}(|Y_{t+p^*-1}|_\infty \mid Y_{t-1}) \le (1 + \tilde{\rho} + \ldots + \tilde{\rho}^{p^*-1})H^* + \tilde{\rho}|Y_{t-1}|_\infty . \qquad (7.13)
$$

Select $\rho \in (\tilde{\rho}^{1/p^*}, 1)$. Then, taking $n(Y_{t-1}) = p^*$, $b = [(1 + \tilde{\rho} + \cdots + \tilde{\rho}^{p^*-1})H^* + 1]/\rho^{p^*} - 1$ and $c = \max\{\tilde{y}, b/(\rho^{p^*} - \tilde{\rho})\}$ for the small set C, it can be verified that

V in (7.9) satisfies (7.2). By Theorem 7.1, $\{y_t\}$ is geometric ergodicity. Thus, by Definition 2, $\{y_t\}$ is strict stationary and ergodic.

For the case that $\phi(\cdot)$ is a polynomial with order γ, we have $\phi(\lambda) = \lambda^\gamma(1 + o(1))$ and so $\phi^{-1}(v) = v^{1/\gamma}(1 + o(1))$. For $\gamma \geq 1$,

$$\begin{aligned}
\lambda_t &= \phi^{-1}\left\{\sum_{i=1}^{r+1}\left(\alpha_i + \sum_{j=1}^{p_{Y,2}} \psi_{y,2,i}^{(j)} g_{y,2}(y_{t-j}) + \psi_{\epsilon,i}^{(0)}\epsilon_t\right) I(z_{t-d} \in (\theta_{i-1}, \theta_i])\right\} \\
&\leq \left\{\max_i \sum_{j=1}^{p_{Y,2}} \left|\psi_{y,2,i}^{(j)}\right| g_{y,2}(y^*) + \max_i\left(\alpha_i + \psi_{\epsilon,i}^{(0)}|\epsilon_t|\right)\right\}^{1/\gamma} (1 + o(1)) \\
&\leq \left\{\max_i \sum_{j=1}^{p_{Y,2}} \left|\psi_{y,2,i}^{(j)}\right|^{1/\gamma} (g_{y,2}(y^*))^{1/\gamma} + \max_i\left(\alpha_i + \psi_{\epsilon,i}^{(0)}|\epsilon_t|\right)^{1/\gamma}\right\}(1 + o(1)) \\
&\leq \left\{\max_i \sum_{j=1}^{p_{Y,2}} \left|\psi_{y,2,i}^{(j)}\right|^{1/\gamma} \max\{b_2|Y_{t-1}|_\infty, G_2\} + \phi^{-1}\left(\max_i\left(\alpha_i + \psi_{\epsilon,i}^{(0)}|\epsilon_t|\right)\right)\right\} \\
&\qquad \times (1 + o(1)), \quad (7.14)
\end{aligned}$$

where the last inequality follows from Condition 1c) and $\phi^{-1}(v) = v^{1/\gamma}(1 + o(1))$. The $(1 + o(1))$ term in (7.14) is negligible if $|Y_{t-1}|_\infty$ is large. Using the same arguments as in the derivation of (7.11), (7.12) and (7.13), the geometric ergodicity can be derived similarly.

The geometric ergodicity of $\{Y_t, Z_t^*\}$ under condition 1a′) can be shown analogously: By Proposition 1, sets of the form $C = \{(Y_t, Z_t^*) : |Y_t|_\infty \leq c, |Z_t^*|_\infty \leq q\}$ for some $c > 0$ are small for $\{Y_t, Z_t^*\}$. Define $V(Y_t, Z_t^*) = 1 + |Y_t|_\infty + |Z_t^*|_\infty$, we have that (7.13) holds by the same derivation. Then, since $0 \leq |Z_t^*|_\infty \leq 1$, by selecting $b = [(1 + \tilde{\rho} + \cdots + \tilde{\rho}^{p^*-1})H^* + 1 + q]/\rho^{p^*} - 1$, $c = \max\{\tilde{y}, b/(\rho^{p^*} - \tilde{\rho})\}$, with other parameters the same as in the proof of $\{Y_t\}$ under condition 1a), the geometric ergodicity of $\{Y_t, Z_t^*\}$ is verified by Theorem 7.1. Hence, $\{Y_t, Z_t^*\}$ is strictly stationary and ergodic, and so does y_t. This finishes the proof. □

Proof of Corollary 1. Denote

$$\rho_z(\gamma, z_{t-d}) = \sum_{i=1}^{r+1}\sum_{j=1}^{p_{Y,2}} \left|\psi_{y,2,i}^{(j)}\right|^{1/\gamma} I(z_{t-d} \in (\theta_{i-1}, \theta_i]).$$

Let $\mathcal{F}_{t-1}$ be the σ-field generated by $\{y_s, \epsilon_s\}_{s \leq t-1}$. From the independence of z_{t-d} and $\{y_s\}_{s<t}$, for concave ϕ in 1′), we have

$$\begin{aligned}
&\mathrm{E}(|g_\lambda(\lambda_t, u_t)| \mid Y_{t-1}) \\
&= \mathrm{E}\left[\mathrm{E}\left[\left|g_\lambda\left\{\phi^{-1}\left(\sum_{i=1}^{r+1}\left(\alpha_i + \sum_{j=1}^{p_{Y,2}} \psi_{y,2,i}^{(j)} g_{y,2}(y_{t-j}) + \psi_{\epsilon,i}^{(0)}\epsilon_{t-q}\right)\right.\right.\right.\right.
\end{aligned}$$

$$\times I(z_{t-d} \in (\theta_{i-1}, \theta_i])\Big), u_t\Big\} \Big| \mid \mathcal{F}_{t-1}\Big] \mid Y_{t-1}\Big]$$

$$\leq \mathrm{E}\left[\max\left\{H, b_\lambda\left(1-\rho_z(1,z_{t-d})\right)\phi^{-1}\left(\frac{\max_i\left(\alpha_i+\psi^{(0)}_{\epsilon,i}|\epsilon_{t-q}|\right)}{1-\rho_z(1,z_{t-d})}\right)\right.\right.$$

$$\left.\left. + b_\lambda\rho_z(1,z_{t-d})\max\{b_2|Y_{t-1}|_\infty, G_2\}\right\} \mid Y_{t-1}\right]$$

$$\leq b_\lambda b_2 \sum_{i=1}^{r+1}\left\{\sum_{j=1}^{p_{Y,2}}\left|\psi^{(j)}_{y,2,i}\right| |Y_{t-1}|_\infty \mathrm{pr}(z_{t-d}\in(\theta_{i-1},\theta_i])\right\} + \widetilde{H}^*_2\,, \tag{7.15}$$

for some constant $\widetilde{H}^*_2$. Thus, it is straightforward to bound $\mathrm{E}(|y_t| \mid Y_{t-1})$ by

$$\begin{aligned}
&\mathrm{E}(|y_t| \mid Y_{t-1})\\
&\leq \sum_{i=1}^{r+1}\left[\left\{\sum_{l=1}^{p_{Y,1}}\left|\psi^{(l)}_{y,1,i}\right| |g_{y,1}(y_{t-l})| + b_\lambda b_2 \sum_{j=1}^{p_{Y,2}}\left|\psi^{(j)}_{y,2,i}\right| |Y_{t-1}|_\infty\right\}\mathrm{pr}(z_{t-d}\in(\theta_{i-1},\theta_i])\right]\\
&\quad + \widetilde{H}^*_2 + \max_i\left(|\omega_i| + \mathrm{E}\left|\psi^{(0)}_{e,i} e_{t-m}\right|\right)\\
&\leq H^*_2 + \sum_{i=1}^{r+1}\left[\left\{b_1\sum_{l=1}^{p_{Y,1}}\left|\psi^{(l)}_{y,1,i}\right| + b_\lambda b_2\sum_{j=1}^{p_{Y,2}}\left|\psi^{(j)}_{y,2,i}\right|\right\}\mathrm{pr}(z_{t-d}\in(\theta_{i-1},\theta_i])\right]|Y_{t-1}|_\infty\\
&= H^*_2 + \left\{\sum_{i=1}^{r+1}\rho_i(1)\mathrm{pr}(z_{t-d}\in(\theta_{i-1},\theta_i])\right\}|Y_{t-1}|_\infty\\
&\leq H^*_2 + \tilde{\rho}|Y_{t-1}|_\infty\,,
\end{aligned}$$

for some $\tilde{\rho}\in(0,1)$ and some constants H^*_2. Using similar arguments in the proof of Theorem 3.1, $\{y_t\}$ is geometrically ergodic and hence strictly stationary and ergodic. If $\phi(\cdot)$ is a polynomial with order $\gamma \geq 1$ in 2′), then by considering $\rho_z(\gamma, z_{t-d})$ instead of $\rho_z(1, z_{t-d})$, similar derivations yield an analogous result to (7.15), with $|\psi^{(j)}_{y,2,i}|$ replaced by $|\psi^{(j)}_{y,2,i}|^{1/\gamma}$. Thus, the strict stationarity and ergodicity are verified. □

Proof of Corollary 2. First consider concave ϕ. For simplicity, we first illustrate the proof for $d = 1$ and assume $\upsilon = \max_{i=2,\ldots,r}\rho_i(1) > 1$. Using similar arguments as in (7.12), $\mathrm{E}(|y_t| \mid Y_{t-1}) \leq H^* + \upsilon|Y_{t-1}|_\infty$ for some H^*. Therefore,

$$\begin{aligned}
&\mathrm{E}(|y_{t+1}| \mid Y_{t-1})\\
=&\mathrm{E}\{\mathrm{E}(|y_{t+1}| \mid y_t, Y_{t-1}) \mid Y_{t-1}\}\\
\leq&\mathrm{E}\left[\sum_{i=1}^{r+1}\rho_i(1)I(y_t\in(\theta_{i-1},\theta_i])\max\{|Y_{t-1}|_\infty, |y_t|\} \mid Y_{t-1}\right] + H^*
\end{aligned}$$

$$
\begin{aligned}
&\leq \Big[\max\{\rho_1(1), \rho_{r+1}(1)\} \mathrm{pr}(\{y_t > \theta_r\} \cup \{y_t \leq \theta_1\} \mid Y_{t-1}) \\
&\quad + \max_{i=2,\ldots,r} \rho_i(1) \mathrm{pr}(\theta_1 < y_t \leq \theta_r \mid Y_{t-1}) \Big] \\
&\qquad \times \max\{\mathrm{E}(|y_t| \mid Y_{t-1}), |Y_{t-1}|_\infty\} + H^* \\
&\leq H^* + [\max\{\rho_1(1), \rho_{r+1}(1)\}(1-\pi_y) + \upsilon\pi_y] \, (H^* + |\upsilon Y_{t-1}|_\infty) \\
&\leq (1+\tilde{\rho}) H^* + \tilde{\rho}\upsilon |Y_{t-1}|_\infty \,, \qquad (7.16)
\end{aligned}
$$

for some $\tilde{\rho} \in (0,1)$. Arguing inductively, analogous to (7.16), for all $k_1 \geq 0$, $0 \leq k_2 \leq p^* - 1$,

$$
\mathrm{E}(|y_{t+k_1 p^* + k_2 + 1}| \mid Y_{t-1}) \leq (1 + \tilde{\rho} + \cdots + \tilde{\rho}^{k_1 p^* + k_2 + 1}) H^* + \tilde{\rho}^{k_1+1} \upsilon |Y_{t-1}|_\infty \,.
$$

As $\tilde{\rho}^{k_1+1}\upsilon < 1$ for sufficiently large k_1, the geometric ergodicity of $\{y_t\}$ can be established using similar arguments in the proof in Theorem 3.1. When ϕ is a polynomial of order $\gamma \geq 1$, we replace $\rho_i(\gamma)$ by $\rho_i(1)$ to derive the result analogous to (7.16). Using similar but more tedious arguments, the results can be derived for $d = 2, \ldots, p^*$ that, $k_1 \geq 0$, $0 \leq k_2 \leq p^* - 1$,

$$
\begin{aligned}
\mathrm{E}(|y_{t+k_1 p^* + k_2 + d}| \mid Y_{t-1}) \leq (1 + \tilde{\rho} + \cdots + \tilde{\rho}^{k_1 p^* + k_2 + 1})(1 + \upsilon + \cdots + \upsilon^{d-1}) H^* \\
+ \tilde{\rho}^{k_1+1} \upsilon^d |Y_{t-1}|_\infty \,.
\end{aligned}
$$

Again, as $\tilde{\rho}^{k_1+1}\upsilon^d < 1$ for sufficiently large k_1, the geometric ergodicity of $\{y_t\}$ follows. □

Proof of Corollary 3. For the TARMA model, the threshold variable y_{t-d} is clearly measurable with respect to the sigma-field generated by $\{X_t, y_t, \ldots, X_{t-q}, y_{t-q}\}$ with $q \geq d$. Let $b_1 = 1$, $b_2 = 0$ and $G_1 = \tilde{y} = 1$. It follows that $|g_{y,1}(y)| \leq b_1|y|$ for all $|y| > \tilde{y}$, and $g_{y,1}(y) \leq G_1$ for all $|y| \leq \tilde{y}$. Together with the irreducibility from Condition 1f), Condition 1 holds. Therefore, by Theorem 3.1 and the fact that $\psi^{(j)}_{y,2,i} = 0$ for all i, the TARMA process is stationary and ergodic if $\max_{i=1,\ldots,r+1} \rho_i(1) = \max_{i=1,\ldots,r+1} \sum_{l=1}^{p_{Y,1}} |\psi^{(l)}_{y,1,i}| < 1$. The same arguments apply to TAR model, with the irreducibility condition guaranteed by Lemma 3.1 instead of Condition 1f). □

Verification of Assumption 4b)

In this section we illustrate the verification of conditions in Assumption 4b) using the model in Example 1 of the simulation studies. First, we have $T(y_t) = y_t$. As Theorem 3.1 indicates that $\{y_t\}$ is strictly stationary and ergodic, $y_t < \infty$ almost surely. Denote $Y_t = (y_t, \ldots, y_{t-3})$ and let $|\cdot|_\infty$ be the infinite norm such that $|Y_t|_\infty = \max\{|y_t|, \ldots, |y_{t-3}|\}$. Note that $\{Y_t\}$ is a Markov process. By (6.1),

$$
\log(\lambda_t) \leq \max_i a_i + \max_i \sum_{j=1}^{p_i} |\psi_{i,j}| \log(|Y_{t-1}|_\infty + 1) \,,
$$

and hence

$$\lambda_t \le e^{\max_i a_i}(|Y_{t-1}|_\infty + 1)^{\max_i \sum_{j=1}^{p_i} |\psi_{i,j}|} \le K_\lambda |Y_{t-1}|_\infty^\rho + c_\lambda\,, \tag{7.17}$$

for some $0 < \rho < 1$ and positive constants K_λ and c_λ. Thus, $\lambda_t < \infty$ almost surely, and $\mathrm{var}(T(y_t) \mid \lambda_t) < \infty$ almost surely.

Next we show the existence of moments of y_t and λ_t. By strict stationarity and ergodicity, we have $\mathrm{E}(y_t) = \mathrm{E}(\lambda_t) < \infty$ and hence $\mathrm{E}(y_t^k) < \infty$ and $\mathrm{E}(\lambda_t^k) < \infty$ for all $k \in (0, 1]$. For $k > 1$, we denote $\tilde{y}_t = y_t^k$ and analogous to [22], we consider a sequence of perturbation $\{\tilde{y}_t^m\}$ given by

$$\begin{cases} \tilde{y}_t^m = y_t^k + c_m \epsilon_t\,, \quad y_t \sim \mathrm{Pois}(\lambda_t)\,, \\ \log(\lambda_t) = \sum_{i=1}^{r+1} \left[\alpha_i + \sum_{j=1}^{p_{Y,2}} \psi_{y,i}^{(j)} \log((\tilde{y}_{t-j}^m)^{1/k} + 1)\right] I((\tilde{y}_{t-4}^m)^{1/k} \in (\theta_{i-1}, \theta_i])\,, \end{cases}$$

where $\{\epsilon_t\}$ are i.i.d. uniform random variables taking values on $(0, 1)$, $\{c_m\}$ is a sequence of positive constants converging to 0. Thus, we may utilize Theorem 7.1 for proving geometric ergodicity of $\tilde{Y}_t^m = (\tilde{y}_t^m, \ldots, \tilde{y}_{t-3}^m)$, and show the existence of kth moment of $\tilde{y}_t^m$. As y_t follows Poisson distribution with mean λ_t, we have $\mathrm{E}(y_t^k \mid \lambda_t) \le c_k \lambda_t^k + b_k$ for all $k > 1$ and some positive constants c_k and b_k. Together with (7.17),

$$\mathrm{E}(y_t^k \mid \tilde{Y}_{t-1}^m) \le c_k \mathrm{E}(\lambda_t^k \mid \tilde{Y}_{t-1}^m) + b_k \le c_k (K_\lambda |\tilde{Y}_{t-1}^m|_\infty^{\rho/k} + c_\lambda)^k + b_k\,.$$

As $\rho \in (0, 1)$, for any constant c_1 and c_2, we have $c_1 y^\rho + c_2 < \rho y$ for all sufficiently large y. Therefore,

$$\begin{aligned} c_k (K_\lambda |\tilde{Y}_{t-1}^m|_\infty^{\rho/k} + c_\lambda)^k &= (c_k^{1/k} K_\lambda |\tilde{Y}_{t-1}^m|_\infty^{\rho/k} + c_k^{1/k} c_\lambda)^k \le (\rho |\tilde{Y}_{t-1}^m|_\infty^{1/k})^k \\ &= \rho^k |\tilde{Y}_{t-1}^m|_\infty\,, \end{aligned}$$

for all $|\tilde{Y}_{t-1}^m|_\infty^{1/k} > y^*$ with some sufficiently large y^*. Meanwhile, there exists a sufficiently large constant H_k such that $c_k (K_\lambda |\tilde{Y}_{t-1}^m|_\infty^{\rho/k} + c_\lambda)^k \le H_k$ for all $|\tilde{Y}_{t-1}^m|_\infty^{1/k} \le y^*$. Hence,

$$\mathrm{E}(y_t^k \mid \tilde{Y}_{t-1}^m) \le \rho^k |\tilde{Y}_{t-1}^m|_\infty + H_k + b_\lambda = \rho^k |\tilde{Y}_{t-1}^m|_\infty + H_k^*\,,$$

where $H_k^* = H_k + b_\lambda$. Thus, we have

$$\mathrm{E}(|\tilde{y}_t^m| \mid \tilde{Y}_{t-1}^m) = \mathrm{E}(y_t^k \mid \tilde{Y}_{t-1}^m) + c_m \mathrm{E}|\epsilon_t| \le \rho^k |\tilde{Y}_{t-1}^m|_\infty + H_k^* + c_m/2\,.$$

By the induction arguments as in (7.12) and (7.13), we have

$$\mathrm{E}(|\tilde{Y}_{t+3}^m|_\infty \mid \tilde{Y}_{t-1}^m) \le (1 + \rho^k + \ldots + \rho^{3k})(c_m/2 + H_k^*) + \rho^k |\tilde{Y}_{t-1}^m|_\infty\,. \tag{7.18}$$

Analogous to (7.13), (7.18) implies (7.2) with $V(\tilde{Y}_{t-1}^m) = |\tilde{Y}_{t-1}^m|_\infty + 1$, and hence the geometric ergodicity of $\{\tilde{y}_t^m\}$ is verified by Theorem 7.1, which further implies $\mathrm{E}(\tilde{y}_t^m) < \infty$. Then, as $c_m \to 0$, $\tilde{y}_t^m \to \tilde{y}_t$ almost surely and hence

$\mathrm{E}(\tilde{y}_t) = \mathrm{E}(y_t^k) < \infty$ for all $k > 1$. In conclusion, $\mathrm{E}(y_t^k) < \infty$ and $\mathrm{E}(\lambda_t^k) < \infty$ for all $k > 0$.

With the existence of all moments of y_t and λ_t, we are ready to verify the remaining conditions. As $\log[\mathrm{E}(e^{uT(y_t)} \mid \lambda_t)] = \lambda_t(e^u - 1)$, by choosing $c_t = \lambda_t(e^u - 1)$, we have $\limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^n c_t^{3/2} < \infty$. Next, choosing

$$\Gamma(Y_{t-1}; y_t) = y_t \sum_{j=1}^{2} \log(y_{t-j} + 1) + e^{K\sum_{j=1}^{2}\log(y_{t-j}+1)} \sum_{j=1}^{2} \log(y_{t-j} + 1)\,,$$

with some $K > 0$, we have $\mathrm{E}[\Gamma^2(Y_{t-1}; y_t)] < \infty$ since $\mathrm{E}(y_t^{2K})$ is finite for all $K > 0$. Therefore, all of the conditions in Assumption 4b) are satisfied.

Proof of Theorems 5.1, 5.2, 5.3 and 5.4 (asymptotic theory of inferences)

From (4.3), the minimum description length is a sum of the negative log-likelihood and the penalties on model complexity. We define

$$\begin{aligned}\mathrm{Pen}(\Theta, p, d) \quad = \quad & \log_2(r) + \log_2(d) + \frac{1}{2}\sum_{i=1}^{r} \log_2(n_i) + \sum_{i=1}^{r+1} \log_2(p_i' + 4) \\ & + \sum_{i=1}^{r+1} \frac{p_i' + 4}{2} \log_2(n_i)\,, \end{aligned} \tag{7.19}$$

as the penalty on the model complexity. Similar to [66], we show the following propositions:

Proposition 2. *Let $\theta_l < \theta_u$ be constants satisfying $(\theta_l, \theta_u) \subseteq B_n = (\theta_{i-1}^0 - k_n^1, \theta_i^0 + k_n^2)$, where $\{k_n^1\}$ and $\{k_n^2\}$ are two positive sequences that converging to 0. Define*

$$L_{n,i}(\Psi_i, \theta_l, \theta_u, d) = \frac{1}{n}\sum_{t=1}^{n} l(\Psi_i; y_t, Y_{t-1}, X_t) I(z_{t-d} \in (\theta_l, \theta_u])\,.$$

In addition, let $L_i = \lim_{n\to\infty} \mathrm{E}(L_{n,i})$, and $L_i^{(k)}$, $L_{n,i}^{(k)}$ be the k^{th} derivative with respect to Ψ_i for L_i, $L_{n,i}$, respectively. Here $L_i^{(0)} = L_i$, $L_{n,i}^{(0)} = L_{n,i}$. Then, under Assumptions 1–3,

$$\sup_{\theta_l, \theta_u \in B_n} \sup_{\Psi_i} |L_i^{(k)}(\Psi_i, \theta_l, \theta_u, d) - L_{n,i}^{(k)}(\Psi_i, \theta_l, \theta_u, d)| \longrightarrow 0 \quad \text{a.s.}\,,$$

for $k = 0, 1, 2$, $i = 1, \ldots, r+1$.

Proof of Proposition 2. We show the case $k = 0$ as an illustrative example. The proofs for $k = 1$ or 2 are similar. First, by the compactness of the parameter space $S(\Psi_i^*)$ of Ψ_i and the ergodic theorem, for any pair of $(\theta_l, \theta_u) \subseteq [\theta_{i-1}^0, \theta_i^0]$,

$$\sup_{\Psi_i \in S(\Psi_i^*)} |L_i(\Psi_i, \theta_l, \theta_u, d) - L_{n,i}(\Psi_i, \theta_l, \theta_u, d)| \longrightarrow 0 \quad \text{a.s.}\,. \tag{7.20}$$

For discrete z_t, (7.20) holds for any combination of (θ_l, θ_u) from observed z_t values. Thus, Proposition 2 holds.

For continuous z_t, (7.20) holds particularly for all subintervals of rational endpoints. Hence for $\theta \in (\theta_{i-1}^0, \theta_i^0)$ and any $\epsilon > 0$, there exists a rational number $w < \theta$ such that

$$\begin{aligned} & \sup_{\Psi_i \in S(\Psi_i^*)} |L_i(\Psi_i, \theta_{i-1}^0, \theta, d) - L_{n,i}(\Psi_i, \theta_{i-1}^0, w, d)| \\ \leq & \sup_{\Psi_i \in S(\Psi_i^*)} \left| \frac{1}{n} \sum_{t=1}^{n} l(\Psi_i; y_t, Y_{t-1}, X_t) I(w < z_{t-d} \leq \theta) \right| + \epsilon/3 \\ \leq & \sup_{\Psi_i \in S(\Psi_i^*)} |\mathrm{E}[l(\Psi_i; y_t, Y_{t-1}, X_t)] Q(w, \theta) + \epsilon/3| + \epsilon/3 \\ < & \epsilon \,, \end{aligned}$$

where $Q(w, \theta) = \mathrm{E}(I(w < z_{t-d} \leq \theta))$. Selecting w close to θ ensures a sufficiently small $Q(w, \theta)$. As the pair (θ_l, θ_u) is closely approximated by some subset with rational number endpoints, (7.20) holds uniformly on all $(\theta_l, \theta_u) \subseteq [\theta_{i-1}^0, \theta_i^0]$. Furthermore, if θ_l or θ_u is outside $[\theta_{i-1}^0, \theta_i^0]$, the almost sure convergence still holds if they are within a shrinkage neighborhood. □

Proposition 3. *For any $(\theta_l, \theta_u) \subseteq B_n = (\theta_{i-1}^0 - k_n^1, \theta_i^0 + k_n^2)$, we define $\Psi_i^* = \arg\max_{\Psi_i} L_i(\Psi_i, \theta_l, \theta_u, d)$ and $\hat{\Psi}_i = \arg\max_{\Psi_i} L_{n,i}(\Psi_i, \theta_l, \theta_u, d)$. Then, under Assumptions 1–3, $\hat{\Psi}_i \to \Psi_i^*$ almost surely.*

Proof of Proposition 3. From the definition of $\hat{\Psi}_i$ and Ψ_i^*, we have

$$L_{n,i}(\hat{\Psi}_i, \theta_l, \theta_u, d) \geq L_{n,i}(\Psi_i^*, \theta_l, \theta_u, d) \,, L_i(\hat{\Psi}_i, \theta_l, \theta_u, d) \leq L_i(\Psi_i^*, \theta_l, \theta_u, d) \,.$$

By decomposing $L_i(\Psi_i^*, \theta_l, \theta_u, d) - L_i(\hat{\Psi}_i, \theta_l, \theta_u, d)$, we have almost surely that

$$\begin{aligned} 0 \leq & L_i(\Psi_i^*, \theta_l, \theta_u, d) - L_i(\hat{\Psi}_i, \theta_l, \theta_u, d) \\ \leq & |L_i(\Psi_i^*, \theta_l, \theta_u, d) - L_{n,i}(\Psi_i^*, \theta_l, \theta_u, d)| + L_{n,i}(\Psi_i^*, \theta_l, \theta_u, d) \\ & - L_{n,i}(\hat{\Psi}_i, \theta_l, \theta_u, d) + |L_{n,i}(\hat{\Psi}_i, \theta_l, \theta_u, d) - L_i(\hat{\Psi}_i, \theta_l, \theta_u, d)| \qquad (7.21) \\ \leq & 0 \,, \end{aligned}$$

where the first and last terms in (7.21) converge to 0 almost surely by Proposition 2. Thus, by the uniqueness of Ψ_i^* under Assumption 2, $\hat{\Psi}_i^* \to \Psi_i^*$ almost surely. □

For any $(\theta_l, \theta_u) \subseteq (\theta_{i-1}^0, \theta_i^0)$, by the theory of Kullback–Leibler distance, $L_i(\Psi_i^0, \theta_l, \theta_u, d^0) \geq L_i(\Psi_i, \theta_l, \theta_u, d)$, where the equality sign holds if and only if $d = d^0$ and $\Psi_i = \Psi_i^0$. This observation is the main idea for proving Theorem 5.1.

Proof of Theorem 5.1. We prove by contradiction. Let A be the probability one set under which Propositions 2 and 3 hold. For each $\omega \in A$, suppose that there exists a subsequence $\{n_m\}$ such that $\hat{r}_{n_m} \to r^*$, $\hat{d}_{n_m} \to d^*$, $\hat{\Theta}_{n_m} \to \Theta^*$, $\hat{\Psi}_{n_m} \to$

Ψ^* almost surely with $\{\Psi^*, \Theta^*, d^*\} \in \Omega_{\mathcal{M}} \times \{1, \ldots, D\}$ under Assumption 1. For simplicity, we omit the subscript "n_m" and replace $\{n_m\}$ by $\{n\}$ when necessary. Suppose that $r^* < r^0$, then there must be some θ^*_{j-1} and θ^*_j such that for some positive integer k,

$$\theta^0_{i-1} < \theta^*_{j-1} \leq \theta^0_i < \theta^0_{i+1} < \cdots < \theta^0_{i+k} < \theta^*_j \leq \theta^0_{i+k+1}\,.$$

In other words, $k+2$ true regimes are pooled into a "working regime" j. We relabel the true regimes as sub-regime $1, ..., k+2$ with thresholds of sub-regime l denoted as $(\theta_{(l-1)}, \theta_{(l)}]$. Hence, the number of observations in the working regime j is n_j. The log-likelihood of the working regime j is

$$n_j L^*_{n_j}(\hat{\Psi}_j, \hat{\theta}_{j-1}, \hat{\theta}_j, \hat{d}) = \sum_{l=1}^{k+2} n_j L^*_{n_{j,l}}(\hat{\Psi}_j, \hat{\theta}_{(l-1)}, \hat{\theta}_{(l)}, \hat{d})\,,$$

where

$$L^*_{n_{j,l}}(\Psi_j, \theta_{(l-1)}, \theta_{(l)}, d) = \frac{1}{n_j} \sum_{t=1}^{n_j} l(\Psi_j; y_t, Y_{t-1}, X_t) I(z_{t-d} \in (\theta_{(l-1)}, \theta_{(l)}])\,,$$

$$L^*_{n_j}(\Psi_j, \theta^*_{j-1}, \theta^*_j, d) = \frac{1}{n_j} \sum_{t=1}^{n_j} l(\Psi_j; y_t, Y_{t-1}, X_t) I(z_{t-d} \in (\theta^*_{j-1}, \theta^*_j])\,,$$

and $\hat{\Psi}_j = \arg\max_{\Psi_j} L^*_{n_j}(\Psi_j, \theta^*_{j-1}, \theta^*_j, \hat{d})$. Similarly, we define

$$L^*_{j,l}(\Psi_j, \theta_{(l-1)}, \theta_{(l)}, d) = \lim_{n_j \to \infty} \mathrm{E}[L^*_{n_{j,l}}(\Psi_j, \theta_{(l-1)}, \theta_{(l)}, d)]\,,$$

and

$$L^*_j(\Psi_j, \theta^*_{j-1}, \theta^*_j, d) = \lim_{n_j \to \infty} \mathrm{E}[L^*_{n_j}(\Psi_j, \theta^*_{j-1}, \theta^*_j, d)]\,.$$

As $n \to \infty$, we have $n_j \to \infty$ and $n_{j,l} \to \infty$. Denote the true parameter in sub-regime l as $\Psi^0_{(l)}$. Consider a sub-regime m, from Propositions 2, 3 and theory of Kullback–Leibler distance,

$$\begin{aligned}\lim_{n \to \infty} L^*_{n_{j,m}}(\hat{\Psi}_j, \theta_{(m-1)}, \theta_{(m)}, \hat{d}) &= L^*_{j,m}(\Psi^*_j, \theta_{(m-1)}, \theta_{(m)}, d^*) \\ &\leq L^*_{j,m}(\Psi^0_{(m)}, \theta_{(m-1)}, \theta_{(m)}, d^0) \quad \text{a.s.}\,.\end{aligned}$$

However, the equality cannot hold for all m under Assumption 2. Hence, we have

$$\lim_{n_j \to \infty} L^*_{n_j}(\hat{\Psi}_j, \hat{\theta}_{j-1}, \hat{\theta}_j, \hat{d}) < \sum_{l=1}^{k+2} L^*_{j,l}(\Psi^0_{(l)}, \theta_{(l-1)}, \theta_{(l)}, d^0) \quad \text{a.s.}\,, \tag{7.22}$$

since at least one part in the summation of (7.22) is not maximized. Furthermore, for one of such sub-regime m, equipped with the ergodic theorem, there exists some $c_m > 0$ such that

$$\begin{aligned} L^*_{j,m}(\Psi^*_j, \theta_{(m-1)}, \theta_{(m)}, d^*) &= \mathrm{E}[l(\Psi^*_j; y_t, Y_{t-1}, X_t) I(z_{t-d^*} \in (\theta_{(m-1)}, \theta_{(m)}])] \\ &= \mathrm{E}[l(\Psi^0_{(m)}; y_t, Y_{t-1}, X_t) I(z_{t-d_0} \in (\theta_{(m-1)}, \theta_{(m)}])] \\ &\quad - c_m \\ &= L^*_{j,m}(\Psi^0_{(m)}, \theta_{(m-1)}, \theta_{(m)}, d^0) - c_m \,. \end{aligned}$$

By the ergodic theorem again, we have

$$L(\Psi^0, \Theta^0, d^0) - L(\Psi^*, \Theta^*, d^*) = O(n) \,,$$

and is positive almost surely. On the other hand, from (7.19), $\mathrm{Pen}(\Theta^0, p^0, d^0) - \mathrm{Pen}(\Theta^*, p^*, d^*)$ is of order $O(\log(n))$. Hence, the decrease in log-likelihood is more rapid. Therefore, $r^* < r^0$ fails to optimize $\mathrm{MDL}(\mathcal{M})$. In general, if one of the working regimes is not nested in a true regime, for example, $\theta^0_{i-1} < \theta^*_{j-1} < \theta^0_i < \theta^*_j < \theta^0_{i+1}$ for some i, j, then the $\mathrm{MDL}(\mathcal{M})$ cannot be smaller than that with sub-regimes $(\theta^0_{i-1}, \theta^*_{j-1}]$, $(\theta^*_{j-1}, \theta^0_i]$ and $(\theta^0_i, \theta^*_j]$. As a result, all working regimes have to be nested in some true regimes, i.e., $\theta^0_{i-1} < \theta^*_{j-1} < \theta^*_j < \theta^0_i$ for all j with some i, which implies $r^* \geq r^0$.

Next, assume that $d^* \neq d^0$ with $r^* \geq r^0$. Consider regime i, from Proposition 2,

$$\begin{aligned} &\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \sum_{i=1}^{r+1} l(\Psi_i; y_t, Y_{t-1}, X_t) I(z_{t-d} \in (\theta_{i-1}, \theta_i]) \\ &\quad = L_i(\Psi_i, \theta_{i-1}, \theta_i, d) < \infty \quad \text{a.s.} \,. \end{aligned}$$

As the estimated regimes are nested in true regimes, $(\theta^*_{j-1}, \theta^*_j) \subseteq [\theta^0_{i-1}, \theta^0_i]$. By the property of Kullback-Leibler distance,

$$L_i(\Psi^0_i, \theta^*_{j-1}, \theta^*_j, d^0) - L_i(\Psi^*_j, \theta^*_{j-1}, \theta^*_j, d^*) \geq 0 \,,$$

for all j and i, where the equality holds if and only if $d^* = d^0$. Hence it implies

$$L(\Theta^0, \Psi^0, d^0) > L(\Theta^*, \Psi^*, d^*) \,, \tag{7.23}$$

where the difference between the two terms is of order $O(n)$. Thus, for the optimal model, $\hat{d} \to d^0$ almost surely.

Moreover, if $\hat{r} > r^0$, at least two of the classified regimes are sub-regimes of a true regime. By Taylor's expansion, for any $(\theta_l, \theta_u] \subseteq (\theta^0_{i-1}, \theta^0_i]$,

$$\begin{aligned} & L_{n,i}(\hat{\Psi}_i, \theta_l, \theta_u, d^0) - L_{n,i}(\Psi^0_i, \theta_l, \theta_u, d^0) \\ = \;& (\hat{\Psi}_i - \Psi^0_i) L^{(1)}_{n,i}(\Psi^0_i, \theta_l, \theta_u, d^0) + (\hat{\Psi}_i - \Psi^0_i)^2 L^{(2)}_{n,i}(\Psi^0_i, \theta_l, \theta_u, d^0) + o(1) \\ = \;& o(1) + (\hat{\Psi}_i - \Psi^0_i)^2 L^{(2)}_{n,i}(\Psi^0_i, \theta_l, \theta_u, d^0) + o(1) \quad \text{a.s.} \,. \end{aligned} \tag{7.24}$$

As

$$L^{(1)}_{n,i}(\hat{\Psi}_i, \theta_l, \theta_u, d^0) - L^{(1)}_{n,i}(\Psi^0_i, \theta_l, \theta_u, d^0) = (\hat{\Psi} - \Psi^0_i) L^{(2)}_{n,i}(\Psi^0_i, \theta_l, \theta_u, d^0) + o(1) \,, \tag{7.25}$$

where the left-hand-side converges to 0 almost surely, by Assumption 2 and Kolmogorov's law of iterated logarithm, we have

$$\limsup_{n\to\infty} \frac{n^{1/2}}{(\log\log(n))^{1/2}} L_{n,i}^{(2)}(\Psi_i^0, \theta_l, \theta_u, d^0) \longrightarrow 2^{1/2}[\mathrm{var}(L_i^{(2)}(\Psi_i^0, \theta_l, \theta_u, d^0))]^{1/2} \quad \text{a.s.}. \tag{7.26}$$

Together with (7.25), we have $|\hat{\Psi} - \Psi_i^0| = O((n^{-1}\log\log(n))^{1/2})$ almost surely. Thus, by (7.24), $L(\Theta^0, \Psi^0, d^0) - L(\Theta^*, \Psi^*, d^*) = O(\log\log(n))$. As $r^* > r^0$ results in $\mathrm{Pen}(\Theta^*, p^*, d^0) > \mathrm{Pen}(\Theta^0, p^0, d^0)$ where $\mathrm{Pen}(\Theta^*, p^*, d^0) - \mathrm{Pen}(\Theta^0, p^0, d^0) = O(\log(n))$, it implies that the MDL criterion is not minimized almost surely. Hence, $\hat{r} \to r^0$ almost surely.

Under the consistency conditions, we can assume $\hat{d} = d^0$ and $\hat{r} = r^0$ holds for sufficiently large n. As all of the estimated regimes are nested in true regimes, $\hat{\Theta} \to \Theta^0$ almost surely. From Proposition 2, in regime i, we have almost surely that $\hat{\Psi}_i \to \Psi_i^*$, $\hat{p}_i \to p_i^*$ for some Ψ_i^* and p_i^*. Suppose that $\Psi_i^* \neq \Psi_i^0$ in one of the regime i. Then, as Kullback-Leibler distance suggests, $L_i(\Psi_i^*, \theta_{i-1}^0, \theta_i^0, d^0) \leq L_i(\Psi_i^0, \theta_{i-1}^0, \theta_i^0, d^0)$. Analogous to the argument in proving $\hat{d} \to d^0$, $\mathrm{MDL}(\mathcal{M})$ is not asymptotically minimized. On the other hand, if $\hat{\Psi} \to \Psi^0$ almost surely, we have $\hat{p} \to p^*$ simultaneously for some p^*. As $\mathrm{Pen}(\Theta^0, p^*, d^0) \geq \mathrm{Pen}(\Theta^0, p^0, d^0)$ where equality holds if and only if $p^* = p^0$, by the strong consistency of order estimation of MDL (see [47] and [48]) with respect to true regimes, $\hat{p} \to p^0$ almost surely. □

Proof of Theorem 5.2.

1) We prove by showing the $O_p(n^{-1})$ convergence speed of $\hat{\theta}_i$ for all i. Under Theorem 5.1, $|\hat{\theta}_i - \theta_i^0| = o(1)$ almost surely; hence it suffices to verify $\mathrm{pr}(|\hat{\theta}_i - \theta_i^0| > c/n) \to 0$ for some $c > 0$. Below we show $\mathrm{pr}(|\hat{\theta}_i - \theta_i^0| I(\theta_i^0 < \hat{\theta}_i) > c/n) \to 0$, and the same argument give $\mathrm{pr}(|\hat{\theta}_i - \theta_i^0| I(\theta_i^0 \geq \hat{\theta}_i) > c/n) \to 0$, and thus the result follows.

From the strong consistency of $\hat{\theta}_i$, we can restrict $|\hat{\theta}_i - \theta_i^0| < \delta$ and $|\hat{\Psi}_i - \Psi_i^0| < \delta$ for some positive δ and sufficiently large n. Denote $Q(a) = \mathrm{E}\{I(z_{t-d^0} \in (\theta_i^0, \theta_i^0 + a])\}$ for some $a > 0$, where $I(\cdot)$ is the indicator function. Then for any $\xi, \delta > 0$, there is some $c > 0$ such that

$$\mathrm{pr}\left(\sup_{c/n < a \leq \delta} \left| \sum_{t=1}^{n} \frac{I(z_{t-d^0} \in (\theta_i^0, \theta_i^0 + a])}{nQ(a)} - 1 \right| < \xi \right) > 1 - \epsilon\,. \tag{7.27}$$

Furthermore, from Assumption 4a), there exists a measurable function $\Gamma_t = \Gamma(Y_{t-1}, X_t, y_t)$ such that $\mathrm{E}(\Gamma_t^2 | z_{t-d^0}) \leq M$ for all $z_{t-d^0} \in [\theta_i^0 - \delta, \theta_i^0 + \delta]$ and some constant $M > 0$. Combined with Assumption 5, the joint process $\{\Gamma_t I(z_{t-d^0} \in [\theta_i^0 - \delta, \theta_i^0 + \delta]), z_{t-d^0} I(z_{t-d^0} \in [\theta_i^0 - \delta, \theta_i^0 + \delta])\}$ is ρ-mixing with summable mixing coefficients. Therefore, we have

$$\text{pr}\left(\sup_{c/n<a\leq\delta}\left|\sum_{t=1}^{n}\frac{\Gamma_t I(z_{t-d^0}\in(\theta_i^0,\theta_i^0+a])-\text{E}\{\Gamma_t I(z_{t-d_0}\in(\theta_i^0,\theta_i^0+a])\}}{nQ(a)}\right|<\xi\right)$$
$$>1-\epsilon\,. \tag{7.28}$$

The derivations of (7.27) and (7.28) follow similar arguments in Proposition 1 of [10] and Theorem 2 of [54]. Let

$$\begin{aligned}\mathcal{H}(\theta_i^0,a) &= \frac{1}{nQ(a)}\sum_{t=1}^{n}\left[l(\hat{\Psi}_i;y_t,Y_{t-1},X_t)-l(\hat{\Psi}_{i+1};y_t,Y_{t-1},X_t)\right]\\ &\quad\times I(z_{t-d^0}\in(\theta_i^0,\theta_i^0+a])\\ &= \frac{1}{nQ(a)}\sum_{t=1}^{n}\Big[\left\{l(\hat{\Psi}_i;y_t,Y_{t-1},X_t)-l(\Psi_i^0;y_t,Y_{t-1},X_t)\right\}\\ &\quad+\left\{l(\Psi_i^0;y_t,Y_{t-1},X_t)-l(\Psi_{i+1}^0;y_t,Y_{t-1},X_t)\right\}\\ &\quad+\left\{l(\Psi_{i+1}^0;y_t,Y_{t-1},X_t)-l(\hat{\Psi}_{i+1};y_t,Y_{t-1},X_t)\right\}\Big]\\ &\quad\times I(z_{t-d^0}\in(\theta_i^0,\theta_i^0+a])\,.\end{aligned}$$

From Assumptions 4a) and 5, (7.27) and (7.28), we have for every $\xi,\delta>0$, there exists a $c>0$ satisfying $c/n<a<\delta$ such that

$$\begin{aligned}\mathcal{H}(\theta_i^0,a) &< \frac{1}{nQ(a)}\sum_{t=1}^{n}\left[l(\Psi_i^0;y_t,Y_{t-1},X_t)-l(\Psi_{i+1}^0;y_t,Y_{t-1},X_t)\right]\\ &\quad\times I(z_{t-d^0}\in(\theta_i^0,\theta_i^0+a])\\ &\quad+\left(|\hat{\Psi}_i-\Psi_i^0|+|\Psi_{i+1}^0-\hat{\Psi}_{i+1}|\right)(\xi+M)+\xi\,,\end{aligned} \tag{7.29}$$

with probability greater than $1-\epsilon$ for some small positive ϵ. Using Assumptions 2, 4 and Lemma 5.35 in [62], for $0<a<\delta$, there exists a constant $\chi<0$ such that

$$\text{E}\left[l(\Psi_i^0;y_t,Y_{t-1},X_t)-l(\Psi_{i+1}^0;y_t,Y_{t-1},X_t)\mid z_{t-d^0}\in(\theta_i^0,\theta_i^0+a]\right]\leq\chi<0\,.$$

Following Theorem 2 of [54], the first term on the right hand side of (7.29) is less than $\chi(1-\xi)$ with some $\xi>0$. Hence, by appropriate choices of δ and ξ such that $2\delta(M+\xi)+\xi+\chi(1-\xi)<0$, we verify that $\mathcal{H}(\theta_i^0,a)<0$ in probability, with δ,ξ and ϵ go to 0 as $n\to\infty$.

Now suppose that $|\theta_i^0-\hat{\theta}_i|>c/n$ holds for some $c>0$ with positive probability. Denote $\tilde{\Theta}=(\tilde{\theta}_1,\ldots,\tilde{\theta})$ with $\tilde{\theta}_j=\hat{\theta}_j$ if $j\neq i$ and $\tilde{\theta}_i=\theta_i^0$, and $\tilde{\Psi}$ is the maximum likelihood estimator under $\tilde{\Theta}$. As $\hat{\Theta}$ is the minimizer of MDL, we have

$$\begin{aligned}&\text{pr}\left(|\theta_i^0-\hat{\theta}_i|I(\theta_i^0<\hat{\theta}_i)>c/n\right)\\ =\;&\text{pr}\left(|\theta_i^0-\hat{\theta}_i|I(\theta_i^0<\hat{\theta}_i)>c/n,\text{MDL}(\hat{\mathcal{M}}(\hat{\Psi},\hat{\Theta},d^0))<\text{MDL}(\tilde{\mathcal{M}}(\tilde{\Psi},\tilde{\Theta},d^0))\right)\\ \leq\;&\text{pr}\left(\text{MDL}(\hat{\mathcal{M}}(\hat{\Psi},\hat{\Theta},d^0))<\text{MDL}(\tilde{\mathcal{M}}(\tilde{\Psi},\tilde{\Theta},d^0))\right)\,,\end{aligned}$$

where $\hat{\mathcal{M}}(\hat{\Psi}, \hat{\Theta}, d^0), \tilde{\mathcal{M}}(\tilde{\Psi}, \tilde{\Theta}, d^0)$ are models with parameters $\{\hat{\Psi}, \hat{\Theta}, d^0\}$ and $\{\tilde{\Psi}, \tilde{\Theta}, d^0\}$, respectively. As $|\theta_i^0 - \hat{\theta}_i| > c/n$, there exists some $a^* \in (c/n, \delta]$ with a sufficiently small δ such that

$$L(\tilde{\Psi}, \tilde{\Theta}, d^0) - L(\hat{\Psi}, \hat{\Theta}, d^0) = -nQ(a^*)\mathcal{H}(\theta_i^0, a^*) = O_p(1)\,. \tag{7.30}$$

Furthermore, $\mathcal{H}(\theta_i^0, a)$ is negative with probability going to 1, the term in (7.30) converges in probability to a positive term. Meanwhile, as $\hat{\theta}_i \to \theta_i^0$ and $\hat{p}_i \to p_i^0$ almost surely, the difference of penalties $\text{Pen}(\tilde{\Theta}, \tilde{p}, d^0) - \text{Pen}(\hat{\Theta}, \hat{p}, d^0) = o(1)$. Therefore, as $n \to \infty$,

$$\text{pr}(\text{MDL}(\hat{\mathcal{M}}(\hat{\Psi}, \hat{\Theta}, d^0)) > \text{MDL}(\tilde{\mathcal{M}}(\tilde{\Psi}, \tilde{\Theta}, d^0))) \to 1\,,$$

and thus $\text{pr}(|\theta_i^0 - \hat{\theta}_i| I(\theta_i^0 < \hat{\theta}_i) > c/n) \to 0$. Therefore, the convergence rate of $\hat{\theta}_i$ is of order $O_p(n^{-1})$ for all $i = 1, \ldots, r^0$, and the proof is complete.

2) We first prove the convergence property of $\text{pr}(\hat{\theta}_i \neq \theta_i^0)$. For simplicity we assume $\hat{\theta}_i$ are integers and $\nu_i = \nu$ in all regimes. Without loss of generality, we show the convergence of $\text{pr}(\hat{\theta}_i > \theta_i^0)$.

Denote $\hat{\Psi}_i = \arg\max_{\Psi_i} L_{n,i}(\Psi_i, \theta_{i-1}^0, \theta_i^0, d^0)$ and $\hat{\Psi}_i^* = \arg\max_{\Psi_i} L_{n,i}(\Psi_i, \theta_{i-1}^0, \tilde{\theta}_i, d^0)$, where $\hat{\Psi}_i^* \to \Psi_i^*$ almost surely for $\Psi_i^* = \arg\max_{\Psi_i} L_{n,i}(\Psi_i, \theta_{i-1}^0, \tilde{\theta}_i, d^0)$ which is neither Ψ_i^0 nor Ψ_{i+1}^0. Under (5.2), we have $\lambda_t^* = \lambda_t(\Psi_i^*; Y_{t-1}, X_t, z_{t-d})$ and $\lambda_t^0 = \lambda_t(\Psi_i^0; Y_{t-1}, X_t, z_{t-d})$, and $\lambda_t^* \neq \lambda_t^0$ almost surely.

For discrete z_t and any $\tilde{\theta}_i > \theta_i^0$, the number of observations with $z_{t-d} \in (\theta_i^0, \tilde{\theta}_i]$ with order $O(n)$ almost surely. If $\hat{\theta}_i \to \tilde{\theta}_i \neq \theta_i^0$, then by the ergodic theorem,

$$n\left[L_{n,i}(\Psi_i^0, \theta_{i-1}^0, \theta_i^0, d^0) - L_{n,i}(\Psi_i^*, \theta_{i-1}^0, \tilde{\theta}_i, d^0)\right] = O(n) \quad \text{a.s.}\,,$$

which overweights the $O(\log(n))$ rate of difference in penalties under θ_i^0 and $\tilde{\theta}_i$ by (7.19). Hence we claim that asymptotically,

$$\begin{aligned} & \text{pr}(\hat{\theta}_i > \theta_i^0) \\ = \; & \text{pr}\left(\sum_{t=1}^{n}[l(\hat{\Psi}_i^*; y_t, Y_{t-1}, X_t) - l(\hat{\Psi}_i; y_t, Y_{t-1}, X_t)] I(z_{t-d} \in (\theta_{i-1}^0, \hat{\theta}_i]) > 0\right). \end{aligned}$$

For any $\tilde{\theta}_i > \theta_i^0$, we divide all observations with $z_{t-d} \in (\theta_{i-1}^0, \tilde{\theta}_i]$ into two partitions: the first partition constitutes all y_t with $z_{t-d} \in (\theta_{i-1}^0, \theta_i^0]$; and the second partition constitutes of all y_t with $z_{t-d} \in (\theta_i^0, \tilde{\theta}_i]$. For the first partition, from Assumption 4b),

$$\begin{aligned} & l(\hat{\Psi}_i^*; y_t, Y_{t-1}, X_t) - l(\hat{\Psi}_i; y_t, Y_{t-1}, X_t) \\ = \; & [l(\Psi_i^*; y_t, Y_{t-1}, X_t) - l(\Psi_i^0; y_t, Y_{t-1}, X_t)](1 + o(1)) \\ = \; & \frac{1}{\nu a_t}[\{T(y_t)\gamma(\lambda_t^*) - b(\lambda_t^*)\} - \{T(y_t)\gamma(\lambda_t^0) - b(\lambda_t^0)\}](1 + o(1))\,. \end{aligned} \tag{7.31}$$

Here λ_t^* and λ_t^0 are deterministic given $\{Y_{t-1}, X_t, z_{t-d}\}$, and $\{T(y_t)\}_{t=1,2,\ldots}$ in (7.31) are mutually independent given λ_t^0. To simplify derivation, we relabel all observations in the first group from $s = 1, \ldots, n_i^0$, and denote $\Lambda_{n_i^0}$ as a realization of $\{\lambda_s^*, \lambda_s^0\}_{s=1,\ldots,n_i^0}$, where $\lambda_s^* \neq \lambda_s^0$, $s = 1, \ldots, n_i^0$ almost surely. Let f_Λ be the joint density of $\{\lambda_s^*, \lambda_s^0\}_{s=1,\ldots,n_i^0}$. By (7.31), for sufficiently large n and n_i^0,

$$\begin{aligned}
&\mathrm{pr}\left(\sum_{t=1}^{n}[l(\hat{\Psi}_i^*; y_t, Y_{t-1}, X_t) - l(\hat{\Psi}_i; y_t, Y_{t-1}, X_t)]I(z_{t-d} \in (\theta_{i-1}^0, \theta_i^0]) > 0\right) \\
= &\int \mathrm{pr}\left(\sum_{s=1}^{n_i^0}\frac{1}{\nu a_s}[\{T(y_s)\gamma(\lambda_s^*) - b(\lambda_s^*)\}\right. \\
&\qquad \left. - \{T(y_s)\gamma(\lambda_s^0) - b(\lambda_s^0)\}](1 + o(1)) > 0 \mid \Lambda_{n_i^0}\right) \\
&\qquad \times f_\Lambda\left(\{\lambda_s^*, \lambda_s^0\}_{s=1,\ldots,n_i^0}\right) d\lambda_1^*\lambda_1^0 \ldots \lambda_{n_i^0}^*\lambda_{n_i^0}^0\,, \qquad (7.32)
\end{aligned}$$

with

$$\begin{aligned}
&\mathrm{pr}\left(\sum_{s=1}^{n_i^0}\frac{1}{\nu a_s}[\{T(y_s)\gamma(\lambda_s^*) - b(\lambda_s^*)\}\right. \\
&\qquad \left. - \{T(y_s)\gamma(\lambda_s^0) - b(\lambda_s^0)\}](1 + o(1)) > 0 \mid \Lambda_{n_i^0}\right) \\
= &\mathrm{pr}\left(\sum_{s=1}^{n_i^0}\frac{1}{\nu a_s}[T(y_s)\{\gamma(\lambda_s^*) - \gamma(\lambda_s^0)\} - \mathrm{E}(T(y_s) \mid \lambda_s^0)\{\gamma(\lambda_s^*) - \gamma(\lambda_s^0)\}\right. \\
&\qquad \left. > \sum_{s=1}^{n_i^0}\frac{1}{\nu a_s}[b(\lambda_s^*) - b(\lambda_s^0) - \mathrm{E}(T(y_s) \mid \lambda_s^0)\{\gamma(\lambda_s^*) - \gamma(\lambda_s^0)\}] \mid \Lambda_{n_i^0}\right) \\
&\times (1 + o(1)) \\
= &\mathrm{pr}\left(\sum_{s=1}^{n_i^0}\frac{1}{\nu a_s}T^*(y_s)\{\gamma(\lambda_s^*) - \gamma(\lambda_s^0)\} > \sum_{s=1}^{n_i^0}\Delta_b(\lambda_s^*, \lambda_s^0) \mid \Lambda_{n_i^0}\right)(1 + o(1))\,, \\
&\qquad (7.33)
\end{aligned}$$

where given $\Lambda_{n_i^0}$ in (7.33), $T^*(y_s) = T(y_s) - \mathrm{E}(T(y_s) \mid \lambda_s^0)$, $s = 1, 2, \ldots$ are mutually independent. Denote

$$\begin{aligned}
\Delta_b(\lambda_s^*, \lambda_s^0) &= \frac{1}{\nu a_s}[b(\lambda_s^*) - b(\lambda_s^0) - \mathrm{E}(T(y_s) \mid \lambda_s^0)\{\gamma(\lambda_s^*) - \gamma(\lambda_s^0)\}] \\
&= \mathrm{E}[l(\Psi_i^0; y_s, Y_{s-1}, X_s)] - \mathrm{E}[l(\Psi_i^*; y_t, Y_{s-1}, X_s)]\,.
\end{aligned}$$

By Assumption 4b), $\mathrm{E}[l(\Psi_i^0; y_s, Y_{s-1}, X_s) - l(\Psi_i^*; y_t, Y_{s-1}, X_s)]^2 < \infty$, and hence

$$\mathrm{E}|\Delta_b(\lambda_s^*, \lambda_s^0)| < \infty\,, \quad \text{and} \quad \mathrm{E}\left[\frac{1}{\nu^2 a_s^2}\mathrm{var}(T(y_s) \mid \lambda_s^0)\{\gamma(\lambda_s^*) - \gamma(\lambda_s^0)\}^2\right] < \infty\,. \tag{7.34}$$

By the property of Kullback–Leibler distance, $\mathrm{E}[l(\Psi_i^0; y_s, Y_{s-1}, X_s)] > \mathrm{E}[l(\Psi_i^*; y_t, Y_{s-1}, X_s)]$ if $\lambda_s^* \neq \lambda_s^0$. Thus $\Delta_b(\lambda_s^*, \lambda_s^0) > 0$ for all y_s with $z_{s-d} \in (\theta_{i-1}^0, \theta_i^0]$. As $\lambda_s^* \neq \lambda_s^0$ almost surely and $\{Y_{t-1}, X_t, z_{t-d}\}$ is ergodic, by (7.34) and the ergodic theorem,

$$\lim_{n_i^0 \to \infty} \frac{1}{n_i^0} \sum_{s=1}^{n_i^0} \Delta_b(\lambda_s^*, \lambda_s^0) = \mathrm{E}(\Delta_b(\lambda_s^*, \lambda_s^0)) > 0 \quad \text{a.s.}\,. \tag{7.35}$$

Meanwhile, as

$$\begin{aligned} & \mathrm{var}\left[\sum_{s=1}^{n_i^0} \frac{1}{\nu a_s} T^*(y_s)\{\gamma(\lambda_s^*) - \gamma(\lambda_s^0)\} \mid \Lambda_{n_i^0}\right] \\ = & \sum_{s=1}^{n_i^0} \frac{1}{\nu^2 a_s^2} \mathrm{var}(T(y_s) \mid \lambda_s^0)\{\gamma(\lambda_s^*) - \gamma(\lambda_s^0)\}^2\,, \end{aligned}$$

analogously we have

$$\begin{aligned} & \lim_{n_i^0 \to \infty} \frac{1}{n_i^0} \mathrm{var}\left[\sum_{s=1}^{n_i^0} \frac{1}{\nu a_s} T^*(y_s)\{\gamma(\lambda_s^*) - \gamma(\lambda_s^0)\} \mid \Lambda_{n_i^0}\right] \\ = & \lim_{n_i^0 \to \infty} \frac{1}{n_i^0} \sum_{s=1}^{n_i^0} \frac{1}{\nu^2 a_s^2} \mathrm{var}(T(y_s) \mid \lambda_s^0)\{\gamma(\lambda_s^*) - \gamma(\lambda_s^0)\}^2 \\ = & \mathrm{E}\left[\frac{1}{\nu^2 a_s^2} \mathrm{var}(T(y_s) \mid \lambda_s^0)\{\gamma(\lambda_s^*) - \gamma(\lambda_s^0)\}^2\right] > 0 \quad \text{a.s.}\,. \end{aligned} \tag{7.36}$$

by (7.34) and the ergodic theorem. Denote

$$\mathcal{X}_{n_i^0} = \frac{\sum_{s=1}^{n_i^0} \Delta_b(\lambda_s^*, \lambda_s^0)}{\left[\sum_{s=1}^{n_i^0} \mathrm{var}\left(\frac{1}{\nu a_s} T^*(y_s)\{\gamma(\lambda_s^*) - \gamma(\lambda_s^0)\} \mid \Lambda_{n_i^0}\right)\right]^{1/2}}\,. \tag{7.37}$$

By (7.35) and (7.36), $(n_i^0)^{-1/2}\mathcal{X}_{n_i^0}$ converges almost surely to some $\tau = \tau(\tilde{\theta}_i) > 0$. By Theorem 10 in Chapter VIII of [46] and Assumption 4b), we have

$$\begin{aligned} & \mathrm{pr}\left(\sum_{s=1}^{n_i^0} \frac{1}{\nu a_s} T^*(y_s)\{\gamma(\lambda_s^*) - \gamma(\lambda_s^0)\} > \sum_{s=1}^{n_i^0} \Delta_b(\lambda_s^*, \lambda_s^0) \mid \Lambda_{n_i^0}\right) \\ = & \{1 - \Phi(\mathcal{X}_{n_i^0})\}\exp\left[\frac{\mathcal{X}_{n_i^0}^3}{(n_i^0)^{1/2}} h_{n_i^0}\left(\frac{\mathcal{X}_{n_i^0}}{(n_i^0)^{1/2}}\right)\right](1 + l_1\tau)\,, \end{aligned} \tag{7.38}$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal random variable, $h_{n_i^0}(x) = \sum_{k=0}^{\infty} a_{kn_i^0} x^k$ is the generalized Cramér series defined on p.220 of [46], and l_1 is some constant. By [46], $h_{n_i^0}(\tau)$ converges as n_i^0 goes to infinity for sufficiently small $|\tau|$. Hence, we may select a sufficiently small constant τ^* such that for sufficiently large n_i^0, $h_{n_i^0}(\tau) \le H$ for some constant H whenever $|\tau| \le \tau^*$. Therefore, for sufficiently large n, if $\tau \le \tau^*$, we have

$$1 - \Phi(\mathcal{X}_{n_i^0}) \le c_1 n^{-\frac{1}{2}} e^{-\frac{1}{2}\tau^2 n}, \quad \text{and} \quad \exp\left[\frac{\mathcal{X}_{n_i^0}^3}{(n_i^0)^{1/2}} h_{n_i^0}\left(\frac{\mathcal{X}_{n_i^0}}{(n_i^0)^{1/2}}\right)\right] \le c_2 e^{\tau^3 n},$$

for some $c_1, c_2 > 0$. Thus, (7.38) is of order $O(n^{-1/2}e^{-an})$ almost surely for some $a = \tau^2/2 - \tau^3 > 0$, by choosing a proper $\tau^* < 1/2$. Otherwise, if $\tau > \tau^*$, we have

$$\begin{aligned}
& \text{pr}\left(\sum_{s=1}^{n_i^0} \frac{1}{\nu a_s} T^*(y_s)\{\gamma(\lambda_s^*) - \gamma(\lambda_s^0)\} > \sum_{s=1}^{n_i^0} \Delta_b(\lambda_s^*, \lambda_s^0) \mid \Lambda_{n_i^0}\right) \\
\le\ & \text{pr}\left(\sum_{s=1}^{n_i^0} \frac{1}{\nu a_s} T^*(y_s)\{\gamma(\lambda_s^*) - \gamma(\lambda_s^0)\} > \frac{\tau^*}{\tau} \sum_{s=1}^{n_i^0} \Delta_b(\lambda_s^*, \lambda_s^0) \mid \Lambda_{n_i^0}\right) \\
=\ & \left\{1 - \Phi\left(\frac{\tau^*}{\tau}\mathcal{X}_{n_i^0}\right)\right\} \exp\left[\frac{(\tau^*)^3}{\tau^3} \frac{\mathcal{X}_{n_i^0}^3}{(n_i^0)^{1/2}} h_{n_i^0}\left(\frac{\tau^* \mathcal{X}_{n_i^0}}{\tau (n_i^0)^{1/2}}\right)\right] (1 + l_1 \tau^*) \\
\le\ & c_1 c_2 (1 + l_1 \tau^*) n^{-\frac{1}{2}} e^{-(\frac{1}{2}(\tau^*)^2 - (\tau^*)^3)n} \quad \text{a.s.}. \qquad (7.39)
\end{aligned}$$

Next, using similar arguments in (7.38), (7.39) is of order $O(n^{-1/2}e^{-a'n})$ almost surely for some $a' = (\tau^*)^2/2 - (\tau^*)^3 > 0$. Thus, given $\Lambda_{n_i^0}$, (7.33) is of order $O(n^{-1/2}e^{-a_1 n})$ almost surely for some $a_1 > 0$. Combining with (7.32), we have

$$\begin{aligned}
& \text{pr}\left(\sum_{t=1}^{n} [l(\hat{\Psi}_i^*; y_t, Y_{t-1}, X_t) - l(\hat{\Psi}_i; y_t, Y_{t-1}, X_t)] I(z_{t-d} \in (\theta_{i-1}^0, \theta_i^0]) > 0\right) \\
=\ & O(n^{-1/2} e^{-a_1 n}).
\end{aligned}$$

Applying similar argument for the second partition, we have

$$\begin{aligned}
& \text{pr}\left(\sum_{t=1}^{n} [l(\hat{\Psi}_i^*; y_t, Y_{t-1}, X_t) - l(\hat{\Psi}_{i+1}; y_t, Y_{t-1}, X_t)] I(z_{t-d} \in (\theta_i^0, \tilde{\theta}_i]) > 0\right) \\
=\ & O(n^{-1/2} e^{-a_2 n}),
\end{aligned}$$

for some $a_2 > 0$. Hence,

$$\begin{aligned}
& \text{pr}(\hat{\theta}_i = \tilde{\theta}_i) \\
\le\ & \text{pr}\left(\sum_{t=1}^{n} [l(\hat{\Psi}_i^*; y_t, Y_{t-1}, X_t) - l(\hat{\Psi}_i; y_t, Y_{t-1}, X_t)] I(z_{t-d} \in (\theta_{i-1}^0, \theta_i^0]) > 0\right)
\end{aligned}$$

$$+\operatorname{pr}\left(\sum_{t=1}^{n}[l(\hat{\Psi}_i^*;y_t,Y_{t-1},X_t)-l(\hat{\Psi}_{i+1};y_t,Y_{t-1},X_t)]I(z_{t-d}\in(\theta_{i-1}^0,\tilde{\theta}_i])>0\right)$$
$$=O(n^{-1/2}e^{-a_1 n})+O(n^{-1/2}e^{-a_2 n})=O(n^{-1/2}e^{-\min(a_1,a_2)n})\,. \tag{7.40}$$

Note that a_1 and a_2 in (7.40) are functions of $\tilde{\theta}_i$. As the choices of $\tilde{\theta}_i$ are infinite in regime 1 and $r+1$, we need to show that a_1 and a_2 are positive uniform in all $\tilde{\theta}_i$. Without loss of generality we prove this for regime $r+1$: we verify both $\inf_{\tilde{\theta}_i} a_1(\tilde{\theta}_i)>0$ and $\inf_{\tilde{\theta}_i} a_2(\tilde{\theta}_i)>0$. From (7.39), showing $\inf_{\tilde{\theta}_i} a_1>0$ is equivalent to showing $\inf_{\tilde{\theta}_i\in[\theta_i^0+1,\theta_{i+1}^0)]}\tau(\tilde{\theta}_i)>0$. In addition, from (7.37), it suffices to show

$$\inf_{\tilde{\theta}_i\in[\theta_i^0+1,\theta_{i+1}^0)}\mathrm{E}(\Delta_b(\lambda_s^*,\lambda_s^0))>0\,,$$

and

$$\sup_{\tilde{\theta}_i\in[\theta_i^0+1,\theta_{i+1}^0)}\mathrm{E}[\mathrm{var}(T(y_t)\mid\lambda_s^0)(\gamma(\lambda_s^*)-\gamma(\lambda_s^0))^2/(va_s)^2]<\infty\,.$$

First, by Assumption 2, we have

$$\begin{aligned}&\mathrm{E}\{l(\Psi_i^*;y_t,Y_{t-1},X_t)I(z_{t-d}\in(\theta_{i-1}^0,\tilde{\theta}_i])\}\\ <\;&\mathrm{E}\{l(\Psi_i^0;y_t,Y_{t-1},X_t)I(z_{t-d}\in(\theta_{i-1}^0,\theta_i^0])\}\\ &+\mathrm{E}\{l(\Psi_{i+1}^0;y_t,Y_{t-1},X_t)I(z_{t-d}\in(\theta_i^0,\tilde{\theta}_i])\}\,.\end{aligned}$$

Therefore, for all $z_{t-d}\in(\theta_{i-1}^0,\theta_i^0]$, there exists some $\beta\in(0,1)$ such that

$$\begin{aligned}\mathrm{E}(l(\Psi_i^*;y_t,Y_{t-1},X_t))\;\le\;&(1-\beta)\mathrm{E}\{l(\Psi_i^0;y_t,Y_{t-1},X_t)\}\\ &+\beta\mathrm{E}\{l(\Psi_{i+1}^0;y_t,Y_{t-1},X_t)\}\,.\end{aligned}\tag{7.41}$$

Given $z_{t-d}\in(\theta_{i-1}^0,\theta_i^0]$, by (7.41) and the property of Kullback-Leibler distance, we have

$$\begin{aligned}&\mathrm{E}\{l(\Psi_i^0;y_t,Y_{t-1},X_t)\}-\mathrm{E}\{l(\Psi_i^*;y_t,Y_{t-1},X_t)\}\\ \ge\;&\mathrm{E}\{l(\Psi_i^0;y_t,Y_{t-1},X_t)\}\\ &-[(1-\beta)\mathrm{E}\{l(\Psi_i^0;y_t,Y_{t-1},X_t)\}+\beta\mathrm{E}\{l(\Psi_{i+1}^0;y_t,Y_{t-1},X_t)\}]\\ =\;&\beta[\mathrm{E}\{l(\Psi_i^0;y_t,Y_{t-1},X_t)\}-\mathrm{E}\{l(\Psi_{i+1}^0;y_t,Y_{t-1},X_t)\}]>0\,.\end{aligned}\tag{7.42}$$

By (7.42), $\inf_{\tilde{\theta}_i\in[\theta_i^0+1,\theta_{i+1}^0)}\mathrm{E}(\Delta_b(\lambda_s^*,\lambda_s^0))>0$. On the other hand, by Assumption 4b), we have

$$\begin{aligned}&\sup_{\tilde{\theta}_i\in[\theta_i^0+1,\theta_{i+1}^0)}\mathrm{E}\left[\frac{1}{\nu^2a_s^2}\mathrm{var}(T(y_s)\mid\lambda_s^0)\{\gamma(\lambda_s^*)-\gamma(\lambda_s^0)\}^2\right]\\ =\;&\sup_{\tilde{\theta}_i\in[\theta_i^0+1,\theta_{i+1}^0)}\mathrm{E}\left[\mathrm{var}\left\{l(\Psi_i^*;y_t,Y_{t-1},X_t)-l(\Psi_i^0;y_t,Y_{t-1},X_t)\mid Y_{t-1},X_t\right\}\right]\\ \le\;&\sup_{\tilde{\theta}_i\in[\theta_i^0+1,\theta_{i+1}^0)}\mathrm{E}\left[\mathrm{E}\left\{l(\Psi_i^*;y_t,Y_{t-1},X_t)-l(\Psi_i^0;y_t,Y_{t-1},X_t)\mid Y_{t-1},X_t\right\}^2\right]\end{aligned}$$

$$\begin{aligned} &\leq \mathrm{E}\left[\mathrm{E}\left\{\Gamma^2(Y_{t-1}, X_t, y_t \mid Y_{t-1}, X_t\right\}\right] \sup_{\tilde{\theta}_i \in [\theta_i^0+1, \theta_{i+1}^0)} |\Psi_i^* - \Psi_i^0|^2 \\ &= \mathrm{E}[\Gamma^2(Y_{t-1}, X_t, y_t)] \sup_{\tilde{\theta}_i \in [\theta_i^0+1, \theta_{i+1}^0)} |\Psi_i^* - \Psi_i^0|^2 , \end{aligned} \tag{7.43}$$

where (7.43) is bounded by Assumption 4b) and the compactness of Ω_Ψ in Assumption 1. Therefore, $\inf_{\tilde{\theta}_i \in [\theta_i^0, \theta_{i+1}^0)} \tau > 0$ and hence $\inf_{\tilde{\theta}_i \in [\theta_i^0, \theta_{i+1}^0)} a_1(\tilde{\theta}_i) > 0$. Analogously, $\inf_{\tilde{\theta}_i \in [\theta_i^0, \theta_{i+1}^0)} a_2(\tilde{\theta}_i) > 0$. Since $\hat{\theta}_i$ takes value in $\{z_{t-d}\}$, the number of choices in $\tilde{\theta}_i$ is $O(n)$ increasing. By (7.40),

$$\mathrm{pr}(\hat{\theta}_i > \theta_i^0) = \sum_{\tilde{\theta}_i > \theta_i^0} \mathrm{pr}(\hat{\theta}_i = \tilde{\theta}_i) = O(n^{1/2} e^{-a^* n}) ,$$

for some $a^* > 0$. The derivation of the same convergence rate of $\mathrm{pr}(\hat{\theta}_i < \theta_i^0)$ is analogous, and hence $\mathrm{pr}(\hat{\theta}_i \neq \theta_i^0) = O(n^{1/2} e^{-a_2^* n})$ for some $a_2^* > 0$. If $\nu_i \neq \nu_{i+1}$ for some i, the derivation is similar albeit more tedious. Finally, as $\mathrm{pr}(\hat{\Theta} \neq \Theta^0) \leq \sum_{i=1}^r \mathrm{pr}(\hat{\theta}_i \neq \theta_i^0)$, the result in 2) follows. □

Proof of Theorem 5.3. First we prove the weak convergence of $n(\hat{\theta}_i - \theta_i^0)$ to M_i^-, where $[M_i^-, M_i^+)$ is the unique random interval which minimizes the process

$$\begin{aligned} \ddot{\ell}_i(\kappa_i) = &\sum_{t=1}^n [l(\hat{\Psi}_{i+1}; y_t, Y_{t-1}, X_t) - l(\hat{\Psi}_i; y_t, Y_{t-1}, X_t)] \\ &\times I(z_{t-d^0} \in (\theta_i^0, \theta_i^0 + \frac{\kappa_i}{n}]) I(\kappa_i \geq 0) \\ &+ \sum_{t=1}^n [l(\hat{\Psi}_i; y_t, Y_{t-1}, X_t) - l(\hat{\Psi}_{i+1}; y_t, Y_{t-1}, X_t)] \\ &\times I(z_{t-d^0} \in (\theta_i^0 + \frac{\kappa_i}{n}, \theta_i^0]) I(\kappa_i < 0) , \end{aligned} \tag{7.44}$$

where $|\hat{\theta}_i - \theta_i^0| = |\kappa_i|/n$. When $\hat{\Psi}_i$ and $\hat{\Psi}_{i+1}$ are replaced by Ψ_i^0 and Ψ_{i+1}^0 in (7.44), we define an analogous process $\tilde{\ell}_i(\kappa_i) = \tilde{\ell}_{1,i}(\kappa_i) I(\kappa_i \geq 0) + \tilde{\ell}_{2,i}(\kappa_i) I(\kappa_i < 0)$ where

$$\tilde{\ell}_{1,i}(\kappa_i) = \sum_{t=1}^n [l(\Psi_{i+1}^0; y_t, Y_{t-1}, X_t) - l(\Psi_i^0; y_t, Y_{t-1}, X_t)] I(z_{t-d^0} \in (\theta_i^0, \theta_i^0 + \frac{\kappa_i}{n}]) ,$$

$$\tilde{\ell}_{2,i}(\kappa_i) = \sum_{t=1}^n [l(\Psi_i^0; y_t, Y_{t-1}, X_t) - l(\Psi_{i+1}^0; y_t, Y_{t-1}, X_t)] I(z_{t-d^0} \in (\theta_i^0 + \frac{\kappa_i}{n}, \theta_i^0]) ,$$

which correspond to the case $\hat{\theta}_i \geq \theta_i^0$ and $\hat{\theta}_i < \theta_i^0$, respectively. Denote $\hat{\Psi}^{(0)} = (\hat{\Psi}_1^{(0)}, \ldots, \hat{\Psi}_{r+1}^{(0)})$ as the maximum likelihood estimates given Θ^0. By Assumptions 1–5 and Lemmas 3 and 4 of [54], we have for all $K > 0$,

$$\sup_{|\hat{\theta}_i - \theta_i^0| < K/n} |\hat{\Psi}_i - \hat{\Psi}_i^{(0)}| = o_p(n^{-1/2}) , \tag{7.45}$$

and

$$\sup_{\kappa_i \le K} |\ddot{\ell}(\kappa_i) - \tilde{\ell}(\kappa_i)| = o_p(1)\,. \tag{7.46}$$

By (7.46), it suffices to study the convergence properties of $\tilde{\ell}_i(\kappa_i)$. We show that $\tilde{\ell}_{1,i}(\kappa_i)$ converges to $\tilde{\ell}^*_{1,i}(\kappa_i)$ in distribution, and the convergence of $\tilde{\ell}_{2,i}(\kappa_i)$ follows similarly. We study the right-continuous version of $\tilde{\ell}^*_{1,i}(\kappa_i)$ with assuming $\kappa_i \ge 0$.

First, we construct a piecewise constant interpolation for the difference in log-likelihood. Taking $\epsilon = 1/n$, we let $\eta^\epsilon(v) = X^\epsilon_{\lfloor nv \rfloor}$ for $0 \le v \le 1$, where $\lfloor \cdot \rfloor$ is the floor function. Note that $X^\epsilon_0 = 0, X^\epsilon_{t+1} = X^\epsilon_t + \zeta^\epsilon_{t+1}$. In addition, define

$$\zeta^\epsilon_t = [l(\Psi^0_{i+1}; y_t, Y_{t-1}, X_t) - l(\Psi^0_i; y_t, Y_{t-1}, X_t)] I(z_{t-d^0} \in (\theta^0_i, \theta^0_i + \kappa_i \epsilon])\,.$$

By construction, we have $\eta^\epsilon(1) = \tilde{\ell}_{1,i}(\kappa_i)$, and $\eta^\epsilon(v) = X^\epsilon_s$ when $v \in [s\epsilon, s\epsilon + \epsilon)$. Using similar arguments for the discontinuities of $\tilde{\ell}_{1,i}(\kappa_i)$ on $(u, u+h]$ in the proof of Lemma 3.2 in [25], $\{\tilde{\ell}_{1,i}(\kappa_i)\}$ is tight with $\kappa_i \ge 0$. Combining with the uniform boundedness of $\eta^\epsilon(v)$ which is verified by the truncation argument in [29], we can apply the operator convergence in [29], by which the weak convergence of $\eta^\epsilon(v)$ can be deduced. Then taking $v = 1$, the weak convergence of $\tilde{\ell}_{1,i}(\kappa_i)$ in the Skorohod metric is established.

Next, we show the weak convergence of $\eta^\epsilon(v)$ to the compound Poisson process $\mathcal{C}(v)$ with intensity $\pi(\theta^0_i)\kappa$. Let $\mathcal{F}_v$ be a sequence of σ-algebras that are generated from $\{\eta^\epsilon(u), u \le v\}$, and $\mathcal{L}$ be a set of progressively measurable functions f with respective to $\mathcal{F}_v$, with $\sup_v \mathrm{E}|f(v)| < \infty$. Then, let $\mathcal{F}^\epsilon_v \subseteq \mathcal{F}_v$ denote the σ-algebra generated by $\{\eta^\epsilon(u), u \le v\}$, and E^ϵ_v denote the conditional expectation under $\mathcal{F}^\epsilon_v$. From the definition in [28], we denote $p\text{-}\lim_{\delta\to 0} f^\delta = f$ for $f^\delta \in \mathcal{L}$ if and only if $\lim_{\delta\to 0} \mathrm{E}|f^\delta - f| = 0$. In addition, define the p-infinitesimal operator $\hat{A}^\epsilon$ by $\hat{A}^\epsilon f^\epsilon = p\text{-}\lim_{\delta\to 0}[\mathrm{E}^\epsilon_v f(v+\delta) - f(v)]/\delta$, which is p-right continuous and in $\mathcal{L}$. Denote $\hat{\mathcal{L}}$ as the space of continuous bounded positive function f with $\lim_{v\to\infty} f(v) \to 0$, and $\hat{\mathcal{L}}^{(2)} \subseteq \hat{\mathcal{L}}$ with compact supports and continuous bounded second derivatives. Then, define the operator A on $\hat{\mathcal{L}}^{(2)}$ as

$$Af(v) = \pi_z(\theta^0_i)\kappa_i \int (f(y+v) - f(v)) q(dy)\,,$$

where $q(dy)$ is the induced probability measure of $l(y_t; Y_{t-1}, X_t, \Psi^0_{i+1}) - l(y_t; Y_{t-1}, X_t, \Psi^0_i)$ conditioning on $z_{t-d^0} = (\theta^0_i)^+$. Let $f \in \hat{\mathcal{L}}^{(2)}$, and for any $\tau^\epsilon > 0$, define

$$f^\epsilon(v) = \frac{1}{\tau^\epsilon} \int_0^{\tau^\epsilon} \mathrm{E}^\epsilon_v[f(\eta^\epsilon(v+s))] ds\,.$$

By construction, $f^\epsilon(v) \in \hat{\mathcal{L}}^{(2)}$ and $p\text{-}\lim_{\tau^\epsilon\to 0} f^\epsilon(v) = f(\eta^\epsilon(v))$. Moreover, from [28],

$$\hat{A}^\epsilon f^\epsilon(v) = \frac{1}{\tau^\epsilon}[\mathrm{E}^\epsilon_v\{f(\eta^\epsilon(v+\tau^\epsilon))\} - f(\eta^\epsilon(v))]\,.$$

Next we show the operator convergence of $\hat{A}^\epsilon f^\epsilon(v)$ to $Af(\eta^\epsilon(v))$. Express

$$\hat{A}^\epsilon f^\epsilon(v) = \frac{1}{\tau^\epsilon} \sum_{k=0}^{\lfloor n(v+\tau^\epsilon)\rfloor - \lfloor nv\rfloor - 1} \mathrm{E}_v^\epsilon\{f(\eta^\epsilon(v+k\epsilon) + \zeta^\epsilon_{k+\lfloor nv\rfloor+1}) - f(\eta^\epsilon(v+k\epsilon))\}\,,$$

and

$$Af(\eta^\epsilon(v)) = \frac{1}{\tau^\epsilon} \sum_{k=0}^{\lfloor n(v+\tau^\epsilon)\rfloor - \lfloor nv\rfloor - 1} \mathrm{E}_v^\epsilon\{f(\eta^\epsilon(v) + \zeta^\epsilon_{k+\lfloor nv\rfloor+1}) - f(\eta^\epsilon(v))\}\,.$$

Without loss of generality, we illustrate the proof with $f(y \mid Y_{t-1}, X_t, z_{t-d})$ is continuous in y, where the derivation for general case is analogous. Denote $m^\epsilon = \lfloor n(v+\tau^\epsilon)\rfloor - \lfloor nv\rfloor$, by Theorem 15.3 of [6], we have

$$\hat{A}^\epsilon f^\epsilon(v) = \sum_{k=0}^{m^\epsilon-1} \mathrm{E}_v^\epsilon\{f(\eta^\epsilon(v) + \zeta^\epsilon_{k+\lfloor nv\rfloor+1}) - f(\eta^\epsilon(v))\}/\tau^\epsilon + o_p(1)\,. \tag{7.47}$$

Meanwhile, for sufficiently large n and $\epsilon = 1/n$,

$$\mathrm{pr}(z_{t-d^0} \in (\theta_i^0, \theta_i^0 + \kappa_i\epsilon]) = \mathrm{pr}(z_{t-d^0} = (\theta_i^0)^+)\kappa_i\epsilon(1+o(1)) = \frac{1}{n}\pi_z(\theta_i^0)\kappa_i(1+o(1))\,.$$

Thus, together with Assumptions 4a) and 5 and equation (48) in Appendix M, [7], for all fixed X with ρ-mixing coefficient sequence $\{\rho(k)\}$, there exists $K^* \geq \mathrm{E}|f(X + \zeta^\epsilon_{k+\lfloor nv\rfloor+1}) - f(X)|$ such that

$$\begin{aligned}
&\frac{1}{\tau^\epsilon}\sum_{k=0}^{m^\epsilon-1} \mathrm{E}\left[\left|\mathrm{E}_v^\epsilon\{f(X+\zeta^\epsilon_{k+\lfloor nv\rfloor+1}) - f(X)\} - \mathrm{E}\{f(X+\zeta^\epsilon_{k+\lfloor nv\rfloor+1}) - f(X)\}\right|\right] \\
&\leq \frac{1}{\tau^\epsilon}\left[\mathrm{E}\{f(X+\zeta^\epsilon_{k+\lfloor nv\rfloor+1}) - f(X)\}^2\right]^{1/2} \sum_{k=0}^{m^\epsilon-1}\rho(k+1) \\
&\leq \frac{1}{\tau^\epsilon}K^*\left\{\mathrm{pr}(Z_{t-d_0} \in (\theta_i^0, \theta_i^0+\kappa_i\epsilon])\right\}^{1/2} \sum_{k=0}^{m^\epsilon-1}\rho(k+1) \\
&= (1+o(1))\frac{K^*}{\sqrt{n\tau^\epsilon}}\left(\pi_z(\theta_i^0)\kappa_i\right)^{1/2} \sum_{k=0}^{m^\epsilon-1}\rho(k+1) \to 0\,,
\end{aligned}$$

by taking $\tau^\epsilon = n^{-b}$ for some $b \in (0, 1/2)$ and $m^\epsilon \to \infty$. Combining with (7.47), we have

$$\hat{A}^\epsilon f^\epsilon(v) = Af(\eta^\epsilon(v)) + o_p(1)\,. \tag{7.48}$$

As $p\text{-lim}_{\tau^\epsilon\to 0} f^\epsilon(v) = f(\eta^\epsilon(v))$, by Theorem 1 of [29], $\eta^\epsilon(v)$ converges weakly to $\mathcal{C}(v)$. On the other hand, $\eta^\epsilon(v)$ is the uniquely solution such that $f(\eta(t)) - \int_0^t Af(\eta(s))ds$ is a martingale for any f with continuous second-derivative and

a compact support ([57])). Choose $v = 1$, then $\tilde{\ell}_{1,i}(\kappa_i)$ converges in distribution to $\tilde{\ell}^*_{1,i}(\kappa_i)$ in the Skorohod metric. Analogously, we have the weak convergence of $\tilde{\ell}_{2,i}(\kappa_i)$ to $\tilde{\ell}^*_{2,i}(\kappa_i)$ in the Skorohod metric; by Cramer–Wold device, the weak convergence of $\tilde{\ell}_i(\kappa_i)$ to $\tilde{\ell}^*_i(\kappa_i)$ is obtained. Thus, Combining with Theorem 3.1 in [55], $n(\hat{\theta}_i - \theta^0_i)$ converges weakly to M^-_i, where $[M^-_i, M^+_i)$ is the unique random interval that minimizes $\tilde{\ell}^*_i(\kappa_i)$.

Next we consider the convergence of $\hat{\Theta}$. Analogously, we construct $\tilde{\ell}(\kappa) = \sum_{i=1}^r \tilde{\ell}_i(\kappa_i)$. Recall that for $\tilde{\ell}^*_i(\kappa_i)$, as suggested by (5.3) and (5.4), it has an intensity $\pi_z(\theta^0_i)$ and jumps

$$\zeta_s(y^*_s; Y^*_{s-1}, X^*_s, \Psi^0_{i+1}, \Psi^0_i) I(\kappa_i \geq 0) + \zeta_s(y^*_s; Y^*_{s-1}, X^*_s, \Psi^0_i, \Psi^0_{i+1}) I(\kappa_i < 0)\,,$$

with a non-degenerate distribution. Meanwhile, we have deduced the weak convergence of $\tilde{\ell}_i(\kappa_i)$ to $\tilde{\ell}^*_i(\kappa_i)$ from operator convergence as (7.48). Employing the idea in Theorem 3.3 of [32], for any constants c_1 and c_2 with either one nonzero, and any vectors $\kappa^{(1)}, \ldots, \kappa^{(4)} = \{\kappa^{(1)}_1, \ldots, \kappa^{(1)}_r\}, \ldots, \{\kappa^{(4)}_1, \ldots, \kappa^{(4)}_r\}$ which are not all equal, the process

$$c_1[\tilde{\ell}(\kappa^{(1)}) - \tilde{\ell}(\kappa^{(2)})] + c_2[\tilde{\ell}(\kappa^{(3)}) - \tilde{\ell}(\kappa^{(4)})]$$

has a positive jump rate with bounded and non-degenerated jump sizes. Furthermore, operator convergence results analogous to (7.48) can be shown for this process, as both A and $\hat{A}^\epsilon$ are linear operators. Thus, by similar arguments in the proof of Theorem 3.3 in [32], the weak convergence of $\tilde{\ell}(\kappa)$ to the compound Poisson process $\tilde{\ell}^*(\kappa)$ is established. Using Theorem 3.1 in [55] again, $n(\hat{\Theta} - \Theta^0)$ converges weakly to M^-, where $[M^-, M^+)$ is the unique minimizer of $\tilde{\ell}^*(\kappa)$.

In the general case, as $\{\eta^\epsilon(v), 0 \leq v \leq 1\}$ is tight, every subsequence has a convergent subsequence, where the convergence can be assumed almost surely by enlarging the probability space. Moreover, as (7.47)–(7.48) hold, by using similar arguments as in [54], $n(\hat{\Theta} - \Theta^0)$ converges weakly to M^-, where $[M^-, M^+)$ is an almost surely minimizer of $\tilde{\ell}^*(\kappa)$. □

Proof of Theorem 5.4. For continuous z_t, the asymptotic independence between $n(\hat{\Theta} - \Theta^0)$ and $n^{1/2}(\hat{\Psi} - \Psi^0)$ follows from $O_p(n^{-1})$ convergence rate of $\hat{\Theta}$, (7.45) and arguments of proving Theorem 2 of [10]. For discrete z_t, as we have obtained a very rapid convergence of $\hat{\Theta}$ to Θ^0 in probability, the asymptotic independence between $n(\hat{\Theta} - \Theta^0)$ and $n^{1/2}(\hat{\Psi} - \Psi^0)$ can be similarly established.

Next we show the asymptotic distribution of model parameter estimates using techniques in [62]. Under Assumption 2 and the strong consistency of model parameters estimates, the estimator $\hat{\Psi}$ satisfies

$$\begin{aligned} L'(\hat{\Psi}, \hat{\Theta}, d^0) &= \left(\frac{\partial L_{n,1}(\hat{\Psi}_1, \theta_0, \hat{\theta}_1, d^0)}{\partial \Psi_1}, \ldots, \frac{\partial L_{n,r^0+1}(\hat{\Psi}_{r^0+1}, \hat{\theta}_{r^0}, \theta_{r^0+1}, d^0)}{\partial \Psi_{r^0+1}} \right)^{\mathrm{T}} \\ &= 0. \end{aligned}$$

By Taylor's expansion,

$$\begin{aligned} L'(\hat{\Psi},\hat{\Theta},d^0) &= L'(\Psi^0,\hat{\Theta},d^0)+L''(\Psi^0,\hat{\Theta},d^0)(\hat{\Psi}-\Psi^0)+o_p(1) \\ &= L'(\Psi^0,\Theta^0,d^0)+L''(\Psi^0,\Theta^0,d^0)(\hat{\Psi}-\Psi^0)+o_p(1)\,, \end{aligned}$$

where the second-order partial derivative is symmetric. Next we verify the regularity conditions in Theorem 5.41 of [62]. First, from (4.1), for all $i \neq j$ we have

$$\frac{\partial^2 L(\Psi,\Theta^0,d^0)}{\partial\Psi_i\partial\Psi_j}=0\,,\quad \mathrm{E}\left[\frac{\partial L(\Psi,\Theta^0,d^0)}{\partial\Psi_i}\left(\frac{\partial L(\Psi,\Theta^0,d^0)}{\partial\Psi_j}\right)^{\mathrm{T}}\right]=0\,.$$

Thus, $L''(\Psi^0,\Theta^0,d^0)$ and $\mathrm{E}\left[L'(\Psi^0,\Theta^0,d^0)\{L'(\Psi^0,\Theta^0,d^0)\}^{\mathrm{T}}\right]$ are block-diagonal. In addition, by ergodic theorem,

$$\left\{L''(\Psi^0,\Theta^0,d^0)\right\}^{-1}\longrightarrow\left[\mathrm{E}\left\{L''(\Psi^0,\Theta^0,d^0)\right\}\right]^{-1},$$

in probability. From Assumption 2, we have the finiteness of $\mathrm{E}[\{\partial L(\Psi^0,\Theta^0,d^0)/\partial\Psi_i\}^2]$ and $\mathrm{E}[L'(\Psi^0,\Theta^0,d^0)\{L'(\Psi^0,\Theta^0,d^0)\}^{\mathrm{T}}]$. Furthermore, Assumption 2 implies the boundedness of third-order derivatives of $L(\Psi,\Theta^0,d^0)$ with respect to Ψ by some integrable function in the neighborhood of Ψ^0. Therefore, as all regularity conditions are satisfied, by Theorem 5.41 in [62], we have $n^{1/2}(\hat{\Psi}-\Psi^0)\sim N(0,\Sigma^*)$, where

$$\begin{aligned}\Sigma^* = &\left[\mathrm{E}\{L''(\Psi^0,\Theta^0,d^0)\}\right]^{-1}\mathrm{E}\left[L'(\Psi^0,\Theta^0,d^0)\{L'(\Psi^0,\Theta^0,d^0)\}^{\mathrm{T}}\right]\\ &\times\left[\mathrm{E}\{L''(\Psi^0,\Theta^0,d^0)\}\right]^{-1}.\end{aligned}$$

As $L'(\Psi^0,\Theta^0,d^0)$ is continuous in Ψ for sufficiently large n, by information matrix equality,

$$\mathrm{E}\left[L'(\Psi^0,\Theta^0,d^0)\{L'(\Psi^0,\Theta^0,d^0)\}^{\mathrm{T}}\right]=-\left[\mathrm{E}\{L''(\Psi^0,\Theta^0,d^0)\}\right]^{-1}.$$

Hence, Σ^* is defined as (5.5) and is diagonal. In particular, $|\hat{\Psi}-\Psi^0|=O_p(n^{-1/2})$. □

References

[1] Alba, E. & Troya, J. M. (1999). A survey of parallel distributed genetic algorithms. *Complexity* **4**, 31–52 MR1688681

[2] Alba, E. & Troya, J. M. (2002). Improving flexibility and efficiency by adding parallelism to genetic algorithms. *Statistics and Complexity* **12**, 91–114 MR1897509

[3] An, H. Z. & Huang, F. C. (1996). The geometrical ergodicity of nonlinear autoregressive models. *Statist. Sin.* **6**, 943–56 MR1422412

[4] Bradley, R. C. (2005). Basic Properties of Strong Mixing Conditions. A Survey and Some Open Questions. *Probab. Surv.* **2**, 107–44 MR2178042

[5] Billingsley, P. (1961). The Lindberg-Lévy theorem for martingales. *Proc. Amer. Math. Soc.* **12**, 788–92 MR0126871
[6] Billingsley, P. (1968). *Convergence of Probability Measures*. New York: Wiley. MR0233396
[7] Billingsley, P. (1999). *Convergence of Probability Measures* 2nd ed. New York: Wiley. MR1700749
[8] Cai, Y. & Stander, J. (2008). Quantile of self-excited threshold autoregressive time series models *J. Time Series Anal.* **29**, 186–202 MR2387486
[9] Chan, K. S. (1990). Deterministic stability, stochastic stability and ergodicity. In *Non-Linear Time Series: A Dynamical System Approach*, Ed. Tong, H. Oxford: Clarendon Press, Appendix 1 MR1079320
[10] Chan, K. S. (1993). Consistency and limiting distribution of the least square estimator of a threshold autoregressive model. *Ann. Statist.* **21**, 520–33 MR1212191
[11] Chan, K. S. & Tong, H. (1985). On the use of the deterministic Lyapunov function for the ergodicity of stochastic difference equations. *Adv. Appl. Prob.* **17**, 666–78 MR0798881
[12] Chan, K. S. & Tong, H. (1990). On likelihood ratio test for threshold autoregression. *J. R. S. S. B* **52**,469–76 MR1086798
[13] Chan, K. S. & Tsay, R. S. (1998) Limiting properties of least square estimator of a continuous threshold model . *Biometrika* **85**, 413–26 MR1649122
[14] Chan, K. S. & Goracci, G. (2019). On the ergodicity of first-order threshold autoregressive moving-average processes positive. *J. Time Series Anal.* **40**, 256–264 MR3915530
[15] Chan, N. H., Yau, C. Y. & Zhang, R. M. (2015). LASSO estimation for threshold autoregressive models. *J. Econometrics* **189**, 285–96 MR3414900
[16] Chan, N. H., Ing, C. K., Li, Y. & Yau, C. Y. (2017). Threshold estimation via group orthogonal greedy algorithm. *J. Bus. Econom. Stat.* **35**, 334–345 MR3622841
[17] Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton: Princeton University Press MR0016588
[18] Davis, R. A., Lee, T. C. M. & Rodriguez-Yam, G. A. (2006). Structural break estimation for nonstationary time series models. *J. Amer. Statist. Assoc.* **101**, 223–39 MR2268041
[19] Davis, R. A. & Yau, C. Y. (2013). Consistency of minimum description length model selection for piecewise stationary time series models. *Electron. J. Stat.* **7**, 381–411 MR3020426
[20] Davis, R. A., Lee, T. C. M. & Rodriguez-Yam, G. A. (2008). Break detection for a class of nonlinear time series models. *J. Time Series Anal.* **29**, 834–67 MR2450899
[21] Doukhan, P. (1994). *Mixing: Properties and Examples*. New York: Springer-Verlag MR1312160
[22] Fokianos, K., Rahbek, A. & Tjøtheim. D. (2009). Poisson autoregression. *J. Amer. Statist. Assoc.* **104**, 1430–9 MR2596998
[23] Gao, X., Ritter J. R. & Zhu, Z. (2013). Where have all IPO gone? *J. Fin. Quan. Anal.* **48**, 1663–92

[24] Ibragimov, I. A. (1963). A central limit theorem for a class of dependent random variables. *Theory of Probability & Its Application* **8**, 83–9 MR0151997
[25] Ibrangimov, I. A. & Khasminskii, R. Z. (1981). *Statistical Estimation: Asymptotic Theory*. New York: Springer-Verlag MR0620321
[26] Konishi, S. & Kitagawa, G. (2008). *Information Criteria and Statistical Modeling*. New York: Springer Science+Business Media, LLC MR2367855
[27] Koop, G. & Potter, S. M. (2003) Bayesian analysis of endogenous delay threshold models. *J. Bus.Econom. Stat.* **21**, 93–103 MR1950381
[28] Kurz, T. G. (1975). Semigroups of conditioned shifts and approximation of Markov process. *Ann. Prob.* **3**, 618–42 MR0383544
[29] Kushner, H. J. (1980). A martingale method for the convergence of a sequence of processes to a jump-diffusion process. *Z. Wahrscheinlichkeitstheorie verw. Gebiete* **53**, 207–19 MR0580914
[30] Lee, T. C. M. (2000). An introduction to coding theory and the two-part minimum description length principle. *Int. Stat. Rev.* **69**, 169–83
[31] Li, D., Li, W. K. & Ling, S. (2011) On the least square estimation of threshold autoregressive and moving-average models. *Statistics and Its Inference* **4**, 183–196 MR2812814
[32] Li, D. & Ling, S. (2012). On the least square estimation of multiple-regime threshold autoregressive models. *J. Econometrics* **167**, 240–53 MR2885449
[33] Li, D., Ling, S. & Zakoïan, J.-M. (2015). Asymptotic inference in multiple-threshold double autoregressive model. *J. Econometrics* **189**, 415–27 MR3414910
[34] Li, D., Ling, S. & Zhang, R. (2016). On a Threshold Double Autoregressive Model. *J. Bus. Econom. Statist.* **34**, 161–75 MR3450051
[35] Li, G. & Li, W. K. (2011) Testing a linear time series model against its threshold extension. *Biometrika* **98**, 243–50 MR2804225
[36] Li, G., Guan, B., Li, W. K. & Yu, P. L. H. (2015) Hysteretic autoregressive time series model. *Biometrika* **102**, 717–23 MR3394287
[37] Liu, J. & Susko, ED (1992) On strict stationarity and ergodicity of a non-linear ARMA model. *J. Appl. Prob* **29**, 363–373 MR1165221
[38] Ling, S. (1999). On the probabilistic properties of a double threshold ARMA conditional heteroskedastic model. *J. Appl. Prob.* **36**, 688–705 MR1737046
[39] Lowry, M. and Schwert, G. W. (2002). IPO market cycles: bubbles or sequential earning? *J. Financ.* **3**, 1171–200
[40] Lowry, M. (2003). Why does IPO volume fluctuate so much? *J. Financ. Econom.* **67**, 3–40
[41] Lu, Q., Lund, R. & Lee, T. C. M. (2010). An MDL approach to the climate segmentation problem. *Ann. Appl. Statist.* **4**, 299–319 MR2758173
[42] McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*, 2nd Ed. London: Chapman and Hall MR3223057
[43] Meyn S. P. & Tweedie R. L. (1993). *Markov Chains and Stochastic Stability*. London: Springer-Verlag. MR1287609
[44] Newey, W. K. & McFadden, D. (1994). Large Sample Estimation and

Hypothesis Testing. In *Handbook of Econometrics, Vol.4,* Ed. Engle, R and McFadden, D. Elsevier Science. MR1315971

[45] Pástor, Ľ. & Veronesi, P. (2005). Rational IPO waves. *J. Financ.* **60**, 1713–57

[46] Petrov, V. V.(1975). *Sums of Independent Random Variables.* Berlin: Springer-Verlag MR0388499

[47] Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry.* Singapore: World Scientific MR1082556

[48] Rissanen, J. (2007). *Information and Complexity in Statistical Modeling.* Springer. MR2287233

[49] Ritter, J. R. (1984). The "hot issue" market of 1980. *J. Bus.* **57**, 215–40

[50] Ritter, J. R. & Welch, I. (2002). A review of IPO activity, pricing and allocations. *J. Financ.* **4**, 1795–828

[51] Ritter, J. R. (2013) Historical US IPO Statistics *Quandl.com* Retrieved June, 2015, from *https://www.quandl.com/data/RITTER/US_IPO_STATS-Historical-US-IPO-Statistics*

[52] Saïdi, Y. & Zakoïan J.-M. (2006). Stationarity and Geometric Ergodicity of a Class of Nonlinear ARCH Models. *Ann. Appl. Probab.*, **4**, 2256–71 MR2288721

[53] Samia, N. I., Chan, K. S. & Stenseth, N. C. (2007). A generalized threshold mixed model for analyzing nonnormal nonlinear time series, with application to plague in Kazakhstan. *Biometrika* **94**, 101–18 MR2367827

[54] Samia, N. I. & Chan, K. S. (2011). Maximum likelihood estimation of a generalized threshold stochastic regression model. *Biometrika* **98**, 433–48 MR2806439

[55] Seijo, E. & Sen, B. (2011) A continuous mapping theorem for the smallest argmax functional. *Electron. J. Stat.* **5**, 421–39 MR2802050

[56] So, M. K. P., Li, W. K. & Lam, K. (2002). A Threshold Stochastic Volatility Model. *J. Forecast.* **21**, 473–500

[57] Strook, D. W. & Varadhan, S. R. S. (1971). Diffusion process with boundary conditions. *Comm. Pure. Appl. Math.* **24**, 147–225 MR0277037

[58] Tong, H. (1978). On a threshold model. In *Pattern Recognition and Signal Processing. NATO ASI series E: Applied Sc. (29).* (ed. C.Chen), 575–86. Amsterdam: Sijthoff & Noordhoff

[59] Tong, H. (1990). *Non-Linear Time Series. A Dynamical System Approach.* New York: Oxford University Press MR1079320

[60] Tong, H. (2007). Birth of the threshold time series model. *Statist. Sinica.* **17**, 8–14

[61] Tong, H. (2011). Threshold models in time series analysis-30 years on. *Stat. Interface* **4**, 107–18 MR2812802

[62] van der Vaart, A. W. (2000). *Asymptotic Statistics.* Cambridge: Cambridge University Press MR1652247

[63] Wong, C. S.& Li, W. K.(1997) Testing for threshold autoregression with conditional heteroskedasticity. *Biometrika* **84**, 407–18 MR1467056

[64] Wu, Y. & Ling, S. (2017). Using genetic algorithms to parameters (d; r) estimation for threshold autoregressive models. *Computational Statistics*

and Data Analysis **35**, 318–333

[65] Yang, Y. & Ling, S. (2017). Inference for heavy-tailed and Multiple-Threshold Double Autoregressive Models. *J. Bus. Econom. Statist.* **35**, 318–333 MR3622840

[66] Yau, C. Y., Tang, C. M. & Lee, T. C. M. (2015). Estimation of multiple-regime threshold autoregressive models with structural breaks. *J. Amer. Statist. Assoc.* **110**, 1175–86 MR3420693

[67] Yu, P.(2012). Likelihood estimation and inference in threshold regression. *J. Econometrics* **167**, 274–94 MR2885451

[68] Yung, C., Çolak, G. & Wang, W. (2008). Cycles in the IPO market. *J. Financ. Econ.* **89**, 192–208