FUNCTIONAL THRESHOLD AUTOREGRESSIVE MODEL

Yuanbo Li[1], Kun Chen[2*], Xunze Zheng[3] and Chun Yip Yau[3]

[1] *University of International Business and Economics*

[2] *Southwestern University of Finance and Economics*

[3] *and The Chinese University of Hong Kong*

Supplementary Material

The Supplementary Material discusses further probabilistic properties of fTAR and fTARX models in Section A, proofs of the asymptotic theory for statistical inferences in Section B, the fMDL estimation procedure of fTARX models in Section C, the genetic algorithm for optimization in Section D, simulation studies in Section E, construction of confidence intervals for thresholds and model parameters in Section F, additional information for the real data analysis in Section G, and the extension of the proposed procedure to infinite-dimensional parameter space in Section H.

A: Further probabilistic properties

A.1 β-mixing of fTAR processes

In this section we study the β-mixing property of the fTAR process $\{Y_k\}$ in (2.1). First, we give the definition of β-mixing of a Markov process.

Definition 1. (Davydov (1973); Doukhan (1994)) *β-mixing:* If $\{Y_t\}$ is a time-homogeneous Markov process with a state space $(\Omega, \mathcal{B})$ and an invariant probability measure $\pi(\cdot)$, then it is β-mixing with an exponentially decaying rate if there is some constant $c > 0$ and $\rho \in (0, 1)$, such that

$$\int \|\mathrm{pr}(Y_{t+n} \in A \mid Y_t = x) - \mathrm{pr}(Y_{t+n} \in A)\|_{TV} \pi(dx) \leq c\rho^n, \ \forall A \in \mathcal{B}, \ n \in \mathbb{Z}^+ ,$$

where $\|\mu - \nu\|_{TV} = 2 \sup_{A \in \mathcal{B}} |\mu(A) - \nu(A)|$ is the total-variation distance between two probability measures μ and ν. In addition, if the initial value Y_0 is generated from $\pi(\cdot)$, β-mixing of the Markov process Y_t implies its strict stationarity and ergodicity.

Theorem S.1. *If* (3.5) *holds and* $\{Y_k\}$ *is spanned by finitely many basis functions, then* $\{Y_k\}$ *of* (2.1) *is β-mixing with an exponentially decaying rate.*

Moreover, if specific knowledge of the threshold variable z_{k-d} is available, less restrictive conditions compared to (3.5) can be established as follows.

Corollary S.1. *Denote*

$$\gamma = \rho_1(1 - \pi(y^*)) + \rho_2\pi(y^*)\,,$$

where $\rho_1 = \max_{i=1,r} \sum_{j=1}^{p_{Y,i}} \|\Psi_{i,j}\|_{\mathcal{L}}$, $\rho_2 = \max_{i=2,\ldots,r-1} \sum_{j=1}^{p_{Y,i}} \|\Psi_{i,j}\|_{\mathcal{L}}$, *and* $\pi(y^*) = \sup_{y^*} \mathrm{pr}(g(Y_k) \in (\theta_1, \theta_{r-1}] \mid Y_{k-1}^* = y^*)$. *If* $z_{k-d} = g(Y_{k-d})$ *for some measurable functional* g, $0 < \rho_1 < 1$ *and* $\gamma < 1$, *then* β*-mixing holds.*

On the other hand, denote

$$\tilde{\gamma} = \sum_{i=1}^{r} \sum_{j=1}^{p_{Y,i}} \|\Psi_{i,j}\|_{\mathcal{L}} \,\mathrm{pr}(z_{k-d} \in (\theta_{i-1}, \theta_i])\,.$$

If $z_{k-d} = X_{k-d,m}$ *is a scalar exogenous variable and* $\tilde{\gamma} < 1$, *then* β*-mixing holds.*

The condition of finitely many basis functions is required for applying the drift condition method in Meyn and Tweedie (1993), which is only applicable to a state space with countably generated sigma-fields, such as $\mathbb{R}^d$, but not for $\mathbb{R}^\infty$. Nevertheless, the β-mixing property is not required in establishing the stationary and ergodicity of the fTAR process, and the consistency of parameter estimation in Section 5. It is only used to for deriving the asymptotic distribution of parameter estimates.

A.2: Extensions to fTARX processes

For the fTARX process $\{Y_k\}$ in (2.4), all theorems and corollaries in Section 3 remain valid under the conditions that the exogenous variables $X_{k,m}$ for $m = 1, \ldots, p_X$ are i.i.d. with finite fourth moment, and $X_{k,m}$ is independent of $\{Y_s\}_{s \leq k-1}$. These conditions are imposed for proving the ergodicity by using the techniques in Stout (1974). In addition, if we abuse notation by using $Y_k^* = (Y_k, \ldots, Y_{k-p+1})^{\mathrm{T}}$ and assume $\{X_{k,m}\}$ is spanned by finitely many basis functions, the conditions on $X_{k,m}$ ensure the Markovian property of Y_k^* in proving the β-mixing property, which in turn implies the β-mixing of Y_k. In general, these conditions are difficult to be relaxed in verifying the Markovian property of Y_k^* due to the complicated interrelationship between $X_{k,m}$ and Y_k. In practice, if the exogenous variables are dependent, then one may apply a prewhitening procedure to obtain approximately independent covariates.

A.3: Proofs of results from Section 3 and Section A.1

Proof of Theorem 1

Without loss of generality, assume $\sigma_i = \sigma$ and $a_i = a$ for some constants σ and a, respectively, and hence $a_i^* = a^* = (a, 0, \ldots, 0)^{\mathrm{T}}$ for all i. The derivation of the general case follows the same procedure but is much tedious

in labeling, and hence is omitted. In addition, assume $d \leq p$ by taking $\Psi_{i,j} = 0$ for $j = p+1, \ldots, d$.

First, as $\tilde{\Psi}_k^* = \sum_{i=1}^r \Psi_i^* I(z_{k-d} \in (\theta_{i-1}, \theta_i])$ by using the definition $\tilde{\Psi}_k^* = \sum_{i=1}^r \Psi_i^* I(z_{k-d} \in (\theta_{i-1}, \theta_i])$, we have

$$\begin{aligned} Y_k^* &= a^* + \tilde{\Psi}_{k-1}^*(Y_{k-1}^*) + \sigma\epsilon_k^* \\ &= a^* + \left(\tilde{\Psi}_{k-1}^* \circ \tilde{\Psi}_{k-2}^*\right)(Y_{k-2}^*) + \tilde{\Psi}_{k-1}^*(a^* + \sigma\epsilon_{k-1}^*) + \sigma\epsilon_k^* \\ &= \sum_{l=1}^{\infty} \left(\tilde{\Psi}_{k-1}^* \circ \tilde{\Psi}_{k-2}^* \circ \cdots \circ \tilde{\Psi}_{k-l}^*\right)(a^* + \sigma\epsilon_{k-l}^*) + a^* + \sigma\epsilon_k^* . \quad \text{(S.1)} \end{aligned}$$

By the definition of operator norm, it can be derived that

$$\|\tilde{\Psi}_{i_2}^* \circ \tilde{\Psi}_{i_1}^*\|_{\mathcal{L}_p} \leq \|\tilde{\Psi}_{i_2}^*\|_{\mathcal{L}_p} \|\tilde{\Psi}_{i_1}^*\|_{\mathcal{L}_p} .$$

Hence, similar to Lemma 3.1 of Bosq (2000), with some $u \geq 1$, for any $l = ua_1 + a_2$ where $a_1 \geq 1$ is nonnegative integer and $0 \leq a_2 < u$, we have

$$\max_{i_1,\ldots,i_l} \|\tilde{\Psi}_{i_l}^* \circ \cdots \circ \tilde{\Psi}_{i_1}^*\|_{\mathcal{L}_p} \leq \left(\max_{i_1,\ldots,i_u} \|\tilde{\Psi}_{i_u}^* \circ \cdots \circ \tilde{\Psi}_{i_1}^*\|_{\mathcal{L}_p}\right)^{a_1} \max_{i_1,\ldots,i_{a_2}} \|\tilde{\Psi}_{i_{a_2}}^* \circ \cdots \circ \tilde{\Psi}_{i_1}^*\|_{\mathcal{L}_p} \leq cb^l ,$$

where $c = (\max_{i_1,\ldots,i_u} \|\tilde{\Psi}_{i_u}^* \circ \cdots \circ \tilde{\Psi}_{i_1}^*\|_{\mathcal{L}_p})^{-1} \max_{i_1,\ldots,i_{a_2}} \|\tilde{\Psi}_{i_{a_2}}^* \circ \cdots \circ \tilde{\Psi}_{i_1}^*\|_{\mathcal{L}_p}$ and $b = \max_{i_1,\ldots,i_u} \|\tilde{\Psi}_{i_u}^* \circ \cdots \circ \tilde{\Psi}_{i_1}^*\|_{\mathcal{L}_p}^{1/u}$. By direct calculations, $\max_{i_1,\ldots,i_u} \|\tilde{\Psi}_{i_u}^* \circ \cdots \circ \tilde{\Psi}_{i_1}^*\|_{\mathcal{L}_p} \leq \max_{i_1,\ldots,i_u} \|\Psi_{i_u}^* \circ \cdots \circ \Psi_{i_1}^*\|_{\mathcal{L}_p} < 1$. Since $a_1 > l/u - 1$ and $\max_{i_1,\ldots,i_u} \|\tilde{\Psi}_{i_u}^* \circ \cdots \circ \tilde{\Psi}_{i_1}^*\|_{\mathcal{L}_p} < 1$, we have $c > 0$ and $0 < b < 1$, i.e., if (3.5) holds, then

$$\exists c > 0 \,, 0 < b < 1 \,, \max_{i_1,\ldots,i_u} \|\tilde{\Psi}_{i_u}^* \circ \cdots \circ \tilde{\Psi}_{i_1}^*\|_{\mathcal{L}_p} \leq cb^u \,, \forall u > 0 \,. \quad \text{(S.2)}$$

By (S.2), $\max_k \sum_{l=1}^{\infty} \|\tilde{\Psi}^*_{k-1} \circ \cdots \circ \tilde{\Psi}^*_{k-l}\|^2_{\mathcal{L}_p} < \infty$. Thus, with this summability of product norms, if $z_{k-d} = X_{k-d,m}$, then the strict stationarity of $\{Y^*_k\}$ follows from (S.1). If $z_{k-d} = g(Y_{k-d})$, then by (S.1), we define $z_{k-d} = g(Y_{k-d}) = g((Id, 0, \ldots, 0)Y^*_{k-d})$, and hence $\tilde{\Psi}^*_k$ is measurable with respect to the sigma-field generated by $\{\epsilon^*_{k-d}, \epsilon^*_{k-d-1}, \ldots\}$. Since ϵ^*_ks are i.i.d, the strict stationarity of $\{Y^*_k\}$ can be shown by the same arguments. Therefore, $\{Y^*_k\}$ and thus $\{Y_k\}$ are strictly stationary.

In addition, the orthogonality of ϵ^*_k entails that

$$\Delta^{m'}_m := \mathrm{E}\left\|\sum_{l=m}^{m'} (\tilde{\Psi}^*_{i_l} \circ \cdots \circ \tilde{\Psi}^*_{i_1})(a^* + \sigma\epsilon^*_{k-l})\right\|^2_p = \sum_{l=m}^{m'} \mathrm{E}\left\|(\tilde{\Psi}^*_{i_l} \circ \cdots \circ \tilde{\Psi}^*_{i_1})(a^* + \sigma\epsilon^*_{k-l})\right\|^2_p,$$

for $1 \le m < m'$. Hence, (S.2) yields that

$$\Delta^{m'}_m \le \sum_{l=m}^{m'} \left\|\tilde{\Psi}^*_{i_l} \circ \cdots \circ \tilde{\Psi}^*_{i_1}\right\|^2_{\mathcal{L}_p} (\|a\|^2 + \sigma^2) \to 0 \quad \text{as } m \text{ and } m' \to \infty\,.$$

Thus from the Cauchy criterion, it follows that Y^*_k converges in $L^2_{H_p}$. In fact, since $\mathrm{E}(\sum_{l=1}^{\infty}(\tilde{\Psi}^*_{i_l} \circ \cdots \circ \tilde{\Psi}^*_{i_1})(a^* + \sigma\epsilon^*_{k-l}) + a^* + \sigma\epsilon^*_k)^2 < \infty$, it follows that the series Y^*_k also converges almost surely. By Stout (1974), as $Y^*_k \in L^2_{H_p}$ almost surely and ϵ^*_ks and $X_{k,m}$s are i.i.d., we have that $\{Y^*_k\}$ and thus $\{Y_k\}$ are ergodic.

For the general case of functional TAR series, we may write $\tilde{a}^*_k = \sum_{i=1}^{r} a^*_i I(z_{k-d} \in (\theta_{i-1}, \theta_i])$, $\tilde{\sigma}_k = \sum_{i=1}^{r} \sigma_i I(z_{k-d} \in (\theta_{i-1}, \theta_i])$, and

$$Y^*_k = \sum_{l=1}^{\infty} \left(\tilde{\Psi}^*_{k-1} \circ \tilde{\Psi}^*_{k-2} \circ \cdots \circ \tilde{\Psi}^*_{k-l}\right)(\tilde{a}^*_{k-l} + \tilde{\sigma}_{k-l}\epsilon^*_{k-l}) + \tilde{a}^*_k + \tilde{\sigma}_k\epsilon^*_k\,.$$

Hence, the strict stationarity and ergodicity of $\{Y_k^*\}$ and thus $\{Y_k\}$ follow the same arguments with mild modifications.

For fTARX series (2.4), note that $X_{k,m}$s are assumed i.i.d. with $\mathrm{E}\|X_{k,m}\|^4 < \infty$ and $\mathrm{E}|X_{k,m}|^4 < \infty$ for function and scalar $X_{k,m}$ terms, respectively. Following the same derivations, there exists some measurable function h such that $Y_k^* = h(\epsilon_k^*, X_{k,m}, z_{k-d}, \epsilon_{k-1}^*, X_{k-1,m}, z_{k-d-1}, \ldots)$. Thus under the additional assumptions that $\max_i \sum_{m=1}^{p_X} \|\Phi_{i,m}\|_{\mathcal{L}} < 1$, the strict stationarity and ergodicity of $\{Y_k\}$ follow from the same arguments. □

Proof of Corollary 1

For two bounded linear operators Ψ_1 and Ψ_2 on H, we have

$$\|\Psi_1 \circ \Psi_2\|_{\mathcal{L}} \leq \|\Psi_1\|_{\mathcal{L}} \|\Psi_2\|_{\mathcal{L}} \,.$$

For two p-dimensional matrix of operators $\Psi_{i_2}^* = (\Psi_{l,k}^{*,i_2})_{1\leq l,k\leq p}$ and $\Psi_{i_1}^* = (\Psi_{l,k}^{*,i_1})_{1\leq l,k\leq p}$ on $x = (x_1, \ldots, x_p)^{\mathrm{T}} \in H_p$, their product norm is defined as $\|\Psi_{i_2}^* \circ \Psi_{i_1}^*\|_{\mathcal{L}_p} = \sup_{\|x\|_p=1} \|\Psi_{i_2}^*(\Psi_{i_1}^*)(x)\|_p$. By using the definition of p-norm of matrix operator, the triangular inequality and the Jensen's inequality, it

follows that

$$\begin{aligned}
\sup_{\|x\|_p=1} \|\Psi^*_{i_2}(\Psi^*_{i_1})(x)\|_p &= \sup_{\|x\|_p=1} \left\| \Psi^*_{i_2} \left(\sum_{j=1}^p \Psi^{*,i_1}_{1,j}(x_j), \cdots, \sum_{j=1}^p \Psi^{*,i_1}_{p,j}(x_j) \right)^{\mathrm{T}} \right\|_p \\
&= \sup_{\|x\|_p=1} \left\{ \sum_{i=1}^p \left\| \sum_{k=1}^p \sum_{j=1}^p \Psi^{*,i_2}_{i,k}(\Psi^{*,i_1}_{k,j})(x_j) \right\|^2 \right\}^{1/2} \\
&\leq \frac{p^{1/2}}{p} \sum_{i=1}^p \sup_{\|x\|_p=1} \left\| \sum_{k=1}^p \sum_{j=1}^p \Psi^{*,i_2}_{i,k}(\Psi^{*,i_1}_{k,j})(x_j) \right\| \\
&\leq p^{1/2} \max_i \sup_{\|x\|_p=1} \sum_{k=1}^p \sum_{j=1}^p \|\Psi^{*,i_2}_{i,k}\|_{\mathcal{L}} \|\Psi^{*,i_1}_{k,j}\|_{\mathcal{L}} \|x_j\| \\
&\leq \left(\max_l \max_i \sum_{j=1}^p \|\Psi^{*,i_l}_{i,j}\|_{\mathcal{L}} \right)^2 ,
\end{aligned}$$

for $l = 1, 2$, and the last quantity is bounded above by 1. Thus, by induction we have

$$\max_{i_1,\ldots,i_u} \left\| \Psi^*_{i_u} \circ \cdots \circ \Psi^*_{i_1} \right\|_{\mathcal{L}_p} < 1 ,$$

which completes the proof. □

Proof of Corollary 2

The proof of this corollary is straightforward: from p.34–35 in Bosq (2000) and p.148 of Riesz and Sz.-Nagy (1990), $\|\Psi\|_{\mathcal{S}} \geq \|\Psi\|_{\mathcal{L}}$ for all bounded linear operators Ψ on H whenever the norms are well defined. Hence, the condition of $\Psi_{i,j}$ in Corollary 2 implies Corollary 1. □

For the proof of Theorem S.1, we first introduce the following definitions and theorem:

Definition 2. *Irreducibility:* a Markov process $\{y_t\}$ on a measurable space $\{\Omega, \mathcal{B}\}$ with transition probability $P^n(y, A) = \mathrm{pr}(y_n \in A \mid y_0 = y)$ 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 3. *Geometric ergodicity:* a Markov process $\{y_t\}$ on $\{\Omega, \mathcal{B}\}$ is geometrically ergodic if there exists an invariant probability measure π on $\mathcal{B}$ such that for all $A \in \mathcal{B}$ and $y \in \Omega$, there exists some $\rho \in (0, 1)$ and $0 < M_y < \infty$ that

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

where $\|\cdot\|_{TV}$ is the total-variation norm. The geometric ergodicity implies that $\{y_t\}$ is β-mixing with an exponentially decaying rate, and thus ergodic. It also implies that $\{y_t\}$ has some unique stationary distribution π (see Bradley (2005); Saïdi and Zakoïan (2006)).

Definition 4. *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 Meyn and Tweedie (1993).

Proof of Theorem S.1:

To prove Theorem S.1, let $\boldsymbol{Y}_k = (\langle y_k, u_1\rangle, \ldots, \langle y_k, u_q\rangle)^{\mathrm{T}}$ is the projection on the true basis. First, we verify that $\{\boldsymbol{Y}_k\}$ is β-mixing with an exponentially decaying rate. By using the Karhunen-Loève expansion, it suffices to prove the geometric ergodicity of $\{\boldsymbol{Y}_k\}$. Denote

$$\boldsymbol{Y}_k = \sum_{i=1}^{r} \left[\boldsymbol{a}_i + \sum_{j=1}^{p_{Y,i}} \boldsymbol{\Psi}_{i,j} \boldsymbol{Y}_{k-j} + \boldsymbol{W}_{k,i} \right] I(z_{k-d} \in (\theta_{i-1}, \theta_i]), \tag{S.3}$$

where $\boldsymbol{W}_{k,i} = \sigma_i(\langle \epsilon_k, u_1\rangle, \ldots, \langle \epsilon_k, u_q\rangle)^{\mathrm{T}}$ has mean $\mathbf{0}$. Since every element of $\boldsymbol{W}_{k,i}$ has an absolutely continuous and almost everywhere positive density on $\mathbb{R}$, $\boldsymbol{W}_{k,i}$ has an almost everywhere continuous density on $\mathbb{R}^{q_Y}$. Next, define $\boldsymbol{Y}_k^* = \mathrm{vec}(\boldsymbol{Y}_k, \ldots, \boldsymbol{Y}_{k-p+1})$, $\boldsymbol{a}_i^* = \mathrm{vec}(\boldsymbol{a}_i, \mathbf{0}, \ldots, \mathbf{0})$, $\boldsymbol{W}_{k,i}^* = \mathrm{vec}(\boldsymbol{W}_{k,i}, \mathbf{0}, \ldots, \mathbf{0})$, and

$$\boldsymbol{\Psi}_i^* = \begin{pmatrix} \boldsymbol{\Psi}_{i,1} & \boldsymbol{\Psi}_{i,2} & \cdots & \boldsymbol{\Psi}_{i,p-1} & \boldsymbol{\Psi}_{i,p} \\ I_q & 0_q & \cdots & 0_q & 0_q \\ 0_q & I_q & \cdots & 0_q & 0_q \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0_q & 0_q & \cdots & I_q & 0_q \end{pmatrix},$$

where I_q and 0_q are $q \times q$ dimensional identity and zero matrices, respectively. For $p_{Y,i} < j \leq p$, $\boldsymbol{\Psi}_{i,j}$ is a zero matrix by the construction of (2.2)–(2.3) and (4.6). To verify the Markovian property, we may write (4.7) as

$$\boldsymbol{Y}_k^* = \sum_{i=1}^{r} \left[\boldsymbol{a}_i^* + \boldsymbol{\Psi}_i^* \boldsymbol{Y}_{k-1}^* + \boldsymbol{W}_{k,i}^* \right] I(z_{k-d} \in (\theta_{i-1}, \theta_i]) \,. \tag{S.4}$$

If $z_{k-d} = g(Y_{k-d})$, we have $z_{k-d} = g^*(\boldsymbol{Y}_{k-d})$ for some measurable function $g^* : \mathbb{R}^q \to \mathbb{R}$ such that $g^*(\boldsymbol{Y}_k) = g(Y_k)$. By (S.4), $\mathrm{pr}(\boldsymbol{Y}_k^* \mid \boldsymbol{Y}_{k-1}^*, \boldsymbol{Y}_{k-2}^*, \ldots) = \mathrm{pr}(\boldsymbol{Y}_k^* \mid \boldsymbol{Y}_{k-1}^*)$, and hence $\boldsymbol{Y}_k^*$ is a time-homogeneous Markov process. If $z_{k-d} = X_{k-d,m}$, as $X_{k-d,m}$s are i.i.d,

$$\mathrm{pr}(\boldsymbol{Y}_k^*, z_{k-d} \mid \boldsymbol{Y}_{k-1}^*, \boldsymbol{Y}_{k-2}^*, \ldots) = \mathrm{pr}(z_{k-d})\mathrm{pr}(\boldsymbol{Y}_k^* \mid \boldsymbol{Y}_{k-1}^*, \boldsymbol{Y}_{k-2}^*, \ldots) = \mathrm{pr}(z_{k-d})\mathrm{pr}(\boldsymbol{Y}_k^* \mid \boldsymbol{Y}_{k-1}^*) \,.$$

Integrating with respect to the density of z_{k-d}, we obtain the Markovian property of $\{\boldsymbol{Y}_k^*\}$.

In addition, then we consider the TVARX model (4.10) with $\boldsymbol{X}_{k,m} = (\langle X_{k,m}, u_1 \rangle, \ldots, \langle X_{k,m}, u_q \rangle)^{\mathrm{T}}$ and $\boldsymbol{X}_k = (\boldsymbol{X}_{k,1}, \ldots, \boldsymbol{X}_{k,p_X})^{\mathrm{T}}$. Thus, $\{\boldsymbol{X}_k\}$ is i.i.d. and independent of $\{\boldsymbol{Y}_s\}_{s \leq k-1}$ with $\mathrm{E}\|\boldsymbol{X}_{k,m}\|_2^4 < \infty$. Then, using the notation of $\boldsymbol{Y}_k^*$ defined previously, we denote $\mathcal{F}_{k-1}$ as the sigma-algebra generated by $\{\boldsymbol{Y}_{k-1}^*, \boldsymbol{Y}_{k-2}^*, \ldots\}$. By our assumptions on $X_{k,m}$, $\boldsymbol{X}_k$ is independent of $\mathcal{F}_{k-1}$. Therefore, if $z_{k-d} = g(Y_{k-d}) = g^*(\boldsymbol{Y}_{k-d})$, then

$$\mathrm{pr}(\boldsymbol{Y}_k^*, \boldsymbol{X}_k \mid \mathcal{F}_{k-1}) = \mathrm{pr}(\boldsymbol{X}_k)\mathrm{pr}(\boldsymbol{Y}_k^* \mid \boldsymbol{X}_k, \mathcal{F}_{k-1}) = \mathrm{pr}(\boldsymbol{X}_k)\mathrm{pr}(\boldsymbol{Y}_k^* \mid \boldsymbol{Y}_{k-1}^*, \boldsymbol{X}_k) \,,$$

and hence

$$\begin{aligned}\mathrm{pr}(\boldsymbol{Y}_k^* \mid \mathcal{F}_{k-1}) = \int_{\boldsymbol{X}_k} \mathrm{pr}(\boldsymbol{Y}_k^*, \boldsymbol{X}_k \mid \mathcal{F}_{k-1})\mathrm{pr}(d\boldsymbol{X}_k) &= \int_{\boldsymbol{X}_k} \mathrm{pr}(\boldsymbol{Y}_k^* \mid \boldsymbol{X}_k, \boldsymbol{Y}_{k-1}^*)\mathrm{pr}(d\boldsymbol{X}_k) \\ &= \mathrm{pr}(\boldsymbol{Y}_k^* \mid \boldsymbol{Y}_{k-1}^*)\,.\end{aligned}$$

Meanwhile, if $z_{k-d} = X_{k-d,m}$ for some scalar exogenous variable $X_{k-d,m}$, then $\mathrm{pr}(X_{k-d,m} \mid \mathcal{F}_{k-1}) = \mathrm{pr}(X_{k-d,m} \mid \boldsymbol{Y}_{k-1}^*)$. Analogously,

$$\begin{aligned}\mathrm{pr}(\boldsymbol{Y}_k^*, \boldsymbol{X}_k, X_{k-d,m} \mid \mathcal{F}_{k-1}) &= \mathrm{pr}(\boldsymbol{Y}_k^* \mid \boldsymbol{Y}_{k-1}^*, \boldsymbol{X}_k, X_{k-d,m})\mathrm{pr}(\boldsymbol{X}_k)\mathrm{pr}(X_{k-d,m} \mid \boldsymbol{Y}_{k-1}^*) \\ &= \mathrm{pr}(\boldsymbol{Y}_k^*, \boldsymbol{X}_k, X_{k-d,m} \mid \boldsymbol{Y}_{k-1}^*)\,,\end{aligned}$$

which implies $\mathrm{pr}(\boldsymbol{Y}_k^* \mid \mathcal{F}_{k-1}) = \mathrm{pr}(\boldsymbol{Y}_k^* \mid \boldsymbol{Y}_{k-1}^*)$. Therefore, $\{\boldsymbol{Y}_k^*\}$ is Markovian. Furthermore, as X_ks and thus $\boldsymbol{X}_k$s are i.i.d., the sufficient conditions for the β-mixing property of a functional TAR process also implies the β-mixing property for a functional TARX series. Hence, it suffices to study the functional TAR series in the proof of Theorem S.1 and Corollary S.1. By Propositions 1–2 in the following, we can verify the irreducibility, aperiodicity and the existence of some small sets with respect to $\{\boldsymbol{Y}_k^*\}$. Then by the Markovian property of $\{\boldsymbol{Y}_k^*\}$, we can verify the geometric ergodicity of $\{\boldsymbol{Y}_k^*\}$. By induction, (4.6), (S.10) and (S.11), we can further extend the result in Proposition 2 to $\|\Psi_{i_u}^* \circ \Psi_{i_{u-1}}^* \circ \cdots \circ \Psi_{i_1}^*\|_{\mathcal{L}_p} = \|\boldsymbol{\Psi}_{i_u}^* \boldsymbol{\Psi}_{i_{u-1}}^* \cdots \boldsymbol{\Psi}_{i_1}^*\|_2$. Thus, by (S.2), $\max_{i_1,\ldots,i_u} \|\boldsymbol{\Psi}_{i_u}^* \cdots \boldsymbol{\Psi}_{i_1}^*\|_2 = cb^u$ for all positive integers u with some $c > 0$ and $0 < b < 1$, and hence $\max_{i_1,\ldots,i_u} \|\boldsymbol{\Psi}_{i_u}^* \cdots \boldsymbol{\Psi}_{i_1}^*\|_2 < 1$ for

all sufficiently large u. First, we have

$$\begin{aligned}
\mathrm{E}(\|\boldsymbol{Y}_k^*\|_2 \mid \boldsymbol{Y}_{k-1}^*) &\leq \max_i \mathrm{E}(\left\|\boldsymbol{a}_i^* + \boldsymbol{\Psi}_i^*(\boldsymbol{Y}_{k-1}^*) + \boldsymbol{W}_{k,i}^*\right\|_2 \mid \boldsymbol{Y}_{k-1}^*) \\
&\leq \max_i [\mathrm{E}\|\boldsymbol{W}_{k,i}^*\|_2 + \|\boldsymbol{a}_i^*\|_2] + \max_i \|\boldsymbol{\Psi}_i^*\|_2 \|\boldsymbol{Y}_{k-1}^*\|_2 \\
&\leq H^* + \max_i \|\boldsymbol{\Psi}_i^*\|_2 \|\boldsymbol{Y}_{k-1}^*\|_2 \,, \qquad \text{(S.5)}
\end{aligned}$$

for some constant H^*. Then, applying the same arguments in (S.5),

$$\begin{aligned}
&\mathrm{E}(\|\boldsymbol{Y}_{k+1}^*\|_2 \mid \boldsymbol{Y}_{k-1}^*) \\
\leq\ & \max_{i_1,i_2} \mathrm{E}\left[\|\boldsymbol{W}_{k+1,i_2}^* + \boldsymbol{a}_{i_2}^* + \boldsymbol{\Psi}_{i_2}^*(\boldsymbol{W}_{k,i_1}^* + \boldsymbol{a}_{i_1}^*) + \boldsymbol{\Psi}_{i_2}^*(\boldsymbol{\Psi}_{i_1}^*(\boldsymbol{Y}_{k-1}^*))\|_2 \mid \boldsymbol{Y}_{k-1}^*\right] \\
\leq\ & \max_{i_2}[\mathrm{E}\|\boldsymbol{W}_{k+1,i_2}^*\|_2 + \|\boldsymbol{a}_{i_2}^*\|_2] + \max_{i_2} \|\boldsymbol{\Psi}_{i_2}^*\|_2 \max_{i_1}[\mathrm{E}\|\boldsymbol{W}_{k,i_1}^*\|_2 + \|\boldsymbol{a}_{i_1}^*\|_2] + \max_{i_1,i_2} \|\boldsymbol{\Psi}_{i_2}^*\boldsymbol{\Psi}_{i_1}^*\|_2 \|\boldsymbol{Y}_{k-1}^*\|_2 \\
\leq\ & (1 + \max_{i_2} \|\boldsymbol{\Psi}_{i_2}^*\|_2) H^* + \max_{i_1,i_2} \|\boldsymbol{\Psi}_{i_2}^*\boldsymbol{\Psi}_{i_1}^*\|_2 \|\boldsymbol{Y}_{k-1}^*\|_2 \,.
\end{aligned}$$

Therefore, by induction,

$$\begin{aligned}
&\mathrm{E}(\|\boldsymbol{Y}_{k+u}^*\|_2 \mid \boldsymbol{Y}_{k-1}^*) \\
\leq\ & H^*(1 + \max_{i_u} \|\boldsymbol{\Psi}_{i_u}^*\|_2 + \cdots + \max_{i_2,\ldots,i_u} \|\boldsymbol{\Psi}_{i_u}^* \cdots \boldsymbol{\Psi}_{i_2}^*\|_2) + \max_{i_1,\ldots,i_u} \|\boldsymbol{\Psi}_{i_u}^* \cdots \boldsymbol{\Psi}_{i_1}^*\|_2 \|\boldsymbol{Y}_{k-1}^*\|_2 \\
\leq\ & H^*(1 + c(1-b^u)/(1-b)) + \rho^* \|\boldsymbol{Y}_{k-1}^*\|_2 \,, \qquad \text{(S.6)}
\end{aligned}$$

where $\rho^* = cb^u < 1$. By choosing $(\rho^*)^{1/u} < \rho < 1$ for some ρ, $n(\boldsymbol{Y}_{k-1}^*) = u$, $\chi = [H^*(1 + c(1-b^u)/(1-b)) + 1]/\rho^u - 1$, $V(\boldsymbol{Y}_k^*) = 1 + \|\boldsymbol{Y}_k^*\|_2$ and a small set $C = \{\boldsymbol{Y}_{k-1}^* : \ \|\boldsymbol{Y}_{k-1}^*\|_2 \leq \chi/(\rho^u - \rho^*)\}$, (S.6) implies

$$\begin{aligned}
\mathrm{E}[V(\boldsymbol{Y}_{k+u}^*) \mid \boldsymbol{Y}_{k-1}^*] &\leq 1 + H^*(1 + c(1-b^u)/(1-b)) + \rho^* \|\boldsymbol{Y}_{k-1}^*\|_2 \\
&\leq \rho^u(1 + \|\boldsymbol{Y}_{k-1}^*\|_2 + \chi I(\boldsymbol{Y}_{k-1}^* \in C)) \,. \qquad \text{(S.7)}
\end{aligned}$$

Hence we deduce the geometric ergodicity of $\{\boldsymbol{Y}_k^*\}$ by Theorem 19.1.3 in Meyn and Tweedie (1993). Since $\{\boldsymbol{Y}_k^*\}$ is Markovian, we obtain that $\{\boldsymbol{Y}_k^*\}$ and thus $\{\boldsymbol{Y}_k\}$ is β-mixing with a geometrically decaying rate ρ. Thus combining with the expansion of $\{Y_k\}$ leads to that $\{Y_k\}$ is β-mixing. $\square$

Proposition 1. *For $\{\boldsymbol{Y}_k^*\}$ satisfying* (S.4)*, it is aperiodic and μ_{pq}-irreducible, where μ_{pq} is the Lebesgue measure on $\mathbb{R}^{pq}$. Furthermore, any non-null compact set is a small set.*

Proof of Proposition 1

We apply the methodologies in Chan and Tong (1985) in our proof. Without loss of generality, we assume $\boldsymbol{a}_i^* = \boldsymbol{a}_i$ and $\boldsymbol{W}_{k,i}^* = \boldsymbol{W}_k^*$ for all i. Denote $P(\boldsymbol{X}, A) = \mathrm{pr}(\boldsymbol{Y}_{k+1}^* \in A \mid \boldsymbol{Y}_k^* = \boldsymbol{X})$ and $P^m(\boldsymbol{X}, A) = \mathrm{pr}(\boldsymbol{Y}_{k+m}^* \in A \mid \boldsymbol{Y}_k^* = \boldsymbol{X})$. Let the density of $\boldsymbol{W}_k = (\langle \epsilon_k, u_1\rangle, \ldots, \langle \epsilon_k, u_q\rangle)^{\mathrm{T}}$ be $f(\cdot)$ in $\mathbb{R}^q$, where $f(\cdot)$ is absolutely continuous and almost everywhere positive.

For $z_{k-d} = g^*(\boldsymbol{Y}_{k-d})$, the regime belonging of $\boldsymbol{Y}_k$ is specified given $\boldsymbol{Y}_{k-1}^*$, and hence $\mathrm{E}(\boldsymbol{Y}_k^* \mid \boldsymbol{Y}_{k-1}^* = \boldsymbol{x})$ is deterministic from (S.4). Then, for any compact set A with non-zero μ_{pq}-measure on $\mathbb{R}^{pq}$, there exists a non-null compact set

$$\mathcal{E}_k(\boldsymbol{x}, A) = \{\boldsymbol{W}_k : \ \mathrm{E}(\boldsymbol{Y}_k^* \mid \boldsymbol{Y}_{k-1}^* = \boldsymbol{x}) + \mathrm{vec}(\boldsymbol{W}_k, \boldsymbol{0}, \ldots, \boldsymbol{0}) \in A\},$$

and an injection exists between $\mathcal{E}_k(\boldsymbol{x}, A)$ and A. Hence, $\mu_{pq}(\mathcal{E}_k(\boldsymbol{x}, A)) > 0$,

and we have

$$\begin{aligned} P(\boldsymbol{x}, A) &= \mathrm{pr}(\boldsymbol{Y}_k^* \in A \mid \boldsymbol{Y}_{k-1}^* = \boldsymbol{x}) \\ &= \int_{\mathcal{E}_k(\boldsymbol{x},A)} f(\boldsymbol{w}) d\boldsymbol{w} \geq \inf_{\boldsymbol{w} \in \mathcal{E}_k(x,A)} f(\boldsymbol{w}) \mu_{pq}(\mathcal{E}_k(\boldsymbol{x}, A)) > 0 \,. \end{aligned} \tag{S.8}$$

Next, consider $P^2(\boldsymbol{x}, A)$. Since there exists some compact set A_2 with non-zero μ_{pq}-measure in $\mathbb{R}^{pq}$, by defining $A' = A_1 \times A \subset \mathbb{R}^{pq} \times \mathbb{R}^{pq}$, we construct a non-null compact set

$$\begin{aligned} \mathcal{E}_k^2(\boldsymbol{x}, A') &= \{\boldsymbol{W}_k, \boldsymbol{W}_{k+1} : \mathrm{E}(\boldsymbol{Y}_k^* \mid \boldsymbol{Y}_{k-1}^* = \boldsymbol{x}) + \mathrm{vec}(\boldsymbol{W}_k, \boldsymbol{0}, \ldots, \boldsymbol{0}) \in A_1 \,, \\ &\quad \mathrm{E}[\boldsymbol{Y}_{k+1}^* \mid \boldsymbol{Y}_k^* = \boldsymbol{x} + \mathrm{vec}(\boldsymbol{W}_k, \boldsymbol{0}, \ldots, \boldsymbol{0})] + \mathrm{vec}(\boldsymbol{W}_{k+1}, \boldsymbol{0}, \ldots, \boldsymbol{0}) \in A\} \,. \end{aligned}$$

By Fubini's Theorem,

$$\begin{aligned} P^2(\boldsymbol{x}, A) &\geq \mathrm{pr}(\boldsymbol{Y}_{k+1}^* \in A, \boldsymbol{Y}_k^* \in A_1 \mid \boldsymbol{Y}_{k-1}^* = \boldsymbol{x}) \\ &= \int_{\boldsymbol{x}_2 \in A_1} P(\boldsymbol{x}_2, A) P(\boldsymbol{x}, d\boldsymbol{x}_2) = \iint_{\mathcal{E}_k^2(\boldsymbol{x}, A')} f(\boldsymbol{w}_1, \boldsymbol{w}_2) d\boldsymbol{w}_1 d\boldsymbol{w}_2 > 0 \,. \end{aligned} \tag{S.9}$$

Thus by induction, $P^m(\boldsymbol{x}, A) > 0$ for all positive integer m and $\boldsymbol{x} \in \mathbb{R}^{pq}$, which proves the μ_{pq}-irreducibility and aperiodicity. For a μ_{pq}-non-null compact set $C \subset \mathbb{R}^{pq}$, as $f(\cdot)$ is absolutely continuous and almost everywhere positive on $\mathbb{R}^{pq}$, $\inf_{\boldsymbol{x} \in C} P(\boldsymbol{x}, A) > 0$, which implies that the set C is small by taking $v_m(A) = \inf_{\boldsymbol{x} \in C} P(\boldsymbol{x}, A)$. As z_{k-d} is independent of $\{\boldsymbol{Y}_k\}$, since

$$P(\boldsymbol{x}, A) = \sum_{i=1}^{r} \mathrm{pr}(\boldsymbol{Y}_k^* \in A \mid \boldsymbol{Y}_{k-1}^* = \boldsymbol{x}, z_{k-d} \in (\theta_{i-1}, \theta_i]) \mathrm{pr}(z_{k-d} \in (\theta_{i-1}, \theta_i]) \,,$$

analogous results of (S.8) and (S.9) can be derived by the same arguments, which verify Proposition 1. □

Next, we verify the geometric ergodicity of $\{\boldsymbol{Y}_k^*\}$. In the following, with a little abuse of notation, we denote $\|A\|_2$ as the entry-wise 2-norm of a matrix A, i.e., the Frobenius norm of a matrix A. Define the function $V(\boldsymbol{Y}_k^*) = 1 + \|\boldsymbol{Y}_k^*\|_2$. We have the following proposition:

Proposition 2. *With the true basis $\{u_l\}$ for $l = 1, \ldots, q_Y$, we have $\|\Psi_{i,j}\|_{\mathcal{L}} = \|\boldsymbol{\Psi}_{i,j}\|_2$ for all i, j, where*

$$\boldsymbol{\Psi}_{i,j} = \begin{pmatrix} \langle \Psi_{i,j}(u_1), u_1 \rangle & \cdots & \langle \Psi_{i,j}(u_1), u_q \rangle \\ \vdots & \ddots & \vdots \\ \langle \Psi_{i,j}(u_q), u_1 \rangle & \cdots & \langle \Psi_{i,j}(u_q), u_q \rangle \end{pmatrix} \tag{S.10}$$

is a $q \times q$ matrix.

Proof of Proposition 2

From (4.6)–(4.7), it follows that

$$\begin{aligned}
\Psi_{i,j}(u_1) &= (\langle u_1, u_1\rangle, \ldots, \langle u_1, u_q\rangle)\, \boldsymbol{\Psi}_{i,j}\, (u_1, \ldots, u_q)^{\mathrm{T}} \\
&= (1, 0, \ldots, 0) \begin{pmatrix} \langle \Psi_{i,j}(u_1), u_1\rangle & \cdots & \langle \Psi_{i,j}(u_1), u_q\rangle \\ \vdots & \ddots & \vdots \\ \langle \Psi_{i,j}(u_q), u_1\rangle & \cdots & \langle \Psi_{i,j}(u_q), u_q\rangle \end{pmatrix} (u_1, \ldots, u_q)^{\mathrm{T}} \\
&= (\langle \Psi_{i,j}(u_1), u_1\rangle, \cdots, \langle \Psi_{i,j}(u_1), u_q\rangle)\, (u_1, \ldots, u_q)^{\mathrm{T}} \\
&= \sum_{l=1}^{q} \langle \Psi_{i,j}(u_1), u_l\rangle u_l\,,
\end{aligned}$$

and so on for $u_2, \ldots, u_q$. Hence, by writing $x = \sum_{l=1}^{q} a_l u_l$,

$$\|\Psi_{i,j}\|_{\mathcal{L}} = \sup_{\|x\|\le 1} \|\Psi_{i,j}(x)\| = \sup_{\sum_{l=1}^{q} a_l^2 \le 1} \left[\sum_{l_1=1}^{q} (\sum_{l_2=1}^{q} a_{l_2} \langle \Psi_{i,j}(u_{l_2}), u_{l_1}\rangle)^2 \right]^{1/2} = \|\boldsymbol{\Psi}_{i,j}\|_2\,. \tag{S.11}$$

□

Proof of Corollary S.1

We prove the geometric ergodicity of $\{\boldsymbol{Y}_k\}$ by the same methodology in the proof of Theorem S.1. First, we illustrate the proof with respect to the case with $z_{k-d} = g(Y_{k-d})$. By Proposition 2, we have $\rho_1 = \max_{i=1,r} \sum_j \|\boldsymbol{\Psi}_{i,j}\|_2 \in (0,1)$, $\rho_2 = \max_{i=2,\ldots,r-1} \sum_j \|\boldsymbol{\Psi}_{i,j}\|_2$. Also, denote

$$\pi^*(\boldsymbol{Y}^*) = \sup_{\boldsymbol{Y}^*} \mathrm{pr}(g^*(\boldsymbol{Y}_{k-d}) \in (\theta_1, \theta_r] \mid \boldsymbol{Y}^*_{k-d-1} = \boldsymbol{Y}^*) = \pi(Y^*)\,,$$

for some $\pi^* : \mathbb{R}^q \to \mathbb{R}$. Therefore,

$$\gamma = \rho_1(1 - \pi^*(\boldsymbol{Y}^*)) + \rho_2 \pi^*(\boldsymbol{Y}^*) < 1\,.$$

Without loss of generality, we assume $d = 1$ and $\rho_2 > 1$. Thus, $\rho_1 < \rho_2$. By (S.3),

$$\begin{aligned}
\mathrm{E}(\|\boldsymbol{Y}_k\|_2 \mid \boldsymbol{Y}^*_{k-1}) &= \mathrm{E}\left(\sum_{i=1}^{r} \|\boldsymbol{a}_i + \sum_{j=1}^{p_{Y,i}} \boldsymbol{\Psi}_{i,j}\boldsymbol{Y}_{k-j} + \boldsymbol{W}_{k,i}\|_2 I(g^*(\boldsymbol{Y}_k) \in (\theta_{i-1}, \theta_i]) \mid \boldsymbol{Y}^*_{k-1}\right) \\
&\leq \max_i\{\mathrm{E}\|\boldsymbol{W}_{k,i}\|_2 + \|\boldsymbol{a}_i\|_2\} + \max_i \sum_{j=1}^{p_{Y,i}} \|\boldsymbol{\Psi}_{i,j}\|_2 \|\boldsymbol{Y}_{k-j}\|_2 \\
&\leq H^* + \rho_2 \max_{j=1,\ldots,p} \|\boldsymbol{Y}_{k-j}\|_2 \,,
\end{aligned}$$

for some constant H^*. Similarly, we have

$$\begin{aligned}
&\mathrm{E}(\|\boldsymbol{Y}_{k+1}\|_2 \mid \boldsymbol{Y}^*_{k-1}) \\
\leq\ & \max_i[\mathrm{E}\|\boldsymbol{W}_{k+1,i}\|_2 + \|\boldsymbol{a}_i\|_2] + \mathrm{E}\left\|\sum_{i=1}^{r}\sum_{j=1}^{p_{Y,i}} \boldsymbol{\Psi}_{i,j}\boldsymbol{Y}_{k-j+1} I(g^*(\boldsymbol{Y}_k) \in (\theta_{i-1}, \theta_i]) \mid \boldsymbol{Y}^*_{k-1}\right\|_2 \\
\leq\ & H^* + \mathrm{E}\left[\sum_{i=1}^{r}\sum_{j=1}^{p_{Y,i}} \|\boldsymbol{\Psi}_{i,j}\|_2 I(g^*(\boldsymbol{Y}_k) \in (\theta_{i-1}, \theta_i]) \mid \boldsymbol{Y}^*_{k-1}\right] \max\{\max_{j=1,\ldots,p} \|\boldsymbol{Y}_{k-j}\|_2, \mathrm{E}(\|\boldsymbol{Y}_k\|_2 \mid \boldsymbol{Y}^*_{k-1})\} \\
\leq\ & H^* + \gamma(H^* + \rho_2 \max_{j=1,\ldots,p} \|\boldsymbol{Y}_{k-j}\|_2) \,. \qquad \text{(S.12)}
\end{aligned}$$

By induction, for all $m_1 \geq 0$, $0 \leq m_2 \leq p-1$,

$$\mathrm{E}(\|\boldsymbol{Y}_{k+m_1p+m_2+1}\|_2 \mid \boldsymbol{Y}^*_{k-1}) \leq H^*(1+\gamma+\cdots+\gamma^{m_1p+m_2+1}) + \gamma^{m_1+1}\rho_2 \max_{j=1,\ldots,p} \|\boldsymbol{Y}_{k-j}\|_2 \,. \qquad \text{(S.13)}$$

Since $\gamma < 1$ and hence $\gamma^{m_1+1}\rho_2 < 1$ for a sufficiently large m_1, similar to (S.6) and (S.7), we choose $V(\boldsymbol{Y}^*_k) = 1 + \|\boldsymbol{Y}^*_k\|_2$ and verify the geometric ergodicity of $\{\boldsymbol{Y}^*_k\}$ by Theorem 19.1.3 in Meyn and Tweedie (1993).

For the general case with $d \geq 1$, for all $m_1 \geq 0$, $0 \leq m_2 \leq p-1$,

$$\mathrm{E}(\|\boldsymbol{Y}_{k+m_1p+m_2+d}\|_2 \mid \boldsymbol{Y}^*_{k-1}) \leq H^*(1+\gamma+\cdots+\gamma^{m_1p+m_2+1})(1+\rho_2+\cdots+\rho_2^{d-1})$$
$$+\gamma^{m_1+1}\rho_2^d \max_{j=1,\ldots,p} \|\boldsymbol{Y}_{k-j}\|_2 \,. \tag{S.14}$$

As $\gamma^{m_1+1}\rho_2^d < 1$ for a sufficiently large integer m_1, the geometric ergodicity of $\{\boldsymbol{Y}^*_k\}$ can also be verified similarly.

For the case $z_{k-d} = X_{k-d,m}$, we choose $V(\boldsymbol{Y}^*_k) = 1 + \|\boldsymbol{Y}^*_k\|_2$. By Proposition 2,

$$\gamma = \sum_{i=1}^{r}\sum_{j=1}^{p_{Y,i}} \|\boldsymbol{\Psi}_{i,j}\|_2 \,\mathrm{pr}(z_{k-d} \in (\theta_{i-1}, \theta_i]) < 1 \,.$$

Hence,

$$\begin{aligned}
& \mathrm{E}(\|\boldsymbol{Y}_k\|_2 \mid \boldsymbol{Y}^*_{k-1}) \\
\leq\ & \max_i\{\mathrm{E}\|\boldsymbol{W}_{k,i}\|_2 + \|\boldsymbol{a}_i\|_2\} + \mathrm{E}\left(\sum_{i=1}^{r}\sum_{j=1}^{p_{Y,i}} \|\boldsymbol{\Psi}_{i,j}\|_2\|\boldsymbol{Y}_{k-j}\|_2 I(z_{k-d} \in (\theta_{i-1},\theta_i]) \mid \boldsymbol{Y}^*_{k-1}\right) \\
\leq\ & H^* + \gamma \max_{j=1,\ldots,p} \|\boldsymbol{Y}_{k-j}\|_2 \,.
\end{aligned}$$

By similar arguments in deriving (S.12) and (S.13),

$$\mathrm{E}(\|\boldsymbol{Y}_{k+m}\|_2 \mid \boldsymbol{Y}^*_{k-1}) \ \leq\ (1+\gamma+\cdots+\gamma^m)H^* + \gamma \max_{j=1,\ldots,p} \|\boldsymbol{Y}_{k-j}\|_2 \,,$$

for any positive integer $m \leq p-1$. Therefore, the geometric ergodicity of $\{\boldsymbol{Y}^*_k\}$ is derived by employing the similar arguments as in (S.14). Lastly, by the geometric ergodicity of $\{\boldsymbol{Y}^*_k\}$, $\{\boldsymbol{Y}_k\}$ and hence $\{Y_k\}$ is β-mixing with an exponentially decaying rate by employing analogous arguments in proving Theorem S.1.

B: Proof of asymptotic theory for statistical inferences

First, the following Proposition is established for proving Theorem 2:

Proposition 3. *Denote* $C(\cdot) = \mathrm{E}[\langle Y_k - \mu, \cdot\rangle(Y_k - \mu)]$ *as the covariance operator in the true model,* $\hat{\mu}_n = \sum_{k=1}^{n} Y_k/n$ *as the estimated mean, and* $\hat{C}_n(\cdot) = \sum_{k=1}^{n}[\langle Y_k - \hat{\mu}_n, \cdot\rangle(Y_k - \hat{\mu}_n)]/n$ *as estimated covariance operator. Under Assumption 1,* $\mathrm{E}\|\hat{\mu}_n - \mu\|^2 = O(n^{-1})$ *and* $\mathrm{E}\|\hat{C}_n - C\|_{\mathcal{L}}^2 = O(n^{-1})$.

Proof of Proposition 3

As X_ks are i.i.d and independent of $\{Y_s\}_{s\leq k-1}$, it suffices to illustrate the proof for functional TAR model. In addition, without loss of generality, we assume $a_i = 0$ and $\sigma_i = 1$ in all regimes (so that $\mu = 0$), and the series start at $k = 0$ with $Y_0 = Y$ for some $\|Y\| < \infty$.

We first verify that $Y_k \in L_H^2$ and then verify $Y_k \in L_H^4$. Suppose that the series start with some $\boldsymbol{Y}_0^* \in L_{H_p}^4$ almost surely. Using the definition of $\tilde{\Psi}_k^*$ as $\tilde{\Psi}_k^* = \sum_{i=1}^{r} \Psi_i^* I(z_{k-d} \in (\theta_{i-1}, \theta_i])$, by Jensen's inequality and the assumption $\max_i \mathrm{E}(\|\tilde{\Psi}_{k+u-1}^* \circ \cdots \circ \tilde{\Psi}_k^*\|_{\mathcal{L}_p}^4 \mid z_{k-d-1} \in (\theta_{i-1}, \theta_i]) < 1$, we have

$$\max_{i=1,\ldots,r} \mathrm{E}\left(\|\tilde{\Psi}_{k+u-1}^* \circ \cdots \circ \tilde{\Psi}_k^*\|_{\mathcal{L}_p}^2 \mid z_{k-d-1} \in (\theta_{i-1}, \theta_i]\right) < 1\,.$$

Then, we write

$$Y_{k+l}^* = \sum_{m=1}^{l-1} (\tilde{\Psi}_{k+l-1}^* \circ \cdots \circ \tilde{\Psi}_{k+l-m}^*)(\epsilon_{k+l-m}^*) + \epsilon_{k+l}^* + (\tilde{\Psi}_{k+l-1}^* \circ \cdots \circ \tilde{\Psi}_k^*)(Y_k^*)\,, \tag{S.15}$$

for all $l \in \mathbb{N}$. From (S.15) for each $i = 1, \ldots, r$,

$$\begin{aligned}
\mathrm{E}\|Y^*_{k+l}\|_p^2 &= \mathrm{E}[\mathrm{E}(\|Y^*_{k+l}\|_p^2 \mid z_{k-d-1} \in (\theta_{i-1}, \theta_i])] \\
&\leq \mathrm{E}\left[\sum_{m=1}^{l-1} \mathrm{E}\left(\|\tilde{\Psi}^*_{k+l-1} \circ \cdots \circ \tilde{\Psi}^*_{k+l-m}\|^2_{\mathcal{L}_p} \mid z_{k-d-1} \in (\theta_{i-1}, \theta_i]\right) \mathrm{E}\|\epsilon^*_{k+l-m}\|_p^2 + \mathrm{E}\|\epsilon^*_{k+l}\|_p^2\right. \\
&\quad \left. + \mathrm{E}\left(\|\tilde{\Psi}^*_{k+l-1} \circ \cdots \circ \tilde{\Psi}^*_{k}\|^2_{\mathcal{L}_p} \|Y^*_k\|_p^2 \mid z_{k-d-1} \in (\theta_{i-1}, \theta_i]\right)\right] \\
&\leq \mathrm{E}\left[\sum_{m=1}^{l-1} \mathrm{E}\left(\|\tilde{\Psi}^*_{k+l-1} \circ \cdots \circ \tilde{\Psi}^*_{k+l-m}\|^2_{\mathcal{L}_p} \mid z_{k-d-1} \in (\theta_{i-1}, \theta_i]\right) + 1\right] \mathrm{E}\|\epsilon^*_k\|_p^2 \\
&\quad + \max_{i=1,\ldots,r} \mathrm{E}\left(\|\tilde{\Psi}^*_{k+l-1} \circ \cdots \circ \tilde{\Psi}^*_{k}\|^2_{\mathcal{L}_p} \mid z_{k-d-1} \in (\theta_{i-1}, \theta_i]\right) \mathrm{E}\|Y^*_k\|_p^2 . \qquad \text{(S.16)}
\end{aligned}$$

By (S.16), it follows that

$$\mathrm{E}\|Y^*_{k+u}\|_p^2 \leq H^* + b\mathrm{E}\|Y^*_k\|_p^2 , \qquad \text{(S.17)}$$

for some constant H^* and $b \in (0, 1)$. By induction on (S.17),

$$\mathrm{E}\|Y^*_{k+l}\|_p^2 \leq (1 + b + \cdots + b^{a_1 - 1})H^* + b^{a_1}\mathrm{E}\|Y^*_{k+a_2}\|_p^2 , \qquad \text{(S.18)}$$

for all $l = a_1 u + a_2$ with integer $a_1 \geq 1$ and $0 \leq a_2 \leq u - 1$. Meanwhile, by (S.15), $\mathrm{E}\|Y^*_{k+a_2}\|_p^2 < \infty$ for $0 \leq a_2 \leq u - 1$ given that $Y^*_k \in L^2_{H_p}$. By choosing $k = 0$ and $a_1 \to \infty$, we have $\mathrm{E}\|Y^*_l\|_p^2 < \infty$ for all l, and hence $Y_k \in L^2_H$.

Moreover, analogous to (S.16) for each $i = 1, \ldots, r$,

$$
\begin{aligned}
& \mathrm{E}\|Y^*_{k+l}\|_p^4 \\
= \;& \mathrm{E}[\mathrm{E}(\|Y^*_{k+l}\|_p^4 \mid z_{k-d-1} \in (\theta_{i-1}, \theta_i])] \\
\leq \;& \mathrm{E}\left[\sum_{m=1}^{l-1} \mathrm{E}\left(\|\tilde{\Psi}^*_{k+l-1} \circ \cdots \circ \tilde{\Psi}^*_{k+l-m}\|^4_{\mathcal{L}_p} \mid z_{k-d-1} \in (\theta_{i-1}, \theta_i]\right) + 1\right] \mathrm{E}\|\epsilon^*_k\|_p^4 \\
& +2\mathrm{E}\left[\mathrm{E}\left(\|(\tilde{\Psi}^*_{k+l-1} \circ \cdots \circ \tilde{\Psi}^*_k)\|^2_{\mathcal{L}_p}\|Y^*_k\|_p^2 \mid z_{k-d-1} \in (\theta_{i-1}, \theta_i]\right)\right] \mathrm{E}\|\epsilon^*_k\|_p^2 \\
& +2\mathrm{E}\left[\sum_{m=1}^{l-1} \mathrm{E}\left(\|\tilde{\Psi}^*_{k+l-1} \circ \cdots \circ \tilde{\Psi}^*_{k+l-m}\|^2_{\mathcal{L}_p} \mid z_{k-d-1} \in (\theta_{i-1}, \theta_i]\right)\right] \left(\mathrm{E}\|\epsilon^*_k\|_p^2\right)^2 \\
& +2\mathrm{E}\left[\sum_{m=1}^{l-1} \mathrm{E}\left(\|\tilde{\Psi}^*_{k+l-1} \circ \cdots \circ \tilde{\Psi}^*_{k+l-m}\|^2_{\mathcal{L}_p}\|\tilde{\Psi}^*_{k+l-1} \circ \cdots \circ \tilde{\Psi}^*_k\|^2_{\mathcal{L}_p}\|Y^*_k\|_p^2 \mid z_{k-d-1} \in (\theta_{i-1}, \theta_i]\right)\right] \mathrm{E}\|\epsilon^*_k\|_p^2 \\
& +2\mathrm{E}\left[\sum_{m_1=2}^{l-1}\sum_{m_2=1}^{m_1-1} \mathrm{E}\left(\|\tilde{\Psi}^*_{k+l-1} \circ \cdots \circ \tilde{\Psi}^*_{k+l-m_2}\|^2_{\mathcal{L}_p}\|\tilde{\Psi}^*_{k+l-1} \circ \cdots \circ \tilde{\Psi}^*_{k+l-m_1}\|^2_{\mathcal{L}_p} \mid z_{k-d-1} \in (\theta_{i-1}, \theta_i]\right)\right] \\
& \times \left(\mathrm{E}\|\epsilon^*_k\|_p^2\right)^2 + \max_{i=1,\ldots,r} \mathrm{E}\left(\|\tilde{\Psi}^*_{k+l-1} \circ \cdots \circ \tilde{\Psi}^*_{k+1}\|^4_{\mathcal{L}_p} \mid z_{k-d-1} \in (\theta_{i-1}, \theta_i]\right) \mathrm{E}\|Y^*_k\|_p^4 .
\end{aligned}
$$

By the assumption $\max_i \mathrm{E}\left(\|\tilde{\Psi}^*_{k+u-1} \circ \cdots \circ \tilde{\Psi}^*_k\|^4_{\mathcal{L}_p} \mid z_{k-d-1} \in (\theta_{i-1}, \theta_i]\right) < 1$, $\mathrm{E}\|Y^*_{k+u}\|_p^4 < \tilde{H}^* + \tilde{b}\mathrm{E}\|Y^*_k\|_p^4$ with some constant $\tilde{H}^*$ and $\tilde{b} \in (0, 1)$. Then, employing the same arguments on (S.18), for all $l = a_1 u + a_2$ with integer $a_1 \geq 1$ and $0 \leq a_2 \leq u - 1$,

$$
\mathrm{E}\|Y^*_{k+l}\|_p^4 \leq (1 + b + \ldots + b^{a_1-1})\tilde{H}^* + \tilde{b}^{a_1}\mathrm{E}\|Y^*_{k+a_2}\|_p^4 .
$$

Hence, by (S.15), $\mathrm{E}\|Y^*_{k+a_2}\|_p^4 < \infty$ for $0 \leq a_2 \leq u - 1$ given that $Y^*_k \in L^4_{H_p}$, and hence we have $\mathrm{E}\|Y^*_l\|_p^4 < \infty$ for all l. Analogously, we have $Y_k \in L^4_H$.

Next, if $z_{k-d} = g(Y_{k-d})$, by (S.1), we may write $Y_k = g_Y(\epsilon_k, \epsilon_{k-1}, \ldots)$ for some measurable function $g_Y : H_\infty \to H$. As $Y_k \in L^2_H$ and $Y_k \in L^4_H$,

by Hörmann and Kokoszka (2010), $\{Y_k\}$ is L^2 and L^4-m-approximable. If $z_{k-d} = X_{k-d,m}$, as $X_{k-d,m}$s are i.i.d., we have

$$z_k = X_{k-d,m} = e_k + \eta\,, \tag{S.19}$$

where $\{e_k\}$ is some sequence of i.i.d. white noises with $\mathrm{E}(e_k^4) < \infty$, and η is some constant. Therefore, denote $e_k^* = (\epsilon_k, e_k)^{\mathrm{T}}$, we have $Y_k = g_Y(e_k^*, e_{k-1}^*, \ldots)$ for some measurable function $g_Y : H_\infty \times \mathbb{R}^\infty \to H$. Therefore, $\{Y_k\}$ is L^2 and L^4-m-approximable. For a functional TARX process, as $X_{k,m}$s are assumed i.i.d. with finite second moment, by a similar decomposition of (S.19) on all $X_{k,m}$, we can also verify that $\{Y_k\}$ is L^2 and L^4-m-approximable.

By Theorem 3.1 in Hörmann and Kokoszka (2010) and Theorems 4.1 and 4.2 in Hörmann and Kokoszka (2012), as Y_k is L^2 and L^4-m-approximable, we have $\mathrm{E}\|\hat{\mu}_n - \mu\|^2 = O(n^{-1})$, and $\mathrm{E}\|\hat{C}_n - C\|_{\mathcal{S}}^2 = O(n^{-1})$ which implies that $\mathrm{E}\|\hat{C}_n - C\|_{\mathcal{L}}^2 = O(n^{-1})$. □

Denote λ_l and ν_l as the true eigenvalue and eigenfunction of the covariance operator C, respectively, i.e., $C\nu_l = \lambda_l \nu_l$. Correspondingly, $\hat{\lambda}_l$ and $\hat{\nu}_l$ are the empirical eigenvalue and eigenfunction for $\hat{C}$, respectively, i.e., $\hat{C}_n \hat{\nu}_l = \hat{\lambda}_l \hat{\nu}_l$. Based on Proposition 3, we have the following Proposition for $\hat{\lambda}_l$ and $\hat{\nu}_l$.

Proposition 4. *By Assumption 1,* $\mathrm{E}\|\hat{c}_l \hat{\nu}_l - \nu_l\|^2 = O(n^{-1})$ *and* $\mathrm{E}|\hat{\lambda}_l - \lambda_l|^2 =$

$O(n^{-1})$, *where* $\hat{c}_l = \text{sign}\langle \hat{\nu}_l, \nu_l \rangle$.

Proof of Proposition 4

By Proposition 3, $\mathrm{E}\|\hat{\mu}_n - \mu\|^2 = O(n^{-1})$ and $\mathrm{E}\|\hat{C}_n - C\|_{\mathcal{L}}^2 = O(n^{-1})$. By Lemma 2.1 and 2.2 in Hörmann and Kokoszka (2012),

$$|\hat{\lambda}_l - \lambda_l| \leq \|\hat{C}_n - C\|_{\mathcal{L}}\,, \quad \text{and} \quad \|\hat{c}_l\hat{\nu}_l - \nu_l\| \leq \frac{2\sqrt{2}}{\alpha_l^*}\|\hat{C}_n - C\|_{\mathcal{L}}\,,$$

where $\alpha_1^* = \lambda_1 - \lambda_2$ and $\alpha_l^* = \min\{\lambda_{l-1} - \lambda_l, \lambda_l - \lambda_{l+1}\}$ for $l = 2, 3, \ldots$. Thus, $\mathrm{E}\|\hat{c}_l\hat{\nu}_l - \nu_l\|^2 = O(n^{-1})$ and $\mathrm{E}|\hat{\lambda}_l - \lambda_l|^2 = O(n^{-1})$. □

We illustrate the proofs of Theorems 2–5 without assuming known basis for the functional TAR process, while the results for the functional TARX process can be obtained by more tedious calculations. When the basis is unknown, one can use the data-driven FPCA to obtain the basis functions, and replace empirical FPC scores $\boldsymbol{Y}_k^e = (\langle Y_k, \hat{\nu}_1\rangle, \ldots, \langle Y_k, \hat{\nu}_q\rangle)^{\mathrm{T}}$ by the theoretical ones $\boldsymbol{Y}_k = (\langle Y_k, \nu_1\rangle, \ldots, \langle Y_k, \nu_q\rangle)^{\mathrm{T}}$ in the proofs. By Proposition 4, $\mathrm{E}|\langle Y_k, \hat{c}_l\hat{\nu}_l\rangle - \langle Y_k, \nu_l\rangle|^2 = O(n^{-1})$ for all l, which leads to

$$\|\boldsymbol{Y}_k - \boldsymbol{Y}_k^e\|_2^2 = O_p(n^{-1})\,,$$

that is, the empirical FPC scores converge to the theoretical ones with a root-n rate. Define

$$L_{n,i}(\boldsymbol{\Psi}_i, \Sigma_{W,i}, p_{Y,i}, \theta_l, \theta_u, q, d) = \frac{1}{n}\sum_{k=1}^{n} l(\boldsymbol{\Psi}_i^q, \Sigma_{W,i}^{q_{\max}}; \boldsymbol{Y}_k, \ldots, \boldsymbol{Y}_{k-p_{Y,i}}) I(z_{k-d} \in (\theta_l, \theta_u])\,, \tag{S.20}$$

and

$$L^e_{n,i}(\boldsymbol{\Psi}_i, \Sigma_{W,i}, p_{Y,i}, \theta_l, \theta_u, q, d) = \frac{1}{n}\sum_{k=1}^{n} l(\boldsymbol{\Psi}^q_i, \Sigma^{q_{\max}}_{W,i}; \boldsymbol{Y}^e_k, \ldots, \boldsymbol{Y}^e_{k-p_{Y,i}}) I(z_{k-d} \in (\theta_l, \theta_u]),$$

where the superscript q and $q_{\max}$ are suppressed from the notations of parameters and covariance respectively in $L_{n,i}$ and $L^e_{n,i}$ to simplify notations, and a separate parameter q is used to reflect the dimension of $\boldsymbol{\Psi}_i$. From (S.20), we have $L^e_{n,i}(\boldsymbol{\Psi}_i, \Sigma_{W,i}, p_{Y,i}, \theta_l, \theta_u, q, d) = L_{n,i}(\boldsymbol{\Psi}_i, \Sigma_{W,i}, p_{Y,i}, \theta_l, \theta_u, q, d) + o_p(1)$ uniformly for any model parameters $\boldsymbol{\Psi}_i$ and $\Sigma_{W,i}$, the thresholds $\boldsymbol{\theta}$ and the model orders $p_{Y,i}$ and d. Hence, replacing the empirical ones by the theoretical counterparts has a negligible effect of the magnitude of the likelihood function. For notational simplicity, the following proofs proceed with the theoretical ones. Note that although the covariance matrix $\Sigma_{W,i}$ is not the parameter of interest, it explicitly appears in the likelihood function because its estimation has an effect on the magnitude and convergence of likelihood. Define

$$L_i(\boldsymbol{\Psi}_i, \Sigma_{W,i}, p_{Y,i}, \theta_l, \theta_u, q, d) = \mathrm{E}[l(\boldsymbol{\Psi}^q_i, \Sigma^{q_{\max}}_{W,i}; \boldsymbol{Y}_k, \ldots, \boldsymbol{Y}_{k-p_{Y,i}}) I(z_{k-d} \in (\theta_l, \theta_u])],$$

and

$$L(\boldsymbol{\Psi}, \boldsymbol{p}_Y, \boldsymbol{\theta}, q, d) = \sum_{k=1}^{n}\sum_{i=1}^{r}[l(\boldsymbol{\Psi}^q_i, \Sigma^{q_{\max}}_{W,i}; \boldsymbol{Y}_k, \ldots, \boldsymbol{Y}_{k-p_{Y,i}}) I(z_{k-d} \in (\theta_{i-1}, \theta_i])].$$

By extending Proposition 1 of Yau et al. (2015) that, for any $(\theta_l, \theta_u) \subseteq B_n = (\theta_{i-1} - c^1_n, \theta_i + c^2_n)$ with $(c^1_n : n \in \mathbb{N}^+)$ and $(c^2_n : n \in \mathbb{N}^+)$ are positive

sequences converging to 0, we have

$$\sup_{(\theta_l,\theta_u)\in B_n} \sup_{\boldsymbol{\Psi}_i,\Sigma_{W,i}} \left| L_{n,i}^{(u)}(\boldsymbol{\Psi}_i, \Sigma_{W,i}, p_{Y,i}, \theta_l, \theta_u, q, d) - L_i^{(u)}(\boldsymbol{\Psi}_i, \Sigma_{W,i}, p_{Y,i}, \theta_l, \theta_u, q, d) \right| \xrightarrow{a.s.} 0\,, \tag{S.21}$$

where $u = 0, 1, 2$ for u-th derivative with respect to $\boldsymbol{\Psi}_i$ and $\Sigma_{W,i}$ of $L_{n,i}$ and L_i, respectively; $L_{n,i}^{(0)} = L_{n,i}$ and $L_i^{(0)} = L_i$. Moreover, define

$$\{\hat{\boldsymbol{\Psi}}_i, \hat{\Sigma}_{W,i}\} = \arg \max_{\boldsymbol{\Psi}_i,\Sigma_{W,i}} L_{n,i}(\boldsymbol{\Psi}_i, \Sigma_{W,i}, p_{Y,i}, \theta_l, \theta_u, q, d)\,,$$

and

$$\{\boldsymbol{\Psi}_i^*, \Sigma_{W,i}^*\} = \arg \max_{\boldsymbol{\Psi}_i,\Sigma_{W,i}} L_i(\boldsymbol{\Psi}_i, \Sigma_{W,i}, p_{Y,i}, \theta_l, \theta_u, q, d)\,.$$

Then by using the uniform convergence of the likelihood function and its first and second derivatives in (S.21), we have

$$\hat{\boldsymbol{\Psi}}_i \xrightarrow{a.s.} \boldsymbol{\Psi}_i^* \quad \text{and} \quad \hat{\Sigma}_{W,i} \xrightarrow{a.s.} \Sigma_{W,i}^*\,. \tag{S.22}$$

Based on the results of equations (S.21)–(S.22), we can prove Theorem 2 of model orders estimation consistency.

Proof of Theorem 2

We prove by contradiction. Let A be the probability of one set under which (S.21) and (S.22) hold. For each $\omega \in A$, by the compactness of the parameter space and (S.22), there exists a subsequence $\{n_m\}$ of n such that $\hat{r}_{n_m} \longrightarrow r^\dagger$, $\hat{d}_{n_m} \to d^\dagger$, $\hat{\boldsymbol{\theta}}_{n_m} \to \boldsymbol{\theta}^\dagger$, and $\hat{q}_{n_m} \xrightarrow{p.} q^\dagger < q$ for some $r^\dagger$, $d^\dagger$, $q^\dagger$

and $\theta^\dagger$, respectively. Also, $\hat{p}_{Y,j} \to p^\dagger_{Y,j}$, $\hat{\Sigma}_{W,i} \to \Sigma^\dagger_{W,i}$ and $\hat{\mathbf{\Psi}}_j \to \mathbf{\Psi}^\dagger_j$ for some $p^\dagger_{Y,j}$, $\Sigma^\dagger_{W,i}$ and $\mathbf{\Psi}^\dagger_j$, respectively. For simplicity, we omit the subscript "n_m" and replace $\{n_m\}$ by $\{n\}$ when necessary.

First, suppose that $r^\dagger < r$, meaning that the number of assumed regimes is smaller than the number of true regimes. Without loss of generality, we assume that there exists a positive integer $b \geq 2$ such that

$$\theta^\dagger_{j-1} \leq \theta_i < \cdots < \theta_{i+b} \leq \theta^\dagger_j \,,$$

representing that several true regimes are merged into a "working regime j". By Assumption 4, at least one of $\mathbf{\Psi}_i \neq \mathbf{\Psi}^\dagger_j, \ldots, \mathbf{\Psi}_{i+b} \neq \mathbf{\Psi}^\dagger_j$ holds.

By the property of the Kullback-Leibler distance, it follows that

$$L_i(\mathbf{\Psi}_i, \Sigma_{W,i}, p_{Y,i}, q, \theta^\dagger_{j-1}, \theta_i, d) \geq L_i(\mathbf{\Psi}^\dagger_j, \Sigma^\dagger_{W,j}, p^\dagger_{Y,i}, \theta^\dagger_{j-1}, \theta_i, q^\dagger, d^\dagger) \,,$$

Similarly, it holds that $L_m(\mathbf{\Psi}_m, \Sigma_{W,m}, p_{Y,m}, \theta_{m-1}, \theta_m, q, d) \geq L_j(\mathbf{\Psi}^\dagger_j, \Sigma^\dagger_{W,j}, p^\dagger_{Y,j}, \theta_{m-1}, \theta_m, q^\dagger, d^\dagger)$ for $m = i+1, \ldots, i+b$. However, the equality cannot hold for all these expressions under Assumption 4. Equipping the ergodic theorem, there exists some $k > 0$ such that

$$L_i(\mathbf{\Psi}_i, \Sigma_{W,i}, p_{Y,i}, \theta^\dagger_{j-1}, \theta_i, q, d) + L_{b+1+i}(\mathbf{\Psi}_{b+1+i}, \Sigma_{W,b+1+i}, p_{Y,b+1+i}, \theta_{i+b}, \theta^\dagger_j, q, d) + \sum_{m=i+1}^{i+b} L_m(\mathbf{\Psi}_m, \Sigma_{W,m}, p_{Y,m}, \theta_{m-1}, \theta_m, q, d) - L_i(\mathbf{\Psi}^\dagger_j, \Sigma^\dagger_{W,j}, p^\dagger_{Y,j}, \theta^\dagger_{j-1}, \theta^\dagger_j, q^\dagger, d^\dagger) = k \,,$$

leading to the difference of $\mathrm{CL}(Y|\mathcal{M})$ having the order $O(\sum_{i=1}^r n_i)$. Mean-

while, denote $\text{Pen}(r, d, \boldsymbol{p}_Y) \equiv \text{CL}(\mathcal{M})$ in (S.43). By Assumption 3, we have

$$\text{Pen}(r^\dagger, d^\dagger, q^\dagger, \boldsymbol{p}_Y^\dagger) - \text{Pen}(r, d, q, \boldsymbol{p}_Y) = o\left(\sum_{i=1}^{r} n_i\right),$$

which is asymptotically dominated by the $O(\sum_{i=1}^{r} n_i)$ rate. Thus, $r^\dagger < r$ fails in minimizing the $\text{MDL}(\hat{\mathcal{M}})$. Moreover, if one of the working regimes is not nested in a true regime, for example, $\theta_{i-1} < \theta_{j-1}^\dagger < \theta_i < \theta_j^\dagger < \theta_{i+1}$ for some i, j. Then by the same analysis, the MDL value can be further minimized in probability by dividing the working regime into three sub-regimes $(\theta_{i-1}, \theta_{j-1}^\dagger]$, $(\theta_{j-1}^\dagger, \theta_i]$ and $(\theta_i, \theta_j^\dagger]$. Hence, for sufficiently large n, all working regimes are nested in some true regimes, i.e., $\theta_{i-1} < \theta_{j-1}^\dagger < \theta_j^\dagger < \theta_i$ for all j with some i, in probability.

Then, assume that $d^\dagger \neq d$. For any $(\theta_{j-1}^\dagger, \theta_j^\dagger) \subseteq (\theta_{i-1}, \theta_i)$, by the property of Kullback-Leibler distance,

$$L_i(\boldsymbol{\Psi}_\mathbf{i}, \Sigma_{W,i}, p_{Y,i}, \theta_{j-1}^\dagger, \theta_j^\dagger, d, q) - L_i(\boldsymbol{\Psi}_\mathbf{i}^\dagger, \Sigma_{W,i}^\dagger, p_{Y,i}^\dagger, \theta_{j-1}^\dagger, \theta_j^\dagger, d^\dagger, q^\dagger) \neq 0\,,$$

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

$$L(\boldsymbol{\Psi}, \boldsymbol{p}_Y, \boldsymbol{\theta}, q, d) - L(\boldsymbol{\Psi}^\dagger, \boldsymbol{p}_Y^\dagger, \boldsymbol{\theta}^\dagger, q^\dagger, d^\dagger,) = O(\sum_{i=1}^{r} n_i)\,,$$

which is strictly positive and dominates $\text{Pen}(r^\dagger, d^\dagger, q^\dagger, \boldsymbol{p}_Y^\dagger) - \text{Pen}(r, d, q, \boldsymbol{p}_Y) = o(\sum_{i=1}^{r} n_i)$ by (4.11) in probability. Hence, we have $d^\dagger = d$. With the similar arguments, we can prove $p_{Y,i}^\dagger \geq p_{Y,i}$.

Next, assume $q^\dagger < q$. Denote the set of all non-zero coefficients in the qth element of $\boldsymbol{a}_i$, the qth row or the qth column in $\boldsymbol{\Psi}_{i,j}$, for $i = 1, \ldots, p_{Y,i}$ and $j = 1, \ldots, p_{Y,i}$, as B. Denote β as the smallest absolute value among all the elements in B. By Assumption 3, β is a positive value. Then,

$$L(\boldsymbol{\Psi}^\dagger, \boldsymbol{p}_Y^\dagger, \boldsymbol{\theta}^\dagger, q^\dagger, d) - L(\boldsymbol{\Psi}, \boldsymbol{p}_Y, \boldsymbol{\theta}, q, d) > \beta\lambda_{min}\mathrm{E}(\boldsymbol{Y}_k\boldsymbol{Y}_k^{\mathrm{T}}) = O(\sum_{i=1}^{r} n_i).$$

On the other hand, the penalty term has order $O(\log n)$. Thus, minimizing the fMDL criterion cannot lead to $q^\dagger < q$.

On the other hand, assume at least one of $r^\dagger > r$, $p_{Y,i}^\dagger > p_{Y,i}$ or $q^\dagger > q$ holds. Without loss of generality, we assume that $\theta_{i-1} < \theta_l = \theta_{j-1}^\dagger < \theta_j^\dagger = \theta_u < \theta_i$ since we have proved that all working regimes are nested in true regimes. By Taylor's expansion and the results of (S.21), it follows that

$$\begin{aligned}
& L_{n,i}(\hat{\boldsymbol{\Psi}}_i, \hat{\Sigma}_{W,i}, \hat{p}_{Y,i}, \theta_l, \theta_u, \hat{q}_n, d) - L_{n,i}(\boldsymbol{\Psi}_i, \Sigma_{W,i}, p_{Y,i}, \theta_l, \theta_u, q, d) \\
= \; & (\hat{\boldsymbol{\Psi}}_i - \boldsymbol{\Psi}_i, \hat{\Sigma}_{W,i} - \Sigma_{W,i})^{\mathrm{T}} L_{n,i}^{(1)}(\boldsymbol{\Psi}_i, \Sigma_{W,i}, p_{Y,i}, \theta_l, \theta_u, q, d) \\
& + (\hat{\boldsymbol{\Psi}}_i - \boldsymbol{\Psi}_i, \hat{\Sigma}_{W,i} - \Sigma_{W,i})^{\mathrm{T}} L_{n,i}^{(2)}(\boldsymbol{\Psi}_i, \Sigma_{W,i}, p_{Y,i}, \theta_l, \theta_u, q, d)(\hat{\boldsymbol{\Psi}}_i - \boldsymbol{\Psi}_i, \hat{\Sigma}_{W,i} - \Sigma_{W,i}) + o(1) \\
= \; & (\hat{\boldsymbol{\Psi}}_i - \boldsymbol{\Psi}_i, \hat{\Sigma}_{W,i} - \Sigma_{W,i})^{\mathrm{T}} L_{n,i}^{(2)}(\boldsymbol{\Psi}_i, \Sigma_{W,i}, p_{Y,i}, \theta_l, \theta_u, q, d)(\hat{\boldsymbol{\Psi}}_i - \boldsymbol{\Psi}_i, \hat{\Sigma}_{W,i} - \Sigma_{W,i}) + o(1)\,.
\end{aligned} \tag{S.23}$$

Since

$$\begin{aligned}
& L_{n,i}^{(1)}(\hat{\boldsymbol{\Psi}}_i, \hat{\Sigma}_{W,i}, \hat{p}_{Y,i}, \theta_l, \theta_u, \hat{q}_n, d) - L_{n,i}^{(1)}(\boldsymbol{\Psi}_i, \Sigma_{W,i}, p_{Y,i}, \theta_l, \theta_u, q, d) \\
= \; & (\hat{\boldsymbol{\Psi}}_i - \boldsymbol{\Psi}_i, \hat{\Sigma}_{W,i} - \Sigma_{W,i}) L_{n,i}^{(2)}(\boldsymbol{\Psi}_i, \Sigma_{W,i}, p_{Y,i}, \theta_l, \theta_u, q, d) + o(1)\,,
\end{aligned} \tag{S.24}$$

with the left-hand-side of (S.24) converges to 0 almost surely, by Kolmogorov's law of iterated logarithm,

$$\begin{aligned}&\limsup_{n\to\infty}\frac{n^{1/2}}{(\log\log(n))^{1/2}}L_{n,i}^{(2)}(\boldsymbol{\Psi}_i,\Sigma_{W,i},p_{Y,i},\theta_l,\theta_u,q,d)\\ \xrightarrow{a.s.}\ &\sqrt{2}[\mathrm{var}(L_i^{(2)}(\boldsymbol{\Psi}_i,\Sigma_{W,i},p_{Y,i},\theta_l,\theta_u,q,d))]^{1/2}\,,\end{aligned}$$

and thus by (S.24), $\|\hat{\boldsymbol{\Psi}}_i-\boldsymbol{\Psi}_i\|_2=O((n^{-1}\log\log(n))^{1/2})$ and $\|\hat{\Sigma}_{W,i}-\Sigma_{W,i}\|_2=O((n^{-1}\log\log(n))^{1/2})$ almost surely. By (S.23),

$$L(\boldsymbol{\Psi}^\dagger,\boldsymbol{p}_Y^\dagger,\boldsymbol{\theta}^\dagger,q^\dagger,d)-L(\boldsymbol{\Psi},\boldsymbol{p}_Y,\boldsymbol{\theta},q,d)=O(\sum_{i=1}^r\log\log(n_i))\,,$$

which is dominated by the penalty rate $O(q^2\sum_{i=1}^r\log n_i)$. Hence, $r^\dagger>r$ or $p_{Y,i}^\dagger>p_{Y,i}$ or $q^\dagger>q$ fails in minimizing the MDL($\hat{\mathcal{M}}$). Hence, $\hat{r}\to r$, $\hat{q}_n\to q$ and $\hat{p}_{Y,i}\to p_{Y,i}$ almost surely; and as all estimated regimes are nested in true regimes in probability, the only possible case that remains is $\hat{\boldsymbol{\theta}}_n\to\boldsymbol{\theta}$ almost surely. Also, by equations (S.20) and (S.22), $\hat{\boldsymbol{\Psi}}_i\to\boldsymbol{\Psi}_i$ and $\hat{\Sigma}_{W,i}\to\Sigma_{W,i}$ almost surely for all i.

Proof of Theorem 3

First, we verify the ρ-mixing property of

$$\{\boldsymbol{Y}_{k-j}I(z_{k-d}\in[\theta_i-\Delta,\theta_i+\Delta]),I(z_{k-d}\in[\theta_i-\Delta,\theta_i+\Delta])\}\,,\qquad\text{(S.25)}$$

for some $0<\Delta<1$. Since for $0<\Delta_1<\Delta_2<1$, $z_{k-d}I(z_{k-d}\in[\theta_i-\Delta_1,\theta_i+\Delta_1])$ and $I(z_{k-d}I(z_{k-d}\in[\theta_i-\Delta_1,\theta_i+\Delta_1])\in[\theta_i-\Delta_1,\theta_i+\Delta_1]))$ are

measurable functions of $z_{k-d}I(z_{k-d} \in [\theta_i - \Delta_2, \theta_i + \Delta_2])$, by Assumption 7, the process $\{I(z_{k-d}I(z_{k-d} \in [\theta_i - \Delta, \theta_i + \Delta]))\}$ is ρ-mixing with summable mixing coefficients $\{\rho(m)\}$. Next, by Assumption 6,

$$I(z_{k-d}I(z_{k-d} \in [\theta_i - \Delta, \theta_i + \Delta]) \in [\theta_i - \Delta, \theta_i + \Delta]) = I(z_{k-d} \in [\theta_i - \Delta, \theta_i + \Delta]),$$

except for the point $z_{k-d} = 0$ with measure 0. By the dominated convergence theorem, we have $\{I(z_{k-d} \in [\theta_i - \Delta, \theta_i + \Delta])\}$ is ρ-mixing with summable mixing coefficients $\{\rho(m)\}$. By Assumption 1, $Y_k \in L_H^4$ and hence $\mathrm{E}\|\boldsymbol{Y}_k\|_2^4 < \infty$. Since

$$\boldsymbol{Y}_{k-j}I(z_{k-d} \in [\theta_i - \Delta, \theta_i + \Delta]) = \boldsymbol{Y}_{k-j}I(z_{k-j-d} \in [\theta_i - \Delta^*, \theta_i + \Delta^*])\frac{I(z_{k-d} \in [\theta_i - \Delta, \theta_i + \Delta])}{I(z_{k-j-d} \in [\theta_i - \Delta^*, \theta_i + \Delta^*])},$$

where $0 < \Delta^* < 1$, (S.25) is ρ-mixing with summable mixing coefficients $\{\rho(m)\}$ for all $j = 0, \ldots, p$ and all sufficiently small Δ.

Then, we use the ρ-mixing property of (S.25) to establish several important inequalities. Denote $Q(z) = \mathrm{E}(I(z_{k-d} \leq z))$ for some constant z. Since $\epsilon_k \in L_H^4$, we have $\mathrm{E}|w_{k,i}^{(l)}|^4 < \infty$ for all l. By the ρ-mixing property of (S.25) and Cauchy-Schwarz inequality, there exists some $M > 0$ for $m \geq 1$

such that

$$
\begin{aligned}
& |\mathrm{Cov}\{w_{k,i+1}^{(l)} I(z_{k-d} \in (z_1, z_2]), w_{k+m,i+1}^{(l')} I(z_{k+m-d} \in (z_1, z_2])\}| \\
= \; & |\mathrm{Cov}(w_{k,i+1}^{(l)}, w_{k+m,i+1}^{(l')}) \mathrm{Cov}(I(z_{k-d} \in (z_1, z_2]), I(z_{k+m-d} \in (z_1, z_2]))| \\
\leq \; & (\mathrm{E}|w_{k,i+1}^{(l)}|^2)^{1/2} (\mathrm{E}|w_{k+m,i+1}^{(l')}|^2)^{1/2} |\mathrm{Cov}\{I(z_{k-d} \in (z_1, z_2]), I(z_{k+m-d} \in (z_1, z_2])\}| \\
\leq \; & |\rho(m)| (\mathrm{E}|w_{k,i+1}^{(l)}|^2) \left[\mathrm{E}\{I(z_{k-d} \in (z_1, z_2])\}^2\right]^{1/2} \left[\mathrm{E}\{I(z_{k+m-d} \in (z_1, z_2])\}^2\right]^{1/2} \\
\leq \; & M|\rho(m)|(Q(z_2) - Q(z_1)) \,. \qquad \text{(S.26)}
\end{aligned}
$$

for all $l, l' = 1, \ldots, q$ and $\theta_i - \Delta \leq z_1 < z_2 \leq \theta_i + \Delta$. Analogously,

$$
|\mathrm{Cov}\{|w_{k,i+1}^{(l)}| I(z_{k-d} \in (z_1, z_2]), |w_{k+m,i+1}^{(l')}| I(z_{k+m-d} \in (z_1, z_2])\}| \leq M|\rho(m)|(Q(z_2) - Q(z_1)) \,, \qquad \text{(S.27)}
$$

and

$$
|\mathrm{Cov}\{(w_{k,i+1}^{(l)})^2 I(z_{k-d^0} \in (z_1, z_2]), (w_{k+m,i+1}^{(l')})^2 I(z_{k+m-d} \in (z_1, z_2])\}| \leq M|\rho(m)|(Q(z_2) - Q(z_1)) \,, \qquad \text{(S.28)}
$$

for all $l, l' = 1, \ldots, q$ and $\theta_i \leq z_1 < z_2 \leq \theta_i + \Delta$.

By Assumption 1, $Y_k \in L_H^4$ and thus $\mathrm{E}\|\boldsymbol{Y}_k\|_2^4 < \infty$. It follows that

$$
\begin{aligned}
& |\mathrm{Cov}\{|w_{k,i+1}^{(l)} y_{k-j,l}| I(z_{k-d} \in (z_1, z_2]), |w_{k+m,i+1}^{(l')} y_{k+m-j,l}| I(z_{k+m-d} \in (z_1, z_2])\}| \\
\leq \; & (\mathrm{E}|w_{k,i+1}^{(l)}|^2)^{1/2} (\mathrm{E}|w_{k+m,i+1}^{(l')}|^2)^{1/2} \mathrm{Cov}\{|y_{k-j,l}| I(z_{k-d} \in (z_1, z_2]), |y_{k+m-j,l}| I(z_{k+m-d} \in (z_1, z_2])\}| \\
\leq \; & |\rho(m)| \mathrm{E}|w_{k,i+1}^{(l)}|^2 \left[\mathrm{E}\{y_{k-j,l} I(z_{k-d} \in (z_1, z_2])\}^2\right]^{1/2} \left[\mathrm{E}\{y_{k+m-j,l} I(z_{k+m-d} \in (z_1, z_2])\}^2\right]^{1/2} \\
\leq \; & M|\rho(m)|(Q(z_2) - Q(z_1)) \,, \qquad \text{(S.29)}
\end{aligned}
$$

with some $M > 0$, for all $l, l' = 1, \ldots, q$, $j = 1, \ldots, p_{Y,i}$ and $\theta_i \leq z_1 < z_2 \leq \theta_i + \Delta$.

For verifying the convergence rate of $\hat{\boldsymbol{\theta}}_n$, it suffices to prove the $O_p(n^{-1})$ convergence rate of $\hat{\theta}_i$ to θ_i. Without loss of generality, assume that $\hat{\theta}_i > \theta_i$, and thus denote $\hat{\theta}_i = \theta_i + a$ for some $a > 0$. Then define $Q(\theta_i, a) = \mathrm{E}(I(\theta_i \leq z_{k-d} \leq \theta_i + a))$. By using (S.26)–(S.29), we are ready to apply the argument in verifying Claim 2 in Chan (1993), from which we have for all $\epsilon > 0$ and $\zeta > 0$, there exist some $c > 0$ such that for all n,

$$\mathrm{pr}\left(\sup_{c/n<a\leq\Delta}\left|\sum_{k=1}^{n}\frac{I(\theta_i < z_{k-d} \leq \theta_i + a)}{nQ(\theta_i, a)} - 1\right| < \zeta\right) > 1 - \epsilon\,, \tag{S.30}$$

$$\mathrm{pr}\left(\sup_{c/n<a\leq\Delta}\left\|\sum_{k=1}^{n}\frac{\boldsymbol{W}_{k,i+1}I(\theta_i < z_{k-d} \leq \theta_i + a)}{nQ(\theta_i, a)}\right\|_2 < \zeta\right) > 1 - \epsilon\,, \tag{S.31}$$

$$\mathrm{pr}\left(\sup_{c/n<a\leq\Delta}\left|\sum_{k=1}^{n}\frac{\boldsymbol{W}_{k,i+1}^{\mathrm{T}}\boldsymbol{Y}_{k-j}I(\theta_i < z_{k-d} \leq \theta_i + a)}{nQ(\theta_i, a)}\right| < \zeta\right) > 1-\epsilon\,, \forall j = 1, \ldots, p_{Y,i}\,, \tag{S.32}$$

and

$$\mathrm{pr}\left(\sup_{c/n<a\leq\Delta}\left\|\sum_{k=1}^{n}\frac{\boldsymbol{W}_{k,i+1}\boldsymbol{W}_{k,i+1}^{\mathrm{T}}I(\theta_i < z_{k-d} \leq \theta_i + a)}{nQ(\theta_i, a)} - \Sigma_{W,i+1}\right\|_2 < \zeta\right) > 1-\epsilon\,. \tag{S.33}$$

Then, we define

$$\begin{aligned} L_n^*(\theta_i, a) &= \frac{1}{Q(\theta_i, a)} \Big[L_{n,i}(\hat{\boldsymbol{\Psi}}_i, \hat{\Sigma}_{W,i}, p_{Y,i}, \theta_{i-1}, \hat{\theta}_i, d) + L_{n,i}(\hat{\boldsymbol{\Psi}}_{i+1}, \hat{\Sigma}_{W,i+1}, p_{Y,i}, \hat{\theta}_i, \theta_{i+1}, d) \\ &\quad - L_{n,i}(\hat{\boldsymbol{\Psi}}_i, \hat{\Sigma}_{W,i}, p_{Y,i}, \theta_{i-1}, \theta_i, d) - L_{n,i}(\hat{\boldsymbol{\Psi}}_{i+1}, \hat{\Sigma}_{W,i+1}, p_{Y,i}, \theta_i, \theta_{i+1}, d) \Big] \\ &= -\frac{1}{nQ(\theta_i, a)} \sum_{k=1}^{n} \Bigg[\log\det(\hat{\Sigma}_{W,i}) - \log\det(\hat{\Sigma}_{W,i+1}) + \hat{\boldsymbol{W}}_{k,i+1}^{\mathrm{T}} (\hat{\Sigma}_{W,i}^{-1} - \hat{\Sigma}_{W,i+1}^{-1}) \hat{\boldsymbol{W}}_{k,i+1} \\ &\quad + \left(\Delta^*_{\boldsymbol{W}_{k,i}} \right)^{\mathrm{T}} \hat{\Sigma}_{W,i}^{-1} \Delta^*_{\boldsymbol{W}_{k,i}} + 2 \hat{\boldsymbol{W}}_{k,i+1}^{\mathrm{T}} \hat{\Sigma}_{W,i}^{-1} \Delta^*_{\boldsymbol{W}_{k,i}} \Bigg] I(z_{k-d} \in (\theta_i, \theta_i + a]) \,, \end{aligned}$$

where

$$\hat{\boldsymbol{W}}_{k,i} = \boldsymbol{Y}_k - \hat{\boldsymbol{a}}_i - \sum_{j=1}^{p_{Y,i}} \hat{\boldsymbol{\Psi}}_{i,j} \boldsymbol{Y}_{k-j} - \sum_{m=1}^{p_X} \hat{\boldsymbol{\Phi}}_{i,m} X_{k,m} \,,$$

and $\Delta^*_{\boldsymbol{W}_{k,i}} = \hat{\boldsymbol{W}}_{k,i} - \hat{\boldsymbol{W}}_{k,i+1}$. Since $\boldsymbol{W}_{k,i+1} \sim N(\boldsymbol{0}, \Sigma_{W,i+1})$, by the Kullback-Leibler distance, we have

$$\mathrm{E}\left[\log\det(\Sigma_{W,i}) - \log\det(\Sigma_{W,i+1}) + \boldsymbol{W}_{k,i+1}^{\mathrm{T}} \Sigma_{W,i}^{-1} \boldsymbol{W}_{k,i+1} - \boldsymbol{W}_{k,i+1}^{\mathrm{T}} \Sigma_{W,i+1}^{-1} \boldsymbol{W}_{k,i+1} \right] > 0 \,. \tag{S.34}$$

Hence, for $a > c/n$ with some constant $c > 0$, by Assumption 4, (S.30), (S.33) and (S.34),

$$\begin{aligned} &-\frac{1}{nQ(\theta_i, a)} \sum_{i=1}^{n} \left[\log\det(\hat{\Sigma}_{W,i}) - \log\det(\hat{\Sigma}_{W,i+1}) + \hat{\boldsymbol{W}}_{k,i+1}^{\mathrm{T}} (\hat{\Sigma}_{W,i}^{-1} - \hat{\Sigma}_{W,i+1}^{-1}) \hat{\boldsymbol{W}}_{k,i+1} \right] \\ &\times I(z_{k-d} \in (\theta_i, \theta_i + a]) < 0 \,, \end{aligned}$$

in probability. In addition, by (S.31) and (S.32),

$$\frac{1}{nQ(\theta_i, a)} \sum_{i=1}^{n} \hat{\boldsymbol{W}}_{k,i+1}^{\mathrm{T}} \hat{\Sigma}_{W,i}^{-1} \Delta^*_{\boldsymbol{W}_{k,i}} I(z_{k-d} \in (\theta_i, \theta_i + a]) \xrightarrow{p} 0 \,, \quad \text{as} \quad n \to \infty \,.$$

Since $\hat{\Sigma}_{W,i}$ is semi-positive definite, $(\Delta^*_{\boldsymbol{W}_{k,i}})^{\mathrm{T}}\hat{\Sigma}^{-1}_{W,i}\Delta^*_{\boldsymbol{W}_{k,i}} \geq 0$ for all k. Therefore, analogous to Proposition 4 in Yau et al. (2015), $L^*(\theta_i, a) < 0$ in probability. Hence, we obtain that

$$\begin{aligned} & L_{n,i}(\hat{\boldsymbol{\Psi}}_i, \hat{\Sigma}_{W,i}, p_{Y,i}, \theta_{i-1}, \hat{\theta}_i, d) + L_{n,i}(\hat{\boldsymbol{\Psi}}_{i+1}, \hat{\Sigma}_{W,i+1}, p_{Y,i}, \hat{\theta}_i, \theta_{i+1}, d) \\ & -L_{n,i}(\hat{\boldsymbol{\Psi}}_i, \hat{\Sigma}_{W,i}, p_{Y,i}, \theta_{i-1}, \theta_i, d) - L_{n,i}(\hat{\boldsymbol{\Psi}}_{i+1}, \hat{\Sigma}_{W,i+1}, p_{Y,i}, \theta_i, \theta_{i+1}, d) \\ = \; & Q(\theta_i, a) L^*_n(\theta_i, a) \end{aligned} \tag{S.35}$$

is strictly negative. If $\hat{\theta}_i < \theta_i$, the result (S.35) can be shown similarly.

From Theorem 2, we know that $\hat{r}_n \geq r$ and for each θ_i there exists a $\hat{\theta}_{i'}$ such that $|\hat{\theta}_{i'} - \theta_i| = o_p(1)$ where $1 \leq i' \leq \hat{r}_n$. To prove Theorem 3, it suffices to show that for each $\epsilon > 0$, there exits a $c > 0$ such that

$$pr(|\hat{\theta}_{i'} - \theta_i| > c/n) < \epsilon\,.$$

By definition, the estimates $(r, d, \hat{\boldsymbol{\theta}}_n, q, \boldsymbol{p}_Y)$ minimizes the fMDL. Thus we have for each $j = 1, \ldots, r$ that $\mathrm{fMDL}(\hat{\boldsymbol{\theta}}_n) \leq \mathrm{fMDL}(\tilde{\boldsymbol{\theta}}^i_n)$, where $\tilde{\boldsymbol{\theta}}^i_n$ is the same as $\hat{\boldsymbol{\theta}}_n$ except that $\hat{\theta}_{i'}$ is replaced by θ_i, and $\mathrm{fMDL}(\hat{\boldsymbol{\theta}}_n)$ and $\mathrm{fMDL}(\tilde{\boldsymbol{\theta}}^i_n)$ are the fMDL evaluated at $(r, d, \hat{\boldsymbol{\theta}}_n, q, \boldsymbol{p}_Y)$ and $(r, d, \tilde{\boldsymbol{\theta}}^i_n, q, \boldsymbol{p}_Y)$, respectively. Therefore, it in turns suffices to show that for each $i = 1, \ldots, r$,

$$pr(\mathrm{fMDL}(\hat{\boldsymbol{\theta}}_n) \leq \mathrm{fMDL}(\tilde{\boldsymbol{\theta}}^i_n), |\hat{\theta}_{i'} - \theta_i| > c/n) < \epsilon\,. \tag{S.36}$$

Given that $|\hat{\theta}_{i'} - \theta_i| > c/n$, the quantity $\mathrm{fMDL}(\tilde{\boldsymbol{\theta}}^i_n) - \mathrm{fMDL}(\hat{\boldsymbol{\theta}}_n)$ is equal to

either one of

$$i) \quad L_n^*(\theta_i, a) I(\hat{\theta}_i > \theta_i)\,,$$

$$ii) \quad L_n^*(\theta_i, a) I(\hat{\theta}_i < \theta_i)\,. \tag{S.37}$$

By using the result of 2, we can obtain that both terms in (S.37) are strictly negative in probability, i.e., $pr(\text{fMDL}(\tilde{\boldsymbol{\theta}}_n^i) - \text{fMDL}(\hat{\boldsymbol{\theta}}_n) < 0) \to 0$, yielding (S.36). □

Proof of Theorem 4

First, we prove the convergence of $n(\hat{\theta}_i - \theta_i)$ to M_i^-; and then the result is elaborated to the convergence of $n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})$. Suppose that $\hat{\theta}_i = \theta_i + \kappa_i/n$ with some $\kappa_i > 0$, and define

$$\hat{\mathcal{P}}_i^*(\kappa_i) = \sum_{k=1}^n \hat{\xi}_k^{(i+1,i)} I(z_{k-d} \in (\theta_i, \theta_i + \kappa_i/n]) + \sum_{k=1}^n \hat{\xi}_k^{(i,i+1)} I(z_{k-d} \in (\theta_i - \kappa_i/n, \theta_i])\,,$$

where

$$\begin{aligned} \hat{\xi}_k^{(i,j)} &= l(\hat{\boldsymbol{\Psi}}_i, \hat{\boldsymbol{\Phi}}_i; \boldsymbol{Y}_k, \ldots, \boldsymbol{Y}_{k-p_{Y,i}}, \boldsymbol{X}_{k,1}, \ldots, \boldsymbol{X}_{k,p_X}) \\ &\quad - l(\hat{\boldsymbol{\Psi}}_j, \hat{\boldsymbol{\Phi}}_j; \boldsymbol{Y}_k, \ldots, \boldsymbol{Y}_{k-p_{Y,j}}, \boldsymbol{X}_{k,1}, \ldots, \boldsymbol{X}_{k,p_X})\,. \end{aligned}$$

and

$$\begin{aligned} \mathcal{P}_i^*(\kappa_i) &= \sum_{k=1}^n \xi_k^{(i+1,i)} I(z_{k-d} \in (\theta_i, \theta_i + \kappa_i/n]) + \sum_{k=1}^n \xi_k^{(i,i+1)} I(z_{k-d} \in (\theta_i - \kappa_i/n, \theta_i]) \\ &= \mathcal{P}_{i,1}^*(\kappa_i) + \mathcal{P}_{i,2}^*(\kappa_i)\,. \end{aligned}$$

where $\xi_k^{(i,j)}$ is given in (5.14). In addition, denote $\hat{\boldsymbol{\Psi}}_i(\theta_i)$ as the estimator of $\boldsymbol{\Psi}_i$ given the true threshold θ_i. Since $\mathrm{E}\|\boldsymbol{Y}_k\|^4 < \infty$, by Taylor's expansion as in Lemma 3 of Samia and Chan (2011) and lengthy calculations, it can be derived that

$$\sup_{|\hat{\theta}_i - \theta_i| < K/n} \|\hat{\boldsymbol{\Psi}}_i - \hat{\boldsymbol{\Psi}}_i(\theta_i)\|_2 = o_p(n^{-1/2}) \,. \tag{S.38}$$

for some $K > 0$. Together with this result and (S.20),

$$\sup_{|\kappa_i| < K} |\hat{\mathcal{P}}_i^*(\kappa_i) - \mathcal{P}_i^*(\kappa_i)| = o_p(1) \,. \tag{S.39}$$

Owing to (S.39), we shall proceed as if $\hat{\mathcal{P}}_i^*(\kappa_i) = \mathcal{P}_i^*(\kappa_i)$. For illustration, we only show $\mathcal{P}_{i,1}^*(\kappa_i)$ converges to $\mathcal{P}_{i,1}(\kappa_i)$ in distribution. By using similar arguments, it can be shown that $\mathcal{P}_{i,2}^*(\kappa_i)$ converges to $\mathcal{P}_{i,2}(\kappa_i)$ in distribution.

Note that $\kappa_i > 0$, thus it suffices to study on the right-continuous version of $\mathcal{P}_{i,1}^*(\kappa_i)$. Let $\epsilon = 1/n$. Consider the following process indexed by ϵ: $\eta^\epsilon(t) = X_{\lfloor nt \rfloor}^\epsilon$ for $t \in [0,1]$, where $\lfloor \cdot \rfloor$ denotes the integer part of the expression inside the bracket; $X_0^\epsilon = 0$, $X_{k+1}^\epsilon = X_k^\epsilon + \omega_{k+1}$, and

$$\omega_k = \xi_k^{(i+1,i)} I(z_{k-d} \in (\theta_i, \theta_i + \kappa_i/n]) \,.$$

Thus, $\eta^\epsilon(1) = \ell_{i,1}^*(\kappa_i)$ and $\eta^\epsilon(v) = X_s^\epsilon$ for any $v \in [s/n, s/n + 1/n)$. In addition, we define $\{\mathcal{C}(v), 0 \le v \le 1\}$ as the compound Poisson process with intensity $\pi_z(\theta_i)$, where $\pi_z(\cdot)$ is the density function of z_1 and the distribution

of jump is the same as the conditional distribution of $\xi_k^{(i+1,i)}$ given $z_{k-d} = {\theta_i}^+$. In the following, we show that $\{\eta^\epsilon(v)\}$ converges weakly to $\{\mathcal{C}(v)\}$, which verifies that $\mathcal{P}_{i,1}^*(\kappa_i)$ converges in distribution to $\mathcal{P}_{i,1}(\kappa_i)$.

By Assumptions 1 and 6–7, the process $\mathcal{P}_{i,1}^*(\kappa_i)$ is tight by applying a similar argument in Lemma 3.2 of Ibragimov and Has'minskii (1981). Then based on the results in Kurz (1975) and Kushner (1980), we use the operator convergence method in Kushner (1980) to prove the weak convergence of $\{\eta^\epsilon(v)\}$. Denote $\mathcal{F}_v^\epsilon$ as the σ-field generated by $\{\eta^\epsilon(v)\}$. Choose τ^ϵ such that $\tau^\epsilon \to 0$ and $\sqrt{n}\tau^\epsilon \to \infty$. For any continuous bounded real-valued function f with compact support and continuous second derivative f, define $f^\epsilon(v) = \frac{1}{\tau^\epsilon}\int_0^{\tau^\epsilon} \mathrm{E}[f(\eta^\epsilon(v+u))|\mathcal{F}_v^\epsilon]\,du$. By using the definition of p-lim in Kurz (1975), let $\hat{A}^\epsilon$ be the p-infinitesimal operator such that $\hat{A}^\epsilon f(v) = p\text{-}\lim_{\delta\to 0} \mathrm{E}(f(v+\delta) - f(v) \mid \mathcal{F}_v^\epsilon)/\delta$. After some direct calculations, it follows that

$$\begin{aligned}
\hat{A}^\epsilon f^\epsilon(v) &= \frac{1}{\tau^\epsilon}\{\mathrm{E}[f^\epsilon(v+\tau^\epsilon)|\mathcal{F}_v^\epsilon] - f^\epsilon(v)\} \\
&= \frac{1}{\tau^\epsilon}\sum_{k=\lfloor nv\rfloor}^{\lfloor n(v+\tau^\epsilon)\rfloor - 1} \mathrm{E}\left[f\left(X_k^\epsilon + \omega_{k+1}\right) - f(X_k^\epsilon)\middle|\mathcal{F}_v^\epsilon\right],
\end{aligned}$$

where the tightness of $\eta^\epsilon(v)$ and $\hat{A}^\epsilon f^\epsilon(v)$ can be shown by using similar arguments in Chan (1993). Then, we define the operator A such that

$$Af(\eta) = \pi(\theta_i)\kappa_i \int \left(f(\eta+u) - f(u)\right) q(du)\,,$$

where $q(du)$ is the probability measure induced by the conditional distribution of $\xi_k^{(i+1,i)}$ given $z_{k-d} = {\theta_i}^+$. By Assumptions 6–7 and applying the same argument of Theorem 3 in Samia and Chan (2011), we have

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

Thus, the limit of any convergent subsequence of $\{\eta^\epsilon(v), 0 \le v \le 1\}$ is the unique solution of

$$f(\eta(t)) - \int_0^t Af(\eta(s))ds\,,$$

which is a martingale (see Strook and Varadhan (1971)). By Theorem 1 in Kushner (1980), $\{\eta^\epsilon(v), 0 \le v \le 1\}$ converges weakly in $D[0,1]$ to $\{\mathcal{C}(v), 0 \le v \le 1\}$, which leads that $\mathcal{P}^*_{i,1}(\kappa_i)$ converges weakly to $\mathcal{P}_{i,1}(\kappa_i)$. Moreover, by using the Cramér-Wold device, $\mathcal{P}^*_i(\kappa_i)$ converges weakly to $\mathcal{P}_i(\kappa_i)$. Finally, by using the Theorem 3.1 in Seijo and Sen (2011), there exists a random interval $[M_-^{(i)}, M_+^{(i)})$ on which the process $\mathcal{P}_i(\kappa_i)$ attains its global minimum and $n(\hat{\theta}_i - \theta_i)$ converges weakly to M_i^-. It is readily seen that $n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})$ converges weakly to $\mathbf{M}$, where $[\mathbf{M}_-, \mathbf{M}_+)$ is a r-dimensional random cube on which the process $l(\boldsymbol{\kappa})$ attains its global minimum. □

Proof of Theorem 5

Denote $\hat{\boldsymbol{\beta}}_i(\hat{\boldsymbol{\theta}})$ as the estimator of $\boldsymbol{\beta}_i$ given the true thresholds $\boldsymbol{\theta}$. By

(S.38) and rearranging the coefficients,

$$\sup_{\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\|_2 < K/n} \|\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_i(\boldsymbol{\theta})\| = o_p(n^{-1/2}) \,. \tag{S.40}$$

In addition, by (S.40),

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i) = \sqrt{n}(\hat{\boldsymbol{\beta}}_i(\boldsymbol{\theta}) - \boldsymbol{\beta}_i) + o_p(1) \,,$$

and thus it suffices to study the asymptotic distribution of $\sqrt{n}(\hat{\boldsymbol{\beta}}_i(\boldsymbol{\theta}) - \boldsymbol{\beta}_i)$. Denote

$$\Gamma_i^e = \mathrm{E}[\mathrm{vec}(1, \boldsymbol{Y}_{k-1}^e, \ldots, \boldsymbol{Y}_{k-p_{Y,i}}^e, \boldsymbol{X}_{k,1}^e, \ldots, \boldsymbol{X}_{k,\boldsymbol{p}_X}^e)(\mathrm{vec}(1, \boldsymbol{Y}_{k-1}^e, \ldots, \boldsymbol{Y}_{k-p_{Y,i}}^e, \boldsymbol{X}_{k,1}^e, \ldots, \boldsymbol{X}_{k,\boldsymbol{p}_X}^e))^{\mathrm{T}}] \,.$$

As $\boldsymbol{Y}_k^e \xrightarrow{p.} \boldsymbol{Y}_k$ and $\boldsymbol{X}_{k,m}^e \xrightarrow{p.} \boldsymbol{X}_{k,m}$, we have $\Gamma_i^e \xrightarrow{p.} \Gamma_i$. By (S.38), Theorem 5.41 in van der Vaart (2000) and Proposition 3.1 in Lütkepohl (2006), $\sqrt{n}(\hat{\boldsymbol{\beta}}_i(\boldsymbol{\theta}) - \boldsymbol{\beta}_i)$ is asymptotically normal with mean $\mathbf{0}$ and covariance matrix $\Gamma_i^{-1} \otimes \Sigma_{W,i}$, and hence the asymptotic distribution of $\sqrt{n}(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)$ follows. Analogous to the arguments in Theorem 2 of Chan (1993), the convergence rate of $\hat{\theta}_i$ is much faster, and hence the asymptotic independence between $n(\hat{\theta}_i - \theta_i)$ and $\sqrt{n}(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)$ for all i follows. □

C: MDL criterion for the functional TARX model

In this section, we extend the MDL criterion for the functional TARX model. Denote $\boldsymbol{p}_X = \{\boldsymbol{p}_{X,1}, \ldots, \boldsymbol{p}_{X,r}\}$ as a collection of binary indicator

vector $\boldsymbol{p}_{X,i} = (p_{X,i,1}, \ldots, p_{X,i,p_X})^{\mathrm{T}}$, where $p_{X,i,m} = 1$ if $X_{k,m}$ is present in regime i and 0 otherwise. Also, define $\boldsymbol{q}_X = (q_{X,1}, \ldots, q_{X,p_X})^{\mathrm{T}}$ as the number of basis functions and $\boldsymbol{\Phi}^{q,\boldsymbol{q}_X} = (\boldsymbol{\Phi}_1^{q,\boldsymbol{q}_X}, \ldots, \boldsymbol{\Phi}_r^{q,\boldsymbol{q}_X})$, where $\boldsymbol{\Phi}_i^{q,\boldsymbol{q}_X} = (\boldsymbol{\Phi}_{i,1}^{q,q_{X,1}}, \ldots, \boldsymbol{\Phi}_{i,p_X}^{q,q_{X,p_X}})$ are the coefficient functions of covariates. Specifying the $\mathrm{CL}(\boldsymbol{Y} \mid \mathcal{M})$ term requires the log-likelihood function of the TVARX model. Analogous to (4.8), the log-likelihood function is

$$\begin{aligned} & L_n(\boldsymbol{\Psi}^q, \boldsymbol{\Phi}^{q,\boldsymbol{q}_X}, r, d, \boldsymbol{\theta}, q, \boldsymbol{p}_Y, \boldsymbol{q}_X, \boldsymbol{p}_X) \\ = & -\frac{1}{2}\sum_{k=1}^{n}\sum_{i=1}^{r}\left[q_{\max}\log 2\pi + \log|\Sigma_{W,i}^{q_{\max}}| + \widetilde{\boldsymbol{Y}}_k^{\mathrm{T}}\Sigma_{W,i}^{q}{}^{-1}\widetilde{\boldsymbol{Y}}_k + \sum_{m=q+1}^{q_{\max}}\frac{y_{k,m}^2}{\sigma_{m,i}^2}\right] \\ & \times I(z_{t-d} \in (\theta_{i-1}, \theta_i]), \end{aligned} \tag{S.41}$$

where $\boldsymbol{\Psi}_i^q = (\boldsymbol{\Psi}_{i,1}^q, \ldots, \boldsymbol{\Psi}_{i,p_{Y,i}}^q)$, $\boldsymbol{\Phi}_i^{q,\boldsymbol{q}_X} = (\boldsymbol{\Psi}_{i,1}^{q,q_{X,1}}, \ldots, \boldsymbol{\Phi}_{i,p_X}^{q,q_{X,p_x}})$ and $\widetilde{\boldsymbol{Y}}_k = \boldsymbol{Y}_{k,q} - \boldsymbol{a}_i^q - \sum_{j=1}^{p_{Y,i}}\boldsymbol{\Psi}_{i,j}^q\boldsymbol{Y}_{k-j,q} - \sum_{m=1}^{p_X}\boldsymbol{\Phi}_{i,m}^{q,q_{X,m}}\boldsymbol{X}_{k,m}$.

Using the log-likelihood function in (S.41), the $\mathrm{CL}(\boldsymbol{Y} \mid \mathcal{M})$ term is specified as

$$\mathrm{CL}(\boldsymbol{Y} \mid \mathcal{M}) = -L_n(\hat{\boldsymbol{\Psi}}^q, \hat{\boldsymbol{\Phi}}^{q,\boldsymbol{q}_X}; \boldsymbol{Y}_{k,q}, \ldots, \boldsymbol{Y}_{k-p_{Y,i},q}, \boldsymbol{X}_{k,1}, \ldots, \boldsymbol{X}_{k,p_X})\log_2 e\,, \tag{S.42}$$

where the estimators $\hat{\boldsymbol{\Psi}}^q, \hat{\boldsymbol{\Phi}}^{q,\boldsymbol{q}_X}$ maximize $L_n(\boldsymbol{\Psi}^q, \boldsymbol{\Phi}^{q,\boldsymbol{q}_X}, r, d, \boldsymbol{\theta}, q, \boldsymbol{q}_X, \boldsymbol{p}_X)$ given a set of model orders $\{r, d, \boldsymbol{\theta}, q, \boldsymbol{p}_Y, \boldsymbol{q}_X, \boldsymbol{p}_X\}$. For the $\mathrm{CL}(\mathcal{M})$ term, compared to (4.11), we also need to encode $\boldsymbol{p}_{X,i}$ and the coefficient matrix

$\mathbf{\Phi}_{i,m}^{q,\boldsymbol{q}_X}$. Since the length p_X of $\boldsymbol{p}_{X,i}$ is common across all models, we only need to consider $\mathbf{\Phi}_{i,m}^{q,q_X,m}$ and define

$$\begin{aligned}\mathrm{CL}(\mathcal{M}) &= \log_2 rd + \frac{1}{2}\sum_{i=1}^{r-1}\log_2 n_i + \sum_{i=1}^{r}\log_2[(p_{Y,i}q + 1 + \boldsymbol{q}_X^{\mathrm{T}}\boldsymbol{p}_{X,i})q] \\ &\quad + \sum_{i=1}^{r}\frac{(p_{Y,i}q + 1 + \boldsymbol{q}_X^{\mathrm{T}}\boldsymbol{p}_{X,i})q}{2}\log_2 n_i\,. \qquad \text{(S.43)}\end{aligned}$$

In summary, combining equations (S.42) and (S.43), the fMDL criterion of a functional TARX model $\mathcal{M}$ is defined as

$$\mathrm{fMDL}(r, d, \boldsymbol{\theta}, q, \boldsymbol{q}_X, \boldsymbol{p}_X) = \mathrm{CL}(\mathcal{M}) + \mathrm{CL}(\boldsymbol{Y} \mid \mathcal{M})\,. \qquad \text{(S.44)}$$

Finally, the optimal functional TARX model can be fitted by $\{\hat{r}_n, \hat{d}_n, \hat{\boldsymbol{\theta}}_n, \hat{q}_n, \hat{\boldsymbol{p}}_Y, \hat{\boldsymbol{q}}_X, \hat{\boldsymbol{p}}_X\}$ which minimizes (S.44).

D: Optimization via the genetic algorithm

The optimization of fMDL in (4.13) is non-standard since the objective function fMDL is not differentiable with respective to the parameters $\{r, d, \boldsymbol{\theta}, q, \boldsymbol{q}_X, \boldsymbol{p}_X\}$. Moreover, note that the likelihood function remains constant on $\theta_i \in (R_j, R_{j+1}]$, $j = 1, \ldots, n-1$, where $\{R_1, \ldots, R_n\}$ are the sorted values of threshold variable z_{k-d} in ascending order. Therefore, the estimated thresholds in the optimization problem can be chosen among $\{R_j\}_{j=1,\ldots,n-1}$. Even so, it still requires a huge combinatorial search. To tackle this problem, we develop

a genetic algorithm (GA) which is found promising in detecting change-points (See Davis et al. (2006), Lu et al. (2010)) and thresholds (See Yau et al. (2015)). We illustrate the implementation of a GA with respect to estimating a fTAR model, where the estimation of a fTARX model can be conducted similarly.

The GA begins with a population of *chromosomes*, in which each chromosome represents a model. The performance of every chromosome is evaluated by the fMDL value of the corresponding model. The better performing chromosomes have higher probabilities of reproducing *offsprings* by *crossover* with another chromosomes so that good features are more likely to be inherited. Alternatively, a parent chromosome may be selected to perform *mutation* with a small probability, which introduces new solutions to seek for the global optimum. For sufficiently many rounds of crossover and mutation, the model corresponding to the best performing chromosome is regarded as the solution to the fMDL optimization problem.

The details of the GA for the fTAR model is given as follows. The algorithm is conducted for a fixed d. The estimated value of d can be obtained by repeating the algorithm for several values of d. First, we generate the initial population:

1) Generate r from a Poisson distribution with mean 2;

2) Sequentially sample $\{\theta_1, \ldots, \theta_{r-1}\}$ uniformly from $\{z_k\}$, and reject and sample again whenever one of the regimes violate the *minimum span condition*, that is, a regime has fewer than $\tau_{n,0}$ observations. This condition is imposed to maintain the estimation accuracy.

3) Sample q uniformly from $1, \ldots, q_{\max}$. For $i = 1, \ldots, r$, generate the autoregressive order $p_{Y,i}$ uniformly from $\{0, 1, ..., p_{\max}\}$, with some upper bound $p_{\max}$.

The chromosome $\boldsymbol{c}$ is stored in the form $\boldsymbol{c} = \{r, p_{Y,1}, (\theta_1, p_{Y,2}), \ldots, (\theta_{r-1}, p_{Y,r}), q\}$, and completely specifies a fTAR model. Then, fMDL for $\boldsymbol{c}$ is calculated by (4.13). The chromosome generation is repeated $N_I \times N_p$ times to obtain N_I distinct subpopulations called *islands* with size N_p each.

Next, in each island, the offsprings are reproduced by the following procedure:

Crossover and Mutation. Crossover and mutation are two alternative methods for generating offsprings. For crossover, two *parent* chromosomes with probability proportional to the reciprocal of the rank of fMDL are selected. First, each $p_{Y,1}$ and q are selected from the two parents with equal probability. Next, every $\{R_i, p_{Y,i+1}\}$ pair in both parents is selected with probability 0.5. Then, the selected pairs $\{R_i, p_{Y,i+1}\}$s are sorted in ascending order to reproduce an offspring $\{r^o, p^o_{Y,1}, (\theta^o_1, p^o_{Y,2}), \ldots, (\theta^o_{r-1}, p^o_{Y,r}), q^o\}$. If

the minimal span condition is violated, delete the $\{\theta^o_{i-1}, p^o_{Y,i}\}$ pair until the condition is fulfilled.

The mutation method shares the same procedure of reproducing an offspring in crossover; however, only one parent is selected from the current population, while the other one is simulated as if in the initial population. The reproduction of a new chromosome follows the same steps in crossover, with the probability of selection from the generated parent being 0.7 for a higher degree of evolution. In order to keep a balance between crossover and mutation, and hence keep a balance between retaining existing solutions and searching for the globally optimal solution, the probability of conducting crossover and mutation are 0.9 and 0.1, respectively. The reproduction are conducted N_p times so that the population in each island remains unchanged.

After generating the offspring generation, the *elitist* step is introduced where the 20 worst performing chromosomes are replaced by the best 20 performing chromosomes in the parent generation. In addition, to explore more possibilities of model orders, one of the order parameter in a randomly selected regime is replaced by a newly generated one for every offspring with probability 0.3.

Migration. After every M_i rounds of iteration, the M_N worst performing

chromosomes in island j is replaced by M_N best performing chromosomes in island $j+1$, where $j = 0, \ldots, I-1$ and the $0th$ island is conventionally defined as the N_Ith island. The migration procedure can accelerate computation by parallel computing, and alleviate the problem of trapping in a sub-optimum solution (Alba and Troya (2002)).

Claim of Convergence. When the best performing chromosome remains unchanged in consecutively 20 rounds, we claim convergence and obtain the optimal model from the parameters in the best performing chromosome.

Empirical findings suggest that the parameters $\tau_{n,0}$=$0.1n$, $p_{max} = 3$, $N_i = 50, N_p = 200, M_i = 4, M_N = 2$ generally give satisfactory optimization performance. These parameters are used in the simulation studies in Section E.

E: Simulation Studies

E.1: Estimation of fTAR models

In this section we conduct simulation experiments to investigate the finite sample performance of the proposed method. Consider a three-regime fTAR model (2.1) with $\{\theta_1, \theta_2\} = \{-0.2, 0.6\}$, $\{p_1, p_2, p_3\} = \{2, 1, 2\}$, and $z_{k-2} = \sum_{t=0}^{100} Y_{k-2}(t/100)/101$. The orthonormal basis functions are Fourier basis

on $[0,1]$. Recall that the true autoregressive model parameters can be represented as $a_i = \sum_{l=1}^{3} a_l^{(i)} u_l$ and $\Psi_{i,j} = \sum_{l=1}^{3}\sum_{l'=1}^{3} a_{l,l'}^{(i,j)} u_l \otimes u_{l'}$ for $i = 1,2,3$ and $j = 1,\ldots,p_{Y,i}$. Hence, the linear operators needed to simulate the functional time series of interest can be represented by a 3×3 dimensional matrix. The corresponding innovations are generated according to $\epsilon_k(t) = \sum_{l=1}^{3} A_k u_l(t)$, where A_k are i.i.d. standard normal random variables. The variance curves are specified as constants over $[0,1]$, that is $\sigma_1 = 0.85$, $\sigma_2 = 0.9$ and $\sigma_3 = 0.7$. The coefficients $\boldsymbol{a}_i = (a_1^{(i)}, a_2^{(i)}, a_3^{(i)})^{\mathrm{T}}$ and $\boldsymbol{\Psi}_{i,j} = (a_{l,l'}^{(i,j)})_{1\le l,l'\le 3}$ are taken as

$$\boldsymbol{a}_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \boldsymbol{\Psi}_{1,1} = \begin{pmatrix} 0.15 & 0 & 0 \\ 0 & 0.15 & 0 \\ 0 & 0 & 0.15 \end{pmatrix}, \boldsymbol{\Psi}_{1,2} = \begin{pmatrix} 0 & 0 & 0.75 \\ 0 & 0.8 & 0 \\ 0.8 & 0 & 0 \end{pmatrix},$$

$$\boldsymbol{a}_2 = \begin{pmatrix} 0.15 \\ 0 \\ 0 \end{pmatrix}, \boldsymbol{\Psi}_{2,1} = \begin{pmatrix} 0.3 & 0 & 0 \\ 0 & 1.1 & 0 \\ 0 & 0 & 0.9 \end{pmatrix},$$

$$\boldsymbol{a}_3 = \begin{pmatrix} 0.3 \\ 0 \\ 0 \end{pmatrix}, \boldsymbol{\Psi}_{3,1} = \begin{pmatrix} 0 & 0 & 0.15 \\ 0 & 0.15 & 0 \\ 0.15 & 0 & 0 \end{pmatrix}, \boldsymbol{\Psi}_{3,2} = \begin{pmatrix} 0 & 0 & 0.8 \\ 0.15 & 0.65 & 0 \\ 0.65 & 0.15 & 0 \end{pmatrix}.$$

If we assume that the basis is unknown, the FPCA method is used to

obtain the basis functions. In every simulation experiment, 200 replications are conducted, and sample sizes of 300, 500, and 1000 are explored. The empirical classification rates of the number of regimes, model orders in each regime, delay parameter value and dimension of components are presented in Table 1. The results for treating the basis as known or unknown are compared. It can be seen that high correct classification rates of the number of regimes as well as model orders are observed in all experiments. No matter known or unknown basis functions are assumed, $d = 2$ and $q = 3$ are selected in nearly all cases. Also, the estimation accuracy is satisfactory for moderate and large sample sizes, and there is a significant improvement as the sample size increases.

Next, we report the estimation results of thresholds and variances of noises based on the replications with a correctly identified threshold model structure in Table 2. Table 2 demonstrates the consistency of threshold estimations under the fMDL criterion. In addition, the fast $O_p(n^{-1})$ convergence rates for threshold estimates and the accurate estimates of σ_i in every regime are realized.

Table 1: Percentages of the correct number of estimated thresholds, the correct order structure specifications, the correct value of delay parameter d and the correct dimension selection ($q = 3$) based on MDL.

$\{u_l\}$	Known				Unknown			
n	Regime (%)	Order (%)	delay parameter (%)	Dimension (%)	Regime (%)	Order (%)	delay parameter (%)	Dimension (%)
300	97.0	97.5	100.0	98.0	99.5	98.0	99.0	99.0
500	99.5	99.5	100.0	99.5	100.0	100.0	100.0	100.0
1000	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

E.2 Optimization performance of the genetic algorithm

In this subsection, we conduct two simulation studies to evaluate the optimization performance of the genetic algorithm. The first study aims to demonstrate that the global optimizer can often be achieved by the genetic algorithm. Following the setting in Section E.1, we randomly select one realization with $n = 500$ and conduct the genetic algorithm. The estimated threshold set for this realization is $(\hat{\theta}_1, \hat{\theta}_2) = (-0.211, 0.590)$. Next, we conduct an extensive grid search to find the global optimal thresholds of the fMDL criterion for this realization. As the true set of thresholds is (-0.2, 0.6), it seems sensible to focus on the search region $[-1, 1] \times [-1, 1]$ with

Table 2: Sample means of estimated thresholds θ_i's and σ_i's, with empirical standard errors in parentheses. Threshold estimations are based on replications with correctly estimated regime structures, and σ_i's estimations are based on replications with correctly estimated model structure.

$\{u_l\}$	Known					Unknown				
n	θ_1	θ_2	σ_1	σ_2	σ_3	θ_1	θ_2	σ_1	σ_2	σ_3
300	-0.214	0.586	0.830	0.874	0.674	-0.213	0.589	0.825	0.877	0.681
	(0.042)	(0.026)	(0.058)	(0.084)	(0.047)	(0.029)	(0.031)	(0.053)	(0.084)	(0.048)
500	-0.208	0.593	0.832	0.892	0.680	-0.204	0.592	0.837	0.881	0.684
	(0.023)	(0.012)	(0.045)	(0.067)	(0.037)	(0.018)	(0.014)	(0.044)	(0.068)	(0.038)
1000	-0.204	0.597	0.842	0.889	0.693	-0.203	0.596	0.842	0.891	0.692
	(0.010)	(0.006)	(0.030)	(0.043)	(0.027)	(0.010)	(0.007)	(0.032)	(0.045)	(0.024)
True	-0.2	0.6	0.85	0.9	0.7	-0.2	0.6	0.85	0.9	0.7

grids of widths 0.01. The optimal set of thresholds obtained from the grid search is $(-0.21, 0.60)$ which is almost the same as the one obtained by the genetic algorithm. We repeat the above procedures for other randomly selected realizations and compute the mean absolute error (MAE) between the estimate thresholds and the global ones. For example, in the above realization, MAE $= (|-0.21+0.211|+|0.60-0.59|)/2 = 0.0055$. The MAE for 100 randomly selected realizations are depicted in Figure 1. Observe that the MAEs are almost zero for all the realizations. This suggests that the estimated thresholds from the genetic algorithm is very close to the

global optimizers of the fMDL criterion.

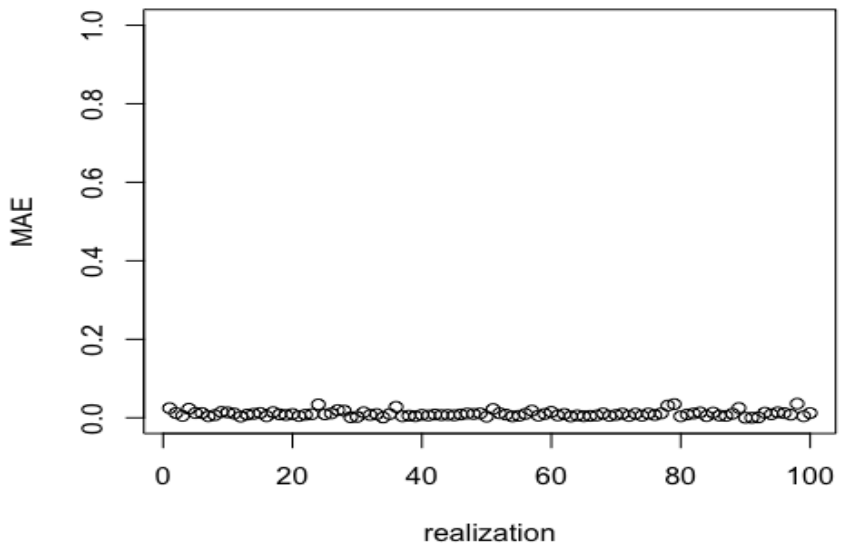

Figure 1: Mean absolute error (MAE) between the estimated thresholds from the genetic algorithm and the global optimizer of the fMDL function from grid search.

The second study aims to demonstrate the stability of the genetic algorithm. Following the setting in Section E.1, we randomly select one realization with $n = 500$ and apply the genetic algorithm for 100 times. The resulting fMDL and the estimated thresholds are summarized in Table 3. Observe that the standard errors of fMDL and estimated thresholds θ_i's are all zero. That is, the solutions of the genetic algorithm converge to the same optimizer for all the 100 simulations. Moreover, we repeat the above

Table 3: Sample means of fMDL and estimated thresholds obtained from the 100 trials of genetic algorithm on the same realization, with empirical standard errors in parentheses.

fMDL	θ_1	θ_2
2285 (0)	-0.248 (0)	0.592 (0)

procedure for 10 randomly selected realizations and summarize the results in Figure 2. From Figure 2, we can observe that the standard errors of the

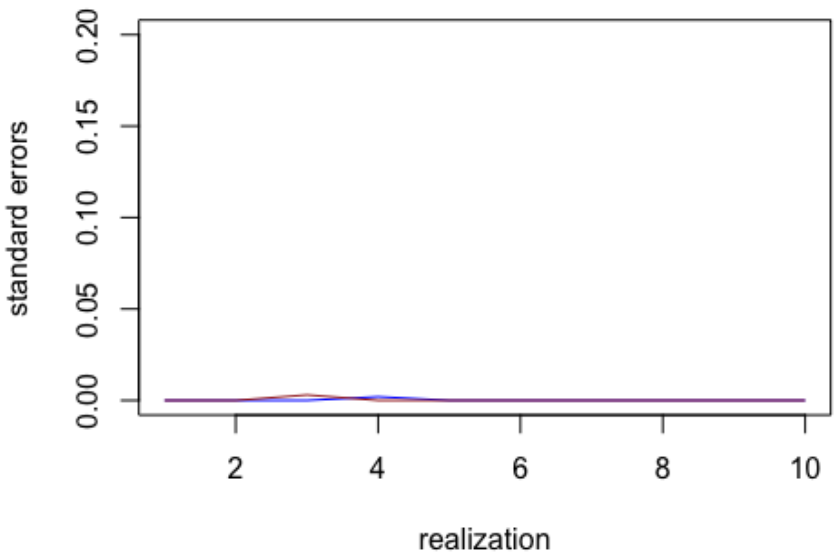

Figure 2: Standard deviations of estimated thresholds θ_1 (blue line) and θ_2 (brown line) over 100 trials of genetic algorithm in each of the 10 realizations.

estimated thresholds are nearly zero for all realizations. This justifies the

stability of the genetic algorithm.

F: Confidence intervals

To quantify the parameter estimation uncertainty, we discuss the construction of confidence intervals for the thresholds and other model parameters. First, we extend the algorithm in Li and Ling (2012) from univariate TAR models to fTAR models for computing the quantiles of $\mathbf{M}_-$ defined in Theorem 4. With these quantiles, confidence intervals for the threshold θ_j's can be constructed. Specifically, for the threshold θ_j, we simulate the two-sided compound Poisson process $\mathcal{P}_j(\kappa)$ in (5.14) on the interval $[-T, T]$ for a large enough $T > 0$. The trajectory of $\mathcal{P}_j(\kappa)$ for $\kappa \in [-T, T]$ is given by

$$\mathcal{P}_j(\kappa) = I(\kappa < 0) \sum_{k=1}^{N_{j,1}} I(U_k > \kappa)\xi_k^{(j+1,j)} + I(\kappa \geq 0) \sum_{l=1}^{N_{j,2}} I(V_l < \kappa)\xi_l^{(j,j+1)} ,$$

where U_k's and V_l's are independently and uniformly distributed on $[-T, 0]$ and $[0, T]$, respectively, $N_{j,1}$ and $N_{j,2}$ are two independent Poisson random variables with the same rate $\pi_z(\theta_j)T$. The jump-size sequences $\{\xi_k^{(j+1,j)}\}_{k=1,\ldots}$ and $\{\xi_l^{(j,j+1)}\}_{l=1,\ldots}$ can be generated from $F_{(j+1,j)}(\cdot|\theta_j)$ and $F_{(j,j+1)}(\cdot|\theta_j)$ by using Algorithm B of Li and Ling (2012) where the density function $\pi_z(\cdot|\mathbf{z_i})$ is replaced by the corresponding empirical density. Then, we take the smallest minimizer of $\mathcal{P}_j(\kappa)$ on $[-T, T]$ as one observed value of $M_-^{(j)}$. Repeating

this procedure gives a number of replicates of $M_-^{(j)}$, and hence the empirical quantiles can be computed.

To investigate the accuracy of the quantile estimation procedure, we conduct the following Monte Carlo simulation experiment. Following the setting in Section E.1, based on the thresholds estimates of 1000 replications, we compute the empirical quantiles of each of $n(\hat{\theta}_j - \theta_j)$ for $j = 1, \ldots, 2$. Specifically, the 5% and 95% empirical quantiles of $n(\hat{\theta}_1 - \theta_1)$ and $n(\hat{\theta}_2 - \theta_2)$ are $(-25.4, 14)$ and $(-21.2, 10.9)$, respectively. In comparison, the 5% and 95% quantiles obtained by the proposed algorithm are $(-26.0, 14.7)$ for $M_-^{(1)}$ and $(-21.0, 10.5)$ for $M_-^{(2)}$, which is very close to the empirical quantiles.

For the confidence intervals of other model parameters, it suffices to estimate the variance matrix of the limiting distribution in Theorem 5. Specifically, Γ_i can be estimated by $\sum_t^{(i)} [\Gamma_{i,t}\Gamma_{i,t}^{\mathrm{T}}]/n_i$, where $\sum_t^{(i)}$ denotes a sum over observations in regime i, $\Gamma_{i,t} = \mathrm{vec}(1, \boldsymbol{Y}_{t-1}, \ldots, \boldsymbol{Y}_{t-p_{Y,i}}, \boldsymbol{X}_{t,1}, \ldots, \boldsymbol{X}_{t,p_X})$, and n_i is the sample size of regime i. Also, $\Sigma_{W,i}$ is estimated by the sample covariance matrix of the residuals in regime i.

G: Estimation results and diagnostic plots for the real data

The estimation results of the functional TARX model are listed as follows:

$$
\begin{aligned}
\hat{\boldsymbol{a}}_1 &= \begin{pmatrix} 0.0741 \\ 0.2090 \\ 0.2377 \end{pmatrix}, \hat{\boldsymbol{\Psi}}_{1,1} = \begin{pmatrix} 0.9952 & 0.2316 & -0.3845 \\ -0.1668 & 1.7421 & -1.4777 \\ -0.2350 & 0.6404 & -0.4059 \end{pmatrix}, \\
\hat{\boldsymbol{\Psi}}_{1,2} &= \begin{pmatrix} 0.0124 & -0.2153 & 0.2446 \\ 0.2223 & 0.7016 & 0.7561 \\ 0.2970 & -0.5845 & 0.6216 \end{pmatrix}, \\
\hat{\boldsymbol{\Phi}}_{1,X_{csfb}} &= \begin{pmatrix} -0.0031 \\ -0.0105 \\ -0.0119 \end{pmatrix}, \hat{\boldsymbol{\Phi}}_{1,X_{gsli}} = \begin{pmatrix} 0.0008 \\ 0.0139 \\ 0.0143 \end{pmatrix}, \\
\hat{\boldsymbol{a}}_2 &= \begin{pmatrix} 0.0564 \\ 0.0969 \\ 0.1238 \end{pmatrix}, \hat{\boldsymbol{\Psi}}_{2,1} = \begin{pmatrix} 1.0191 & -0.0435 & -0.1161 \\ 0.3155 & 1.0623 & -1.0275 \\ 0.3545 & 0.1948 & -0.2023 \end{pmatrix}, \\
\hat{\boldsymbol{\Phi}}_{2,X_{csfb}} &= \begin{pmatrix} -0.0026 \\ -0.0004 \\ -0.0037 \end{pmatrix}, \hat{\boldsymbol{\Phi}}_{2,X_{vix}} = \begin{pmatrix} -0.0007 \\ -0.0158 \\ -0.0119 \end{pmatrix}, \\
\hat{\boldsymbol{a}}_3 &= \begin{pmatrix} 0.0035 \\ -0.0126 \\ 0.1129 \end{pmatrix}, \hat{\boldsymbol{\Psi}}_{3,1} = \begin{pmatrix} 0.9238 & -0.0093 & -0.0577 \\ -0.0763 & 1.0532 & -0.8582 \\ -0.1531 & 0.0776 & 0.01714 \end{pmatrix}, \hat{\boldsymbol{\Phi}}_{3,X_{csfb}} = \begin{pmatrix} 0.0062 \\ 0.0282 \\ 0.0240 \end{pmatrix}.
\end{aligned}
\tag{S.45}
$$

For model diagnostics, we plot the residuals of the intermediate-step

vector TARX model and vector ARX model in Figures 3 and 4, respectively.

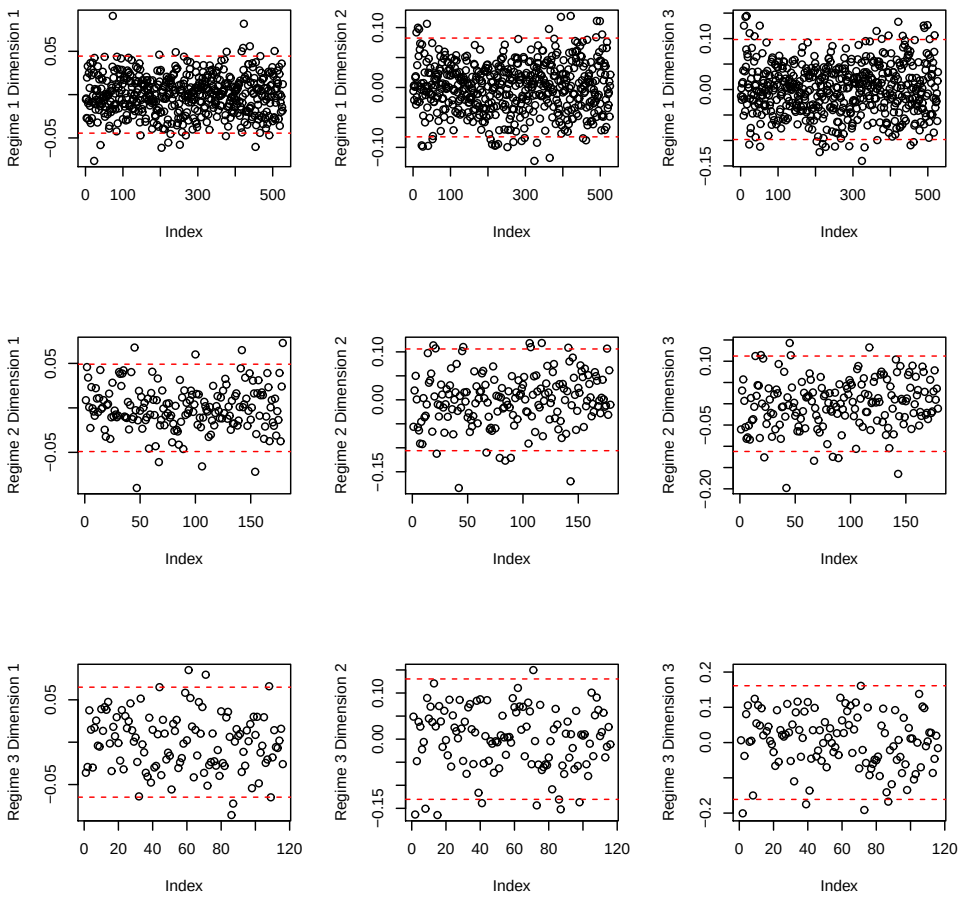

Figure 3: Residual plots from vector TARX modeling for the functional TARX model. Horizontal dashed line: two-sided 95% confidence interval for a white noise sequence of residuals.

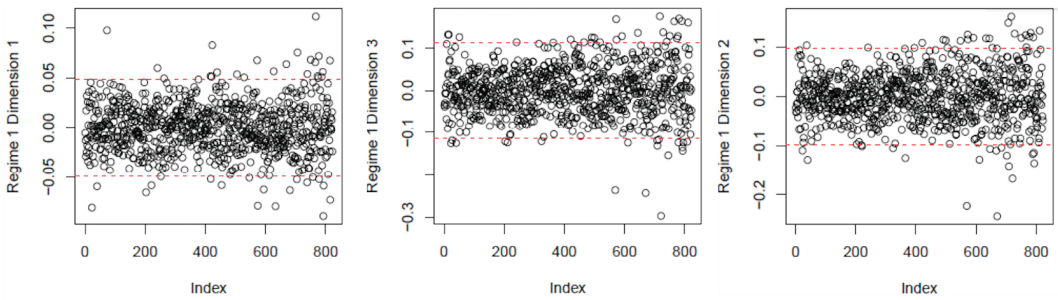

Figure 4: Residual plots from vector ARX modeling for the fARX model. Horizontal dashed line: two-sided 95% confidence interval for a white noise sequence of residuals.

H: Extension to infinite-dimensional parameter space

The estimation and model selection methods proposed in Section 4 require a finite number of basis functions generating a_i and $\Psi_{i,j}$; see Assumption 3. In this section, using a similar approach as in Cerovecki et al. (2019), we extend the proposed method to the general case where a_i and $\Psi_{i,j}$ are parametrised by infinite-dimensional parameters. For simplicity, we only consider the extension of the fTAR model. The extension of the fTARX model can be developed analogously.

Given an infinite orthonormal basis $\{u_1, u_2, \dots\}$, we assume that the true autoregressive model parameters can be represented as $a_i = \sum_{l=1}^{\infty} a_l^{(i)} u_l$ and $\Psi_{i,j} = \sum_{l',l=1}^{\infty} \langle x, u_l \rangle a_{l,l'}^{(i,j)} u_{l'}$ for $i = 1, \dots, r$ and $j = 1, \dots, p_{Y,i}$, where $a_l^{(i)} = \langle a_i, u_l \rangle$ and $a_{l,l'}^{(i,j)} = \langle \Psi_{i,j}(u_l), u_{l'} \rangle$. The $a_l^{(i)}$'s and $a_{l,l'}^{(i,j)}$'s can be collec-

tively regarded as an infinite-dimensional parameter $\boldsymbol{\beta}_i$ which parametrize a_i and $\Psi_{i,j}$. Following Cerovecki et al. (2019), we assume that $\boldsymbol{\beta}_i \in \Omega_\beta$, which is a compact subset of the set of square summable sequences l^2.

For the estimation of parameters, we propose a fMDL criterion

$$\text{fMDL} = \log_2(rd) + \sum_{i=1}^{r-1} \log_2 n_i + \sum_{i=1}^{r} \frac{(p_{Y,i}N+1)N}{2} \log_2 n_i$$
$$+ \sum_{i=1}^{r} \log_2(p_{Y,i}N+1)N + \sum_{i=1}^{r}\sum_{k=1}^{n} l_k I(z_{k-d} \in (\theta_{i-1}, \theta_i]) \log_2 e\,,$$

where $l_k = \sum_{l=1}^{N} \omega_l \left(y_{k,l} - a_l^{(i)} - \sum_{j=1}^{P_{Y,i}} \sum_{l'=1}^{N} a_{l,l'}^{(i,j)} y_{k-j,l'} \right)^2$, $(\omega_l)_{l\geq 1}$ is a positive and summable sequence of weights, and N is a positive integer. Intuitively, only the first N entires of $\{a_l^{(i)}\}_{l=1,\ldots,}$ and the upper-left $N \times N$ block of $\{a_{l,l'}^{(i,j)}\}_{l,l'=1,\ldots}$ are estimated. Hence, the parameter estimators can be defined as

$$\{\hat{\boldsymbol{\beta}}_{1,n}^N, \ldots, \hat{\boldsymbol{\beta}}_{r,n}^N, \hat{\boldsymbol{\theta}}_n^N, \hat{d}_n^N, \hat{r}_n^N, \hat{\boldsymbol{p}}_{Y,n}^N\} := \underset{\boldsymbol{\beta}_1 \in \Omega_\beta^N, \ldots, \boldsymbol{\beta}_r \in \Omega_\beta^N, \boldsymbol{\theta}, d, r, \boldsymbol{p}_Y}{\operatorname{argmin}} \text{fMDL}\,, \quad \text{(S.46)}$$

where $\Omega_\beta^N \subset \Omega_\beta$ is the subspace of all sequences with entries equaling zero except for except for $a_1^{(i)}, \ldots, a_N^{(i)}$ and $(a_{l,l'}^{(i,1)})_{1\leq l,l'\leq N}, \ldots, (a_{l,l'}^{(i,p_{Y,i})})_{1\leq l,l'\leq N}$.

When N increases, more entries in $\boldsymbol{\beta}_i$ are estimated in (S.46). Thus, it is expected that, when N diverges with the sample size, nearly all the information of $\boldsymbol{\beta}_i$ are captured by the estimation procedure (S.46). Indeed, following Proposition 3 of Cerovecki et al. (2019), it can be shown that the estimators in (S.46) is strongly consistent when N increases to infinity with

the sample size, under some additional mild conditions similar to their Assumptions A8 and A9. For the choice of the weight (ω_l), we refer to Assumption A9 in Cerovecki et al. (2019).

Bibliography

Alba, E. and Troya, J. M. (2002). Improving flexibility and efficiency by adding parallelism to genetic algorithms. *Statistics and Computing* **12**, 91–114.

Bosq, D. (2000). *Linear Processes in Function Spaces*. Springer-Verlag, New York.

Bradley, R. C. (2005). Basic Properties of Strong Mixing Conditions. A Survey and Some Open Questions. *Probabability Surveys* **2**, 107–144.

Cerovecki, C., Francq, C., Hörmann, S. and Zakoian, J. M. (2019). Functional GARCH models: The quasi-likelihood approach and its applications. *Journal of Econometrics* **209**, 353–375.

Chan, K. S. and Tong, H. (1985). On the use of the deterministic Lyapunov function for the ergodicity of stochastic difference equations. *Advances in Applied Probability* **17**, 666–678.

Chan, K. S. (1993). Consistency and limiting distribution of the least

squares estimator of a threshold autoregressive model. *The Annals of Statistiscs* **21**, 520–533.

Davis, R. A., Lee, T. C. M. and Rodriguez-Yam, G. A. (2006). Structural break estimation for nonstationary time series models. *Joural of the American Statistical Association* **101**, 223–239.

Davydov, Y. A. (1973). Mixing conditions for Markov chains. *Theory of Probability and its Applications* **18**, 312–328.

Doukhan, P. (1994). *Mixing: Properties and Examples*. Springer-Verlag, New York.

Hörmann, S. and Kokoszka, P. (2010). Weakly dependent functional data. *The Annals of Statistics* **38**, 1845–1884.

Hörmann, S. and Kokoszka, P. (2012). Functional time series. *Handbook of Statistics: Time Series Analysis–Methods and Applications* **110**, 157–186.

Ibragimov, I. A. and Has'minskii, R. Z. (1981). *Statistical Estimation: Asymptotic Theory*. Springer-Verlag, New York.

Kurz, T. G. (1975). Semigroups of conditioned shifts and approximation of Markov process. *Annals of Probability* **3**, 618–642.

Kushner, H. J. (1980). A martingale method for the convergence of a se-

quence of processes to a jump-diffusion process. *Z. Wahrscheinlichkeitstheorie verw. Gebiete* **53**, 207–219.

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

Lu, Q. Q., Lund, R. and Lee, T. C. M. (2010). An MDL approach to the climate segmentation problem. *The Annals of Applied Statistics* **4**, 299–319.

Lütkepohl, H. (2006). *New Introduction to Multiple Time Series Analysis*. Springer-Verlag, New York.

Meyn, S. P. and Tweedie, R. L. (1993). *Markov Chains and Stochastic Stability*. Springer-Verlag, London.

Riesz, F. and Sz.-Nagy, B. (1990). *Functional Analysis*. Dover, New York.

Saïdi, Y. and Zakoïan J.-M. (2006). Stationarity and Geometric Ergodicity of a Class of Nonlinear ARCH Models. *Annals of Applied Probabability* **4**, 2256–2271.

Samia, N. I. and Chan, K. S. (2011). Maximum likelihood estimation of a generalized threshold stochastic regression model. *Biometrika* **98**, 433–448.

Seijo, E. and Sen, B. (2011). A continuous mapping theorem for the smallest argmax functional. *Electronic Journal of Statistics* **5**, 421–439.

Stout, W. F. (1974). *Almost Sure Convergence*. Academic Press, New York.

Strook, D. W. and Varadhan, S. R. S. (1971). Diffusion process with boundary conditions. *Communications on Pure and Applied Mathematics* **24**, 147–225.

van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

Yau, C. Y., Tang, C. M. and Lee, T. C. M. (2015). Estimation of multiple-regime threshold autoregressive models with structural breaks. *Journal of the American Statistical Association* **110**, 1175–1186.