

## MULTIVARIATE REDUCED-RANK NONLINEAR TIME SERIES MODELING

Ming-Chung Li and Kung-Sik Chan

*The EMMES Corporation and The University of Iowa*

*Abstract:* Panels of nonlinear time series data are increasingly collected in scientific studies, and a fundamental problem is to study the common dynamic structures of such data. We propose a new model for exploring the common dynamic structure in multivariate nonlinear time series. The basic idea is that the panel of time series are driven by an underlying low-dimensional nonlinear principal component process that is modeled as some nonlinear function of the past lags of the time series. In particular, we consider in some detail the REduced-rank Threshold AutoRegressive (RETAR) model whose nonlinear principal component process is a piecewise linear vector-valued function of past lags of the panel of time series. We propose an estimation scheme for the RETAR model and derive the large sample properties of the estimator. We illustrate the RETAR model using a modern panel of eight Canada lynx series, and demonstrate a classification of lynx series that is broadly similar to that reported by Stenseth et al. (1999), who used a different approach.

*Key words and phrases:* Common dynamic structure, consistency, maximum likelihood estimation, nonlinear principal component, threshold autoregressive model, weighted least squares.

### 1. Introduction

Multivariate time series data can arise in several areas. In particular, panels of time series data are increasingly collected in various fields and often analyzed with multivariate time series techniques. A common procedure assumes series from each site, or each component series, as driven by a (possibly distinct) model from a common parametric family of nonlinear models with innovations being perhaps contemporaneously correlated, but without serial dependence. This approach was adopted by Stenseth et al. (1999) in modeling a panel of Canada lynx pelt data. They modeled each component lynx pelt series by a second-order Threshold AutoRegressive (TAR) model (Tong (1990)) with delay equal to two, as motivated by a predator-prey model. They studied the spatial variation in the nonlinear dynamics of the data which might be attributed to two competing hypotheses: an ecology-based classification according to forest types, or a climate-based classification according to weather pattern induced by the North Atlantic Oscillation (NAO). Stenseth et al. (1999) found that the lynx dynamics

could be classified according to three climatic regions, Pacific-maritime, Continental and Atlantic-maritime, over each of which the lynx series share similar dynamics during the decrease phase of the series.

An exhaustive search for patterns in the nonlinear dynamics of a panel of time series is generally infeasible. Here we propose a new approach for exploring common structure. The basic idea is that a panel of time series may be driven by a few latent (nonlinear) principal component processes (which we sometimes refer to as factor processes). In the case of linear Vector Auto-Regressive Moving-Average (VARMA) processes, this idea is most conveniently implemented via the reduced rank approach; see Reinsel and Velu (1998). Many nonlinear models, including the TAR model, are conditionally linear, i.e., given some (nonlinearity) parameters, the model resembles a linear 'regression' model with the regressors being some (nonlinear) functions of past lags of the observations and the nonlinearity parameters. Our new approach generalizes the method of reduced-rank regression to conditionally linear models. The new model is referred to as the REduced-rank Nonlinear AutoRegressive (RENAR) model. In particular, we specialize the RENAR model to the case of threshold model, resulting in the REduced-rank Threshold AutoRegressive (RETAR) model.

The organization of this paper is as follows. We introduce the RENAR approach and the RETAR model in Section 2. Statistical estimation of a RETAR model is discussed in Section 3. Some sufficient conditions for ergodicity and stationarity of a RETAR model are derived in Section 4. The estimation procedure proposed in Section 3 is essentially Maximum likelihood (ML) estimation assuming that the innovations are homogeneous and Gaussian. For the case of known error covariance, ML estimation can be carried out by weighted least squares (LS) estimation with the inverse of the error covariance being the weight. In Section 5, we derive the consistency property of the weighted LS estimator of a stationary ergodic RETAR model. In particular, the threshold estimator is shown to be super-consistent, i.e., of  $O_P(1/T)$  from the true value where  $T$  is the sample size. In Section 6, we derive the asymptotic distribution of the weighted least squares estimator with the weight being the inverse of the error covariance matrix, and we discuss how to extend this asymptotic result to the ML estimator proposed in Section 3. We illustrate the RETAR model with the modern panel of Canada lynx data in Section 7, which essentially confirms the climate-based classification reported by Stenseth et al. (1999). We conclude in Section 8. Proofs of all theorems can be found in the online supplement (<http://www3.stat.sinica.edu.tw/statistica>).

## 2. The Model

Let  $\{Y_{st}\}$ ,  $s = 1, \dots, m$ ,  $t = 1, \dots, T$  be a panel of time series data where  $s$  may denote a site, for example, and  $t$  denotes time. A general scheme for

modeling  $Y_t = (Y_{1t}, \dots, Y_{mt})$  is to write

$$Y_t = h(W_t) + \epsilon_t,$$

where  $h(\cdot)$  can be some parametric or non-parametric function, the random vector  $W_t$  may include past lags of  $Y_t$ , and the  $\epsilon_t$  denote the innovations; more generally,  $W_t$  may contain other covariates. See Tong (1990) for a review of nonlinear time series analysis. Here we consider the case of a conditionally linear model, that is,  $h(W_t) = BW_t(\delta)$  where  $W_t(\delta)$  is a function of past lags and/or covariates as well as a parameter vector  $\delta$ . This class of models is quite general as it includes the scheme of fitting a parametric model series by series with contemporaneously correlated innovations. To avoid the curse of dimensionality, we suppose that  $B$  is of reduced rank.

To fix ideas, we specialize the above scheme to TAR models. Recall that a two-regime first order TAR model, also denoted as TAR(2; 1, 1), takes the form

$$Y_{st} = \beta_{s0} + \beta_{s1}Y_{s,t-1} + \{\beta_{s2} + \beta_{s3}Y_{s,t-1}\}I(Y_{s,t-d} > \gamma_s) + \epsilon_{st} \quad (2.1)$$

$$= \beta_{s0} + \beta_{s1}X_{s1t} + \beta_{s2}X_{s2t} + \beta_{s3}X_{s3t} + \epsilon_{st}, \quad (2.2)$$

where  $X_{s1t} = Y_{s,t-1}$ ,  $X_{s2t} = I(Y_{s,t-d} > \gamma_s)$ ,  $X_{s3t} = Y_{s,t-1}I(Y_{s,t-d} > \gamma_s)$ ,  $\epsilon_{st} = \{\sigma_{s1} + \sigma_{s2}I(Y_{s,t-d} > \gamma_s)\}e_t$ , and  $\{e_t\}$  are i.i.d. with zero mean and unit variance. Note that, given  $\gamma_s$  and  $d$ , the TAR model is linear in the  $X$ 's. Equation (2.1) is a common form used to represent a TAR model. The advantage of (2.2) is that it allows us to treat the TAR model as a classical linear model. Note that in (2.2) the "covariates" also contain the unknown parameters  $d, \gamma_1, \dots, \gamma_m$ .

Consider the above system of TAR(2; 1, 1) models in matrix form:

$$\begin{aligned} Y_t = \begin{pmatrix} Y_{1t} \\ Y_{2t} \\ \vdots \\ Y_{mt} \end{pmatrix} &= \begin{pmatrix} \beta_{10} \\ \beta_{20} \\ \vdots \\ \beta_{m0} \end{pmatrix} + \begin{pmatrix} \beta_{11} & 0 & \dots & 0 \\ 0 & \beta_{21} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \beta_{m1} \end{pmatrix} \begin{pmatrix} Y_{1,t-1} \\ Y_{2,t-1} \\ \vdots \\ Y_{m,t-1} \end{pmatrix} \\ &+ \begin{pmatrix} \beta_{12} & 0 & \dots & 0 \\ 0 & \beta_{22} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \beta_{m2} \end{pmatrix} \begin{pmatrix} I(Y_{1,t-1} > \gamma_1) \\ I(Y_{2,t-1} > \gamma_2) \\ \vdots \\ I(Y_{m,t-1} > \gamma_m) \end{pmatrix} \\ &+ \begin{pmatrix} \beta_{13} & 0 & \dots & 0 \\ 0 & \beta_{23} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \beta_{m3} \end{pmatrix} \begin{pmatrix} Y_{1,t-1}I(Y_{1,t-1} > \gamma_1) \\ Y_{2,t-1}I(Y_{2,t-1} > \gamma_2) \\ \vdots \\ Y_{m,t-1}I(Y_{m,t-1} > \gamma_m) \end{pmatrix} \end{aligned}$$

$$+ \begin{pmatrix} \{\sigma_{11} + \sigma_{12}I(Y_{1,t-1} > \gamma_1)\}e_{1t} \\ \{\sigma_{21} + \sigma_{22}I(Y_{2,t-1} > \gamma_2)\}e_{2t} \\ \vdots \\ \{\sigma_{m1} + \sigma_{m2}I(Y_{m,t-1} > \gamma_m)\}e_{mt} \end{pmatrix}.$$

The diagonal coefficient matrices of the above model imply that there are no direct relationships between the  $Y$ 's from different sites. Such relationships may, however, be modeled by using nondiagonal coefficients, so that the model becomes

$$\begin{aligned} Y_t &= \mu + C_1 X_{1t} + C_2 X_{2t} + C_3 X_{3t} + \epsilon_t \\ &= \mu + C X_t + \epsilon_t, \end{aligned} \quad (2.3)$$

where  $\mu = (\beta_{10}, \dots, \beta_{m0})'$ ,  $C = (C_1, C_2, C_3)$ ,  $X_t = (X'_{1t}, X'_{2t}, X'_{3t})'$ , and there are  $m + 3 \times m^2$  coefficient parameters, a large number even for a moderate  $m$ . Note that the  $X$ 's depend on lags 1 and  $d$  of the  $Y$ 's, as well as the nonlinearity parameters  $\gamma$ 's and  $d$ 's.

A useful approach for reducing the number of explanatory variables in linear regression and vector autoregressive models is to invoke the *reduced-rank model*; see Robinson (1973), Izenman (1980), Velu, Reinsel and Wichern (1986), or Reinsel and Velu (1998). Here we borrow the essential idea of reduced-rank regression to reduce the number of unknown parameters by replacing the matrix  $C$  in the full rank model (2.3) by the product of two smaller-rank matrices  $A$  and  $B$  (of equal rank):

$$Y_t = \mu + C X_t + \epsilon_t = \mu + A B X_t + \epsilon_t,$$

where  $Y_t, \mu, \epsilon_t$  are  $m$  dimensional vectors,  $A$  is an  $m \times r$  full rank matrix, and  $B$  is a  $r \times 3m$  full rank matrix. The reduced-rank model bears a resemblance to factor analysis, and it also subsumes the case of principal component analysis and canonical correlation analysis; see Reinsel and Velu (1998). It has the advantage of explicitly modeling the underlying principal component as  $B X_t$ , thereby reducing the  $3m$ -dimensional covariate process  $X_t$  to an  $r$ -dimensional factor process  $B_j X_t$ , where  $B = (B'_1, \dots, B'_r)'$ . Here  $r \leq m$ , but the interesting case of  $r < m$  implies  $C$  is of reduced-rank.

The preceding discussion leads us to propose the *REduced-rank Nonlinear AutoRegressive* (RENAR) model

$$Y_t = \mu + A B X_t(\theta) + \epsilon_t, \quad t = 1, 2, \dots, \quad (2.4)$$

where the  $Y_t$  are  $m$ -dimensional and the  $X_t(\theta)$  are  $n$ -dimensional vector covariates which depend on a vector parameter  $\theta$ , and may consist of past lags of  $Y$ ;  $A$  and  $B$  are, respectively,  $m \times r$  and  $r \times n$  full-rank coefficient matrices;  $\epsilon_t \sim (0, \Sigma_{\epsilon\epsilon} = \Sigma)$  and independent of past  $X$ 's and  $Y$ 's. Some examples of the RENAR model

include EXPAR, STAR and piecewise polynomial models; see Tong (1990). Here, motivated by the threshold autoregressive model, we focus on the case that the  $X$ 's are piecewise linear functions of past lags of  $Y$ 's, although the general idea is applicable to other conditionally linear processes.

The decomposition of  $C = AB$  in (2.4) is not unique because  $C = AP^{-1}PB$  for any nonsingular matrix  $P$ , i.e,  $A$  and  $B$  can be "rotated" arbitrarily. Hence to determine  $A$  and  $B$  uniquely, we must impose some normalization conditions, to be elaborated upon below.

For the case of known  $\theta$  and known rank of  $A$ , the remaining unknown parameters can be estimated by the weighted least square estimator minimizing the objective function (with a positive definite weight matrix  $\Gamma$ )

$$L = tr \left[ \sum_{t=1}^T \Gamma^{\frac{1}{2}} [Y_t - \mu - ABX_t(\theta)] [Y_t - \mu - ABX_t(\theta)]' \Gamma^{\frac{1}{2}} \right]. \quad (2.5)$$

This has the closed-form solution (Reinsel and Velu (1998))

$$A = \Gamma^{-\frac{1}{2}} [V_1, \dots, V_r] = \Gamma^{-\frac{1}{2}} V, \quad B = V' \Gamma^{\frac{1}{2}} S_{yx} S_{xx}^{-1},$$

where  $S_{yx}$  is the sample covariance matrix of  $Y_t$  and  $X_t(\theta)$ , and  $S_{xx}$  is the sample variance matrix of  $X_t(\theta)$ ;  $V_j$  is the normalized eigenvector corresponding to the  $j$ th largest eigenvalue  $\lambda_j^2$  of  $\Gamma^{1/2} S_{yx} S_{xx}^{-1} S_{xy} \Gamma^{1/2}$ ,  $j = 1, 2, \dots, r$ .

**Remarks**

1. The normalization conditions for the decomposition of  $C = AB$  require the eigenvectors  $V_j$ 's to satisfy  $V_j' V_j = 1$  and  $V_i' V_j = 0$  if  $i \neq j$  or, equivalently, to require  $A$  and  $B$  to satisfy the conditions

$$BS_{xx}B' = \Lambda^2, \quad A' \Gamma A = I_r.$$

Owing to the  $r^2$  restrictions, there are  $mr + nr - r^2$  free independent parameters in  $A$  and  $B$ .

2. The estimator  $C = C^{(r)}$  is given by

$$\begin{aligned} C^{(r)} &= A^{(r)} B^{(r)} \\ &= \Gamma^{-\frac{1}{2}} \left( \sum_{j=1}^r V_j V_j' \right) \Gamma^{\frac{1}{2}} S_{yx} S_{xx}^{-1} = P_{\Gamma} S_{yx} S_{xx}^{-1}. \end{aligned}$$

Note that  $P_{\Gamma}$  is an idempotent matrix for any  $\Gamma$  but need not be symmetric. When  $r = m$ ,  $\sum_{j=1}^m V_j V_j' = I_m$  and  $C^{(m)}$  becomes a full-rank  $m \times n$  matrix.

Consider the RENAR model with threshold-type X-components, in which case the model is called the *REduced-rank Threshold AutoRegressive* (RETAR)

model. For the case of two regimes with AR orders  $p_1$  and  $p_2$ , and delay parameter  $d$ , the RETAR model postulates that

$$\begin{aligned}
Y_t &= \mu + C_1 Y_{t-1} + C_2 Y_{t-2} + \cdots + C_{p_1} Y_{t-p_1} + C_{p_1+1} I(Y_{t-d} > \gamma) \\
&\quad + C_{p_1+1+1} Y_{t-1} I(Y_{t-d} > \gamma) + \cdots + C_{p_1+1+p_2} Y_{t-p_2} I(Y_{t-d} > \gamma) + \epsilon_t \\
&= \mu + C_1 X_{1t} + C_2 X_{2t} + \cdots + C_k X_{kt} + \epsilon_t \\
&= \mu + C X_t + \epsilon_t \\
&= \mu + A B X_t + \epsilon_t,
\end{aligned} \tag{2.6}$$

where  $Y_t, \mu, X_{jt}$  are  $m$ -vectors,  $k = p_1 + 1 + p_2$ ,  $X_t$  is a  $km$ -dimensional vector, and  $C_j, C, A, B$  are  $m \times m, m \times km, m \times r$  and  $r \times km$  dimensional matrices, respectively. The notations  $I(Y_{t-d} > \gamma)$  and  $Y_{t-j} I(Y_{t-d} > \gamma)$  are defined componentwise, i.e.,

$$I(Y_{t-d} > \gamma) = (I(Y_{1,t-d} > \gamma_1), \dots, I(Y_{m,t-d} > \gamma_m))', \tag{2.7}$$

$$Y_{t-j} I(Y_{t-d} > \gamma) = (Y_{1,t-j} I(Y_{1,t-d} > \gamma_1), \dots, Y_{m,t-j} I(Y_{m,t-d} > \gamma_m))'. \tag{2.8}$$

The innovations  $\{\epsilon_t\}$  are assumed i.i.d. with zero mean and covariance matrix  $\Sigma$ . The delay parameter is assumed to be identical for all series. Clearly, the model can be generalized to the case of non-constant delay with the delay for the  $j$ th series being  $d_j$ .

### 3. Estimation of the RETAR Model

Consider the estimation of the nonlinearity parameters  $(d, \gamma, p_1, p_2)$ . The vector *delay* parameter  $d = (d_1, \dots, d_m)'$  is an  $m$ -dimensional vector with components in  $\{1, \dots, D\}$ ;  $D$  some known upper bound. Initial estimates of  $d$  and  $\gamma = (\gamma_1, \dots, \gamma_m)'$  can be obtained by fitting separate TAR models to each series. These initial estimates are refined via an iterative scheme to be described below. The orders  $(p_1, p_2)$  can be selected to be the maximum of the estimate obtained from series by series estimation. The order of a nonlinear autoregressive process can be consistently estimated by cross-validation; see Cheng and Tong (1992) and Chan and Tong (2001). Alternatively, these orders can be estimated by some classical model selection criterion. In practice substantive knowledge may supplement these objective criteria in selecting the order and/or delay parameter.

Henceforth, we write  $X_t$  for  $X_t(\theta)$ , for simplicity. Given the nonlinear parameters  $(d, \gamma, p_1, p_2)$ , we can apply reduced-rank regression techniques to estimate  $\mu, A$  and  $B$ . We propose the following algorithm to estimate the RETAR model (the estimator so obtained is the ML estimator for the case of known rank for  $A$  and  $B$ , assuming homogeneous Gaussian errors):

- Step 1: Obtain an initial estimate of  $\{d_s, s = 1, \dots, m\}$  and  $\{\gamma_s, s = 1, \dots, m\}$  by fitting separate TAR models to each series using, for example, the method of conditional least squares.
- Step 2: The corrected Bartlett test statistics (likelihood ratio tests) are computed in order to determine the rank of  $A$  and  $B$ . Specifically, we set the rank to be the smallest  $r$  for which we do not reject the null hypothesis  $H_0 : \text{rank} = r$  v.s.  $H_1 : C$  is of full rank. For the preceding test, the corrected Bartlett statistic is

$$\mathcal{M} = -\left[T - n + \frac{(n - m - 1)}{2}\right] \sum_{j=r+1}^m \log(1 - \hat{\rho}_j^2),$$

where  $\hat{\rho}_j$  is the  $j$ th largest sample canonical correlation between  $Y$  and  $X$ .  $\mathcal{M}$  is asymptotically  $\chi^2$  with d.f.  $= (m - r)(n - r)$  under  $H_0$ ;  $m$  is the dimension of  $Y_t$  and  $n$  that of  $X_t$ . Apply the reduced-rank regression algorithm to obtain the least squares estimator of  $\mu, A$  and  $B$ . As detailed in Reinsel and Velu (1998), the least squares estimators of  $\mu, A$  and  $B$  are the arguments which minimize the objective function

$$MD \equiv \text{tr} \left[ \sum_{t=1}^T \Gamma^{\frac{1}{2}} (Y_t - \mu - ABX_t)(Y_t - \mu - ABX_t)' \Gamma^{\frac{1}{2}} \right],$$

where  $\Gamma = \tilde{\Sigma}_{\epsilon\epsilon}^{-1} = T[(Y - \tilde{\mu} - \tilde{C}X)(Y - \tilde{\mu} - \tilde{C}X)']^{-1}$ ,  $X = (X_1, \dots, X_T)$ ,  $Y = (Y_1, \dots, Y_T)$ ,  $\tilde{C} = YX'(XX')^{-1}$  and  $\tilde{\mu} = \bar{Y} - \tilde{C}\bar{X}$ . In other words, the weight is the inverse of the error covariance matrix from the full-rank model.

- Step 3: Update the parameter estimates of  $(\gamma, d)$  by a grid search. Vary  $\gamma_1$  but with the other parameters of  $(\gamma, d)$  including the rank found in Step 2 being fixed, and compare the objective function ( $MD$ ) for different  $\gamma_1$ , say  $\gamma_{11}, \dots, \gamma_{15}$ . Update  $\gamma_1$  by the one with the smallest  $MD$ . Then we vary  $\gamma_2$  with the other parameters being fixed, and update it by the one with the smallest  $MD$ . Similarly we can update the estimates for  $\gamma_3, \dots, \gamma_m$  and  $d_1, \dots, d_m$ .
- Step 4: Repeat Steps 2 and 3 until the change in the objective function  $MD$  is less than a tolerance limit.

For a fixed  $\Gamma$  and known rank of  $A$ , the weighted least squares (WLS) estimator of  $\mu, A, B, \gamma$  and  $d$  obtained by minimizing the loss function (2.5) can be determined by the preceding algorithm, with  $\Gamma$  and  $r$  fixed. It will be shown that the weighted least squares estimators are consistent and the estimator of  $\gamma$  is super-consistent, that is, it has the convergence rate  $O_p(1/T)$  under some conditions. In particular, the above consistency property enables us to adapt the

techniques from Velu et al. (1986) to derive the limiting distribution of the WLS estimator  $(\hat{A}, \hat{B})$ ; see also Robinson (1973) and Reinsel and Velu (1998). These results can then be extended to show that the estimator obtained by the algorithm defined by Steps 1 to 4 enjoys the limiting properties of the WLS estimator with weight  $\Gamma = \Sigma_{\epsilon\epsilon}^{-1}$ .

We remark that in Step 2, an alternative approach is to determine the rank as a multiple-decision problem, see Anderson (1971, pp.270-276). We can also compute some information-based criteria, for example AIC or BIC, for comparison with the rank chosen by the likelihood ratio tests.

#### 4. Ergodicity and Stationarity of the RETAR Model

In this section, we study the ergodicity of the RETAR model. Recall that an ergodic Markov process is asymptotically stationary and admits a unique stationary distribution. Some early works on the ergodic theory of Markov chain are Foster (1953) and Tweedie (1975, 1976). For applications in non-linear time series, see Petrucci and Woolford (1984), Chan and Tong (1985), Chan, Petrucci, Tong and Woolford (1985), Tjøstheim (1990), Tong (1990), Chan (1993) and Cline and Pu (2001). For comprehensive reviews, see Nummelin (1984) and Meyn and Tweedie (1993).

In the one-dimensional case, Petrucci and Woolford (1984) found a necessary and sufficient condition for the ergodicity of a TAR(2; 1, 1) model. This turns out to be the condition for the stability of the skeleton obtained from suppressing the innovations in the TAR model; see Chan and Tong (1985) for further discussions of the link between stability and ergodicity. Here we are interested in applying this link between stability and ergodicity to find similar sufficient and necessary condition for the ergodicity of RETAR models. We start with the simplest case of model (2.6), with  $m = 2$ ,  $n = 6$ ,  $\mu = 0$ ,  $p_1 = p_2 = 1$ ,  $d = 1$  and  $r = 1$ . Without loss of generality, we can assume the threshold parameter  $\gamma = 0$ . Furthermore, the intercept term does not generally affect the ergodicity of the RETAR model, except for the boundary case; see Remark 1 below Theorem 1. Henceforth in this section, the intercept term is suppressed. Define

$$\begin{aligned} R_1 &= \{(x, y) : x > 0, y > 0\}, & R_2 &= \{(x, y) : x \leq 0, y > 0\}, \\ R_3 &= \{(x, y) : x \leq 0, y \leq 0\}, & R_4 &= \{(x, y) : x > 0, y \leq 0\}. \end{aligned}$$

Let  $Y_t = (Y_{1,t}, Y_{2,t})'$ . Because  $C$  is of unit rank, write  $C = \vec{a} \vec{b}'$ , where  $\vec{a} = (a_1, a_2)'$  and  $\vec{b} = (b_1, b_2, b_3, b_4, b_5, b_6)'$ .



The RETAR model can then be expressed as follows:

$$\begin{aligned}
 Y_t &= \vec{a}(b_1, b_2, b_3, b_4, b_5, b_6) \begin{pmatrix} Y_{1,t-1} \\ Y_{2,t-1} \\ I(Y_{1,t-1} > 0) \\ I(Y_{2,t-1} > 0) \\ Y_{1,t-1}I(Y_{1,t-1} > 0) \\ Y_{2,t-1}I(Y_{2,t-1} > 0) \end{pmatrix} + \epsilon_t \\
 &= \begin{cases} \vec{a}[(b_3 + b_4) + (b_1 + b_5, b_2 + b_6)Y_{t-1}] + \epsilon_t & \text{if } Y_{t-1} \in R_1 \\ \vec{a}[b_4 + (b_1, b_2 + b_6)Y_{t-1}] + \epsilon_t & \text{if } Y_{t-1} \in R_2 \\ \vec{a}[(b_1, b_2)Y_{t-1}] + \epsilon_t & \text{if } Y_{t-1} \in R_3 \\ \vec{a}[b_3 + (b_1 + b_5, b_2)Y_{t-1}] + \epsilon_t & \text{if } Y_{t-1} \in R_4. \end{cases} \quad (4.1)
 \end{aligned}$$

For notational convenience, we write

$$Y_t = \begin{cases} C_1 Y_{t-1} + \epsilon_t & \text{if } Y_{t-1} \in R_1 \\ C_2 Y_{t-1} + \epsilon_t & \text{if } Y_{t-1} \in R_2 \\ C_3 Y_{t-1} + \epsilon_t & \text{if } Y_{t-1} \in R_3 \\ C_4 Y_{t-1} + \epsilon_t & \text{if } Y_{t-1} \in R_4, \end{cases} \quad (4.2)$$

where the intercept terms are suppressed,  $C_i = \vec{a} \vec{b}'_i$  with  $\vec{a}, \vec{b}_i \in R^2$ , and  $\vec{b}'$  denotes the transpose of  $\vec{b}$ . (These  $C_i$  are different from those defined in model (2.6)). Note that the  $\vec{b}'_i$ s are defined in (4.1), e.g.,  $\vec{b}'_1 = (b_1 + b_5, b_2 + b_6)$ . The sample space  $R^2$  satisfies  $R^2 = \bigcup_{i=1}^4 R_i$ .

It is well known that if  $C_1 = C_2 = C_3 = C_4$ , then a necessary and sufficient condition for the ergodicity of  $\{Y_t\}$  is that  $\lambda(C_1) < 1$  where  $\lambda(A)$  denotes the spectral norm (the largest eigenvalue in absolute value) of a square matrix  $A$ . A natural conjecture is that a sufficient condition for the ergodicity of (4.2) is that  $\max_{1 \leq i \leq 4} \lambda(C_i) < 1$ . However, this need not be true as counter-examples exist where each subsystem is stable while the whole system is unstable; see Example 2.9 in Tong (1990).

The next theorem gives a set of sufficient conditions for the geometric ergodicity of the model (4.2). It can be proved by similar arguments as in Chan and Tong (1985).

**Theorem 1.** *Let  $\{Y_t\}$  be defined by (4.2). Suppose the error  $\epsilon_t$  has a density which is positive everywhere and  $E|\epsilon_t| < \infty$ . Let  $\vec{a}$  belong to the  $j$ th quadrant, i.e.,  $\vec{a} \in R_j$ , where  $j = j(\vec{a}) \in \{1, 2, 3, 4\}$ . A sufficient condition for the geometric ergodicity of the process  $\{Y_t\}$  is that*

$$\lambda(C_j) < 1, \quad \lambda(C_{j+2}) < 1, \quad \lambda(C_j C_{j+2}) < 1, \quad (4.3)$$

where the addition is defined modulus 4. Also, the skeleton obtained by suppressing the noise term from model (4.2) is stable if and only if the conditions (4.3) are satisfied.

### Remarks

1. If  $\lambda(C_j) > 1$  or  $\lambda(C_{j+2}) > 1$ , then the process  $\{Y_t\}$  is not ergodic. Also, if  $\lambda(C_j) < 0, \lambda(C_{j+2}) < 0$  and  $\lambda(C_j)\lambda(C_{j+2}) > 1$ , then the process  $\{X_t\}$  is not ergodic. The proofs of these claims are given in the online supplement (<http://www3.stat.sinica.edu.tw/statistica>). The ergodicity of the RETAR model for the boundary case in which  $\lambda(C_j) = 1, \lambda(C_{j+2}) = 1$  or  $\lambda(C_j C_{j+2}) = 1$  is still an open problem.
2. It is readily checked that  $\lambda(C_j) = \vec{b}'_j \vec{a}$  and  $\lambda(C_j C_{j+2}) = \vec{b}'_j \vec{a} \vec{b}'_{j+2} \vec{a}$ .
3. In the decomposition of  $C = \vec{a} \vec{b}'$ , we may attach a negative sign to  $\vec{a}$  and  $\vec{b}$  simultaneously. Although in the estimation step we need to impose a suitable restriction on the sign of  $\vec{a}$  and  $\vec{b}$  for identification purpose, it is readily seen that the conclusion for the above theorem is invariant with respect to the signs of  $\vec{a}$  and  $\vec{b}$ .
3. The key to deriving the preceding result is that the coefficient matrices  $C_i$  admit a common eigenvector  $\vec{a}$ . Such an eigenvector 'controls' the dynamical behavior so we can easily analyze this dynamical system. This is similar to the one-dimensional case where the scalar 1 is the common eigenvector of each coefficient. In particular, the preceding theorem can be readily generalized to the higher dimensional case with all coefficient matrices being of unit-rank.

For the general case with non-unit rank, we have the following result. Let  $\|A\| = (\sum_i \sum_j a_{ij}^2)^{1/2}$  be the Euclidean norm of a matrix  $A$ .

**Theorem 2.** *Suppose that the  $m$ -dimensional random vector  $\{Y_t\}$  satisfies the piecewise linear equations*

$$Y_t = c_i + \sum_{j=1}^p A_{ij} Y_{t-j} + e_t \quad \text{if } Y_{t-d} \in \Omega_i,$$

where  $\Omega_i, 1 \leq i \leq q$ , is a partition of  $R^m$ . If  $\max_i \sum_j \|A_{ij}\| < 1$  and each element of  $e_t$  possesses a first absolute moment, then  $\{Y_t\}$  is geometrically ergodic.

### Remarks

1. This result is a generalization of Lemma 3.1 of Chan and Tong (1985).
2. In place of the Euclidean norm, other matrix norms can be adopted. See the conditions of Theorem 5.6.2 in Graybill (1983).

## 5. Strong Consistency of the Weighted Least Square Estimators of A RETAR Model

We have defined the *REduced-rank Threshold AutoRegressive* (RETAR) model as

$$Y_t = \mu + ABX_t + \epsilon_t, \tag{5.1}$$

where  $Y_t = (Y_{1t}, \dots, Y_{mt})'$  and  $X_t = (Y'_{t-1}, \dots, Y'_{t-p}, I(Y_{t-d} > \gamma), (Y_{t-1}I(Y_{t-d} > \gamma))', \dots, (Y_{t-p}I(Y_{t-d} > \gamma))')'$ . The definition of  $I(Y_{t-d} > \gamma)$  and  $Y_{t-j}I(Y_{t-d} > \gamma)$  are given in (2.7) and (2.8). The matrices  $A$  and  $B$  are of rank  $r$  and must satisfy some normalization conditions to be stated below. The rank  $r$  can be identified by, for example, the Bartlett test. Thus we assume the rank  $r$  is fixed in this and the next sections. The innovations  $\epsilon_t$  are i.i.d. with zero mean, covariance matrix  $\Sigma_{\epsilon\epsilon}$  and assumed to be independent of the past  $Y_{t-1}, Y_{t-2}, \dots$ . Let  $n = \dim(X_t) = 2mp + m$ . A general parameter in  $\Omega = R^m \times R^{r(m+n-r)} \times R^m \times R^m$  is denoted by  $\theta = (\mu', (\text{vec}(A))', \text{vec}(B))', \gamma', d')'$ , and the true parameter is  $\theta_0 = (\mu'_0, (\text{vec}(A_0))', \text{vec}(B_0))', \gamma'_0, d'_0)'$ .

The weighted least squares estimator  $\hat{\theta}_T = (\hat{\mu}', (\text{vec}(\hat{A}))', \text{vec}(\hat{B}))', \hat{\gamma}', \hat{d}')'$  is any measurable choice of  $\theta \in \Omega$  which minimizes the objective function

$$L_T(\theta) = \sum_{t=1}^T \text{tr} \left[ \Gamma^{\frac{1}{2}} (Y_t - \mu - ABX_t) (Y_t - \mu - ABX_t)' \Gamma^{\frac{1}{2}} \right].$$

**Remark.** Two convenient choices of  $\Gamma$  are  $\Gamma = I$  or  $\Gamma = \tilde{\Sigma}_{\epsilon\epsilon}^{-1} = T[(Y - \tilde{\mu} - \tilde{C}X)(Y - \tilde{\mu} - \tilde{C}X)']^{-1}$ , where  $\tilde{\mu} = \bar{Y} - \tilde{C}\bar{X}$  and  $\tilde{C} = YX'(XX')^{-1}$  is the full-rank least squares estimate of  $C$ , with  $X$  defined in terms of some initial but fixed threshold estimate.

The estimators of  $A$  and  $B$  are required to satisfy the normalization conditions

$$BS_{xx}B' = \text{diagonal matrix}, \quad A'\Gamma A = I_r. \quad (5.2)$$

**Theorem 3.** *Suppose that  $(Y_t)$  satisfying (5.1) is stationary ergodic, with finite second moments, and that the stationary distribution of  $(Y_1, \dots, Y_{p+1})'$  admits a positive density everywhere. The errors  $\epsilon_t$  are i.i.d. with an absolutely continuous distribution and finite second moment. Assume the matrices  $A$  and  $B$  satisfy the normalization conditions (5.2). Then the estimator  $\hat{\theta}_T \rightarrow \theta_0$  a.s. Also, for the residual covariance matrix,  $\hat{\Sigma} = \Sigma_{\epsilon\epsilon} + o_P(1)$ , where  $\hat{\Sigma} = \sum_{t=1}^T (Y_t - \hat{\mu} - \hat{A}\hat{B}X_t)(Y_t - \hat{\mu} - \hat{A}\hat{B}X_t)'/T$ .*

We note that the estimator obtained from Steps 2 and 3 in Section 3, with a fixed  $\Gamma$ , is a weighted least squares estimator of  $\theta$ . We now study the convergence rate of the threshold parameter  $\gamma$ . For simplicity, we state the result for the case that  $m = 2, p = d = 1$  although the extension to the case of higher order ( $p > 1$ ) and more than two regimes ( $m > 2$ ) is straightforward. Let  $\mathbf{Y}_t = (Y_{1t}, Y_{2t})'$ . Then  $(\mathbf{Y}_t)$  is a Markov chain. Denote its  $l$ -step transition probability by  $P^l(\mathbf{y}, A)$ , where  $\mathbf{y} \in R^2$  and  $A$  is a Borel set. The following set of regularity conditions is required for deriving the convergence rate of the threshold estimator of the RETAR model.

**Condition 1.**  $(\mathbf{Y}_t)$  admits a unique invariant measure  $\pi(\cdot)$  such that there exist constants  $K$  and  $\rho < 1$ , such that for all  $\mathbf{y} \in R^2, t \in \mathcal{N}$ ,  $\|P^t(\mathbf{y}, \cdot) - \pi(\cdot)\| \leq K(|\mathbf{y}|+1)\rho^t$ , where  $\|\cdot\|$  and  $|\cdot|$  denote the total variation norm and the Euclidean norm, respectively.

**Condition 2.** The error  $\epsilon_t$  are absolutely continuous with a uniformly continuous and positive pdf. Furthermore,  $E(\epsilon_{it}\epsilon_{jt}\epsilon_{kt}\epsilon_{lt}) < \infty$  for all positive  $i, j, k$  and  $l$ , where  $\epsilon_{it}$  is the  $i$ th component of  $\epsilon_t$ .

**Condition 3.**  $(\mathbf{Y}_t)$  is stationary with a bounded pdf  $\pi(\cdot)$ , where  $\sup_y \pi(y) < K$  for some  $K > 0$ . Also,  $E(Y_{it}Y_{jt}Y_{kt}Y_{lt}) < \infty$  for all positive  $i, j, k$  and  $l$ , where  $Y_{it}$  is the  $i$ th component of  $Y_t$ .

**Condition 4.** The autoregressive (conditional mean function of  $Y_t$ ) function is discontinuous. If  $\gamma_0 = (0, 0)'$ , then the autoregressive function is discontinuous iff

$$A_0B_0(0, 0, 1, 0, 0, 0)' \neq 0, \quad (5.3)$$

$$A_0B_0(0, 0, 0, 1, 0, 0)' \neq 0. \quad (5.4)$$

Note that the threshold can be transformed to 0 by considering  $Y'_t = Y_t - r_0$ , it follows a RETAR model with generally different  $\mu_0$  and  $B_0$ .

Similar to Proposition 1 of Chan (1993a), we have the following super-consistent property for the threshold parameter estimator.

**Theorem 4.** *Suppose Conditions 1 to 4 hold. Then  $\hat{\gamma}_T = \gamma_0 + O_P(1/T)$ .*

#### Remarks

1. Condition 4 can be derived as follows. Let

$$X^{(1)} = (x_1, x_2, 1, 1, x_1, x_2)', \quad (5.5)$$

$$X^{(2)} = (x_1, x_2, 0, 1, 0, x_2)', \quad (5.6)$$

$$X^{(3)} = (x_1, x_2, 0, 0, 0, 0)', \quad (5.7)$$

$$X^{(4)} = (x_1, x_2, 1, 0, x_1, 0)'. \quad (5.8)$$

Since the autoregressive function is discontinuous, we have

$$A_0B_0 \lim_{x_2 \rightarrow 0^+} X^{(1)} \neq A_0B_0 \lim_{x_2 \rightarrow 0^-} X^{(4)}, \quad \text{for fixed } x_1 > 0,$$

$$A_0B_0 \lim_{x_2 \rightarrow 0^+} X^{(2)} \neq A_0B_0 \lim_{x_2 \rightarrow 0^-} X^{(3)}, \quad \text{for fixed } x_1 \leq 0,$$

$$A_0B_0 \lim_{x_1 \rightarrow 0^+} X^{(1)} \neq A_0B_0 \lim_{x_1 \rightarrow 0^-} X^{(2)}, \quad \text{for fixed } x_2 > 0,$$

$$\begin{aligned}
 A_0 B_0 \lim_{\substack{x_1 \rightarrow 0^+ \\ x_2 \rightarrow 0^+}} X^{(4)} &\neq A_0 B_0 \lim_{\substack{x_1 \rightarrow 0^- \\ x_2 \rightarrow 0^-}} X^{(3)}, \quad \text{for fixed } x_2 \leq 0, \\
 A_0 B_0 \lim_{\substack{x_1 \rightarrow 0^+ \\ x_2 \rightarrow 0^+}} X^{(1)} &\neq A_0 B_0 \lim_{\substack{x_1 \rightarrow 0^- \\ x_2 \rightarrow 0^-}} X^{(3)}, \\
 A_0 B_0 \lim_{\substack{x_1 \rightarrow 0^- \\ x_2 \rightarrow 0^+}} X^{(1)} &\neq A_0 B_0 \lim_{\substack{x_1 \rightarrow 0^+ \\ x_2 \rightarrow 0^-}} X^{(2)}.
 \end{aligned}$$

Equivalently,

$$\begin{aligned}
 A_0 B_0(x_1, 0, 1, 1, x_1, 0)' &\neq A_0 B_0(x_1, 0, 1, 0, x_1, 0)', \\
 A_0 B_0(x_1, 0, 0, 1, 0, 0)' &\neq A_0 B_0(x_1, 0, 0, 0, 0, 0)', \\
 A_0 B_0(0, x_2, 0, 1, 0, x_2)' &\neq A_0 B_0(0, x_2, 1, 1, 0, x_2)', \\
 A_0 B_0(0, x_2, 0, 0, 0, 0)' &\neq A_0 B_0(0, x_2, 1, 0, 0, 0)', \\
 A_0 B_0(0, 0, 1, 1, 0, 0)' &\neq A_0 B_0(0, 0, 0, 0, 0, 0)', \\
 A_0 B_0(0, 0, 0, 1, 0, 0)' &\neq A_0 B_0(0, 0, 1, 0, 0, 0)',
 \end{aligned}$$

which in turn, are equivalent to (5.3) and (5.4) in Condition 4.

2. For the general case with order  $p$ ,  $m$  time series,  $d \leq p$  and assuming the true threshold parameter  $\gamma_0 = 0$ , Conditions 1 and 3 need only be modified with  $\mathbf{Y}_t = (Y_{1t}, \dots, Y_{mt}, Y_{1,t-1}, \dots, Y_{m,t-1}, \dots, Y_{1,t-p}, \dots, Y_{m,t-p})'$ . Also, equations (5.3) and (5.4) of Condition 4 are replaced by the following:

$$(0'_{mp}, J'_i, 0'_{(d-1)m}, (J_1 \gamma_0)', \dots, (J_m \gamma_0)', 0'_{(p-d)m})' \neq 0 \quad \text{for all } i = 1, \dots, m,$$

where  $J_i$  is an  $m$ -dimensional zero vector except that its  $i$ th element equals 1, and  $0_k$  is a  $k$ -dimensional zero vector. The notation  $J_i \gamma_0$  is defined as componentwise multiplication.

## 6. Asymptotic Distribution of the Estimators of the RETAR Model

In this section, we show that under suitable conditions the weighted least squares estimators  $\hat{A}$  and  $\hat{B}$  with  $\Gamma = \Sigma_{\epsilon\epsilon}^{-1}$  are asymptotically normal with a distribution the same as that for the case when  $\gamma$  is known. Write the parameters of  $A$  and  $B$  as  $A = [\alpha_1, \dots, \alpha_r]$  and  $B = [\beta_1, \dots, \beta_r]'$ , where  $\alpha$ 's and  $\beta$ 's are the column vectors of  $A$  and  $B$  respectively. The corresponding estimators of  $A$  and  $B$  are written as  $\hat{A} = [\hat{\alpha}_1, \dots, \hat{\alpha}_r]$  and  $\hat{B} = [\hat{\beta}_1, \dots, \hat{\beta}_r]'$ .

**Theorem 5.** *Suppose Conditions 1 to 4 hold. Let  $\Sigma_{yx} = \Sigma'_{xy} = \text{Cov}(Y_t, X_t)$ , and  $\Sigma_{xx} = \text{Cov}(X_t)$  be nonsingular. The matrix  $\Sigma_{\epsilon\epsilon}$  is assumed to be positive definite. Let the true parameters  $(\text{vec}(A)', \text{vec}(B)') \in \Theta$ , a compact set defined by the normalization conditions. Then, with  $\Gamma = \Sigma_{\epsilon\epsilon}^{-1}$ , the weighted least squares*

vector variates  $T^{1/2}(\hat{\alpha}_j - \alpha_j)$  and  $T^{1/2}(\hat{\beta}_j - \beta_j)$  ( $j = 1, \dots, r$ ) have a joint limiting distribution as  $T \rightarrow \infty$  that is singular multivariate normal with null mean vectors, and

$$\begin{aligned}
E\{T(\hat{\alpha}_j - \alpha_j)(\hat{\alpha}_l - \alpha_l)'\} &\rightarrow \begin{cases} \sum_{k \neq j=1}^m \frac{\lambda_j^2 + \lambda_k^2}{(\lambda_j^2 - \lambda_k^2)^2} \alpha_k \alpha_k' & (j = l) \\ -\frac{\lambda_j^2 + \lambda_l^2}{(\lambda_j^2 - \lambda_l^2)^2} \alpha_l \alpha_l' & (j \neq l) \end{cases} \\
E\{T(\hat{\beta}_j - \beta_j)(\hat{\beta}_l - \beta_l)'\} &\rightarrow \begin{cases} \sum_{k \neq j=1}^r \frac{3\lambda_j^2 - \lambda_k^2}{(\lambda_j^2 - \lambda_k^2)^2} \beta_k \beta_k' + \Sigma_{xx}^{-1} & (j = l) \\ -\frac{\lambda_j^2 + \lambda_l^2}{(\lambda_j^2 - \lambda_l^2)^2} \beta_l \beta_l' & (j \neq l) \end{cases} \quad (6.1) \\
E\{T(\hat{\alpha}_j - \alpha_j)(\hat{\beta}_l - \beta_l)'\} &\rightarrow \begin{cases} 2\lambda_j^2 \sum_{k \neq j=1}^r \frac{1}{(\lambda_j^2 - \lambda_k^2)^2} \alpha_k \beta_k' & (j = l) \\ -\frac{2\lambda_l^2}{(\lambda_j^2 - \lambda_l^2)^2} \alpha_l \alpha_l' & (j \neq l), \end{cases}
\end{aligned}$$

where the  $\lambda_j^2$ ,  $j = 1, \dots, r$  are the eigenvalues of the matrix  $\Gamma^{1/2} \Sigma_{yx} \Sigma_{xx}^{-1} \Sigma_{xy} \Gamma^{1/2}$  with  $\Gamma = \Sigma_{cc}^{-1}$ .

We have abused the notation in (6.1) in that the limiting expressions are the asymptotic covariance matrices of  $T^{1/2}(\hat{\alpha}_j - \alpha_j)$ ,  $T^{1/2}(\hat{\beta}_j - \beta_j)$  and their cross product term; we do not assert convergence in moments for the parameter estimators.

It can be checked by routine but tedious analysis, using the techniques in the proof presented in the online supplement (<http://www3.stat.sinica.edu.tw/statistica>), that for known rank of  $C$  and Gaussian homogeneous errors,  $\hat{A}$  and  $\hat{B}$  from the estimation method in Section 3 enjoy the same asymptotic distribution stated in Theorem 5. Furthermore,  $\hat{\gamma}$  equals  $\gamma_0 + O_P(1/T)$  and  $\hat{d}$  converges a.s. to  $d_0$ .

## 7. Application

The classical Canada lynx data set consists of the annual record of the numbers of the lynx pelts collected in the Mackenzie River district of Northwest Canada for the period 1821-1934, inclusively. These lynx counts are known to fluctuate periodically, with asymmetrical cycles consisting of sharp and large peaks. Previous studies (see Section 7.2 of Tong (1990) for a review) suggest that the threshold model can adequately model several nonlinear features of the lynx data. Although these data have been analyzed by many authors (see, e.g., Tong (1977), Tong and Lim (1980), Lim (1987), Tong (1990) and Lin and Pourahmadi (1998)), most have focused on the classical lynx data from the Mackenzie River district. Further lynx records are available (Stenseth et al. (1999)). Altogether there are two (old and modern) panels of lynx series, labeled L1-L14 and L15-L22, respectively. Series L3 corresponds to the classical lynx data. Here, we

concentrate on the modern series L15-L22 (Figure 1) that span the years 1920 to 1994.

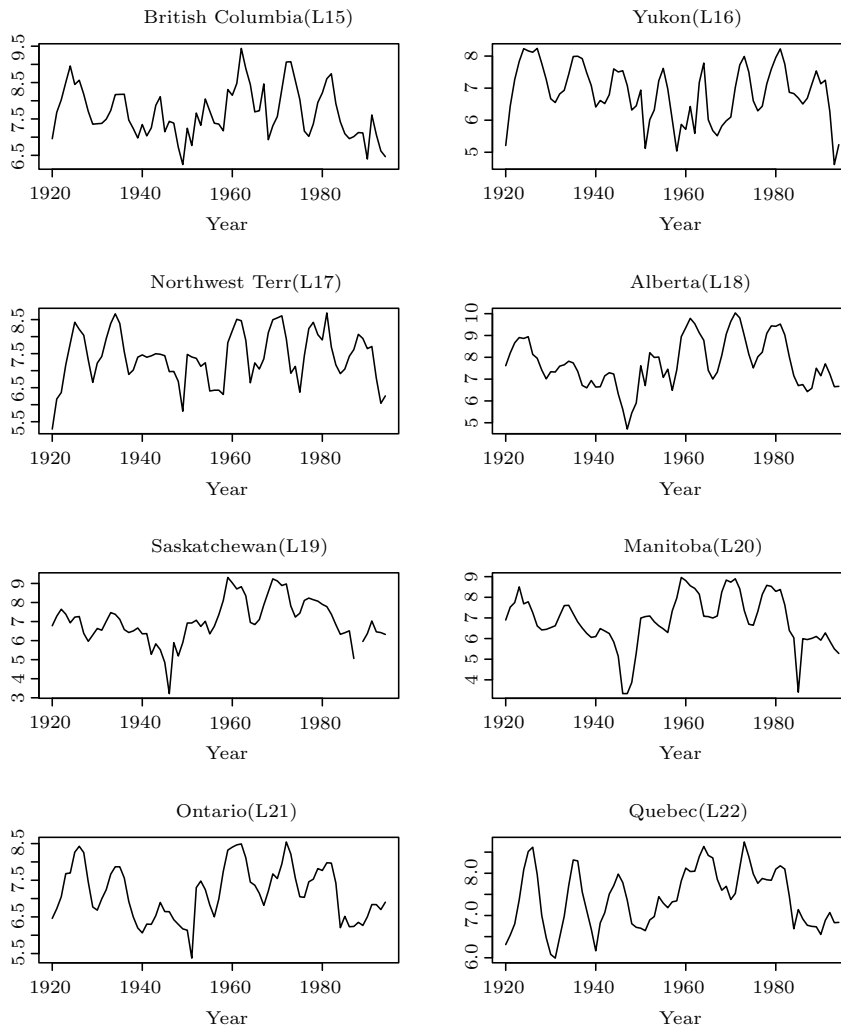


Figure 1. Recent Lynx Series by Provinces.

The dynamics of the lynx data may be structured by two eco-climatic zones: the open boreal forest and the closed boreal forest. Alternatively, it may be structured by three geo-climatic zones (Pacific-maritime, Continental and Atlantic-maritime) via the regional influence of a single large-scale climate system called the North Atlantic Oscillation; see Stenseth et al. (1999). Based on a comparative study of the two preceding hypotheses and the hypothesis of no common structure, Stenseth et al. (1999) concluded that there exists a structural simi-

larity in the lynx data throughout Canada and this structural similarity may be classified by three geo-climatic zones: 'Pacific-maritime', 'Continental' and 'Atlantic-maritime'.

An exhaustive search for common structure in the panel of lynx data seems numerically infeasible. Here, we aim to demonstrate that the RETAR model provides an useful approach to empirically exploring the common structure in a panel of time series. As the old panel of lynx data consists of time series from different time spans, we focus on the recent lynx series L15-L22. As an illustration, we fit a RETAR model to the modern lynx panel. First we set  $(p_1, p_2, d) = (2, 2, 2)$ . The autoregressive orders  $(p_1, p_2)$  and the delay parameter  $d$  are chosen as those considered by Chan, Tong and Stenseth (2004) and Stenseth et al. (1999). Under this framework, they developed some methods for testing for common structures in a panel of threshold model. A natural logarithm transformation is used to stabilize the variance. Due to the presence of zero values in the data, 1 is added to each datum before the transformation. Table 1 reports the rank of  $A(B)$  in the RETAR model as selected by using the corrected Bartlett criterion, at the end of the final parameter iterate of the method introduced in Section 3. It appears from the result that the rank is 3. The AIC and BIC criteria select the rank to be 4 and 1 respectively. It is known that the AIC criterion tends to over-fit the model and the BIC criterion tends to under-fit the model. Henceforth in this example we set the rank to be three. Note that for all the tests, the alternative hypothesis is that the model is of full rank, i.e., rank equals eight (recall that  $m = \dim(Y) = 8$ ,  $n = \dim(X) = 40$ ). We have also experimented with adding 0.1 or 0.01 to the data before log-transformation, but rank 3 is strongly suggested by the likelihood ratio test in both cases. Since the residual covariance is 'smallest' in the case with 1 added to the data before the log-transformation, we henceforth report the analysis with the  $\log(1 + x)$  transformation applied to the lynx data. Also the threshold parameter estimate is given by  $\hat{\gamma} = (7.420, 7.083, 7.944, 7.238, 6.581, 6.392, 6.500, 7.071)$ .

Table 1. Summary of results for LR tests on the rank of the coefficient matrix for the log transformed lynx data, with 1 added to the data before the log-transformation. The alternative hypothesis is  $r=8$ . The determinant of the residual covariance is 1.112e-04.

r=Rank	$\mathcal{M}$ =LR statistic	d.f.	p-value
1	398.9674	df= 273	p= 0.000
2	289.1617	df= 228	p= 0.004
3	194.8913	df= 185	p= 0.295
4	132.5408	df= 144	p= 0.744
5	82.8078	df= 105	p= 0.946
6	45.8200	df= 68	p= 0.982
7	15.6245	df= 33	p= 0.996



After having estimated the parameters of the RETAR model, the next step is to see what kind of information we can infer from the fitted model. Certainly the fact that the rank equals 3 suggests that we may split the eight lynx series into three groups. Recall that for identifiability reasons the coefficient matrix  $C$  is decomposed as the product of two matrices  $A$  and  $B$  satisfying  $BS_{xx}B' = \Lambda^2$  and  $A'\Gamma A = I_r$ . The selection of the normalization condition is subjective and often chosen for mathematical convenience. Rotating the estimates of  $A$  and  $B$  obtained from a set of normalization conditions may lead to more interpretable estimates. A convenient method of rotation is the varimax method; see Mardia, Kent and Bibby (1979).

In Table 2 we report the varimax-rotated  $\hat{A}$ , the estimated factor loading matrix with standard errors. From Table 2 we see the first factor loads heavily on sites 18 (Alberta), 19 (Saskatchewan) and 20 (Manitoba), and note that these are provinces in the southern part of the Continental-climatic region. Factor 2, loading heavily on site 22 (Quebec), represents a province belonging to the Atlantic-maritime region, while factor 3, loading heavily on sites 15 (British Columbia), 16 (Yukon Territory), 17 (N.W. Territory) and 21 (Ontario), represents provinces in the Pacific-maritime region, except for Ontario.

Another perspective on the classification of the lynx series may be obtained by rotating  $\hat{A}$  and  $\hat{B}$  so that  $\hat{B}X$  is orthonormalized; see Table 3 for the rotated  $\hat{A}$  so obtained. The Euclidean distance of any two rows of the  $\hat{A}$  then measure the dissimilarity of the dynamics of the two corresponding lynx series. Figure 2 reports the hierarchical cluster analysis of the eight series based on  $\hat{A}$ , using the options of average Euclidean distance and complete linkage. It now appears that we have two major groups: Group 1 consists of Alberta, Saskatchewan and Manitoba, which can be interpreted as Continental group; Group 2 consists of Ontario, British Columbia, N.W. Territory, Quebec and Yukon, with the last two provinces possibly classified into a third group. Generally speaking, the second cluster represents a maritime group. It seems that Stenseth et al. (1999) obtained a finer classification by studying the lynx dynamics over the decrease phase.

Scatter plots of each of the response variable versus the first three predictive index variables (Li (2000)) suggest that the response variables of L18, L19 and L20 have a quite strong positive linear relationship with  $x_1^* = \hat{\beta}'_1 X$ , and that L22 is strongly related to the second index variable  $x_2^* = \hat{\beta}'_2 X$ , while L15 and L16 are strongly related to the third index variable  $x_3^* = \hat{\beta}'_3 X$ . In Figure 3 we show the time series plots of the first three latent principal component processes  $\hat{\beta}'_i X_t, 1 \leq i \leq 3$ . These time series plots are similar to the time series plots of L19, L22 and L15, which means, for example, the dynamics of the series in the first group is heavily related to the dynamics of series L19. Finally we note that the RETAR model may lead to more accurate forecasts than forecasting each

Table 2. Maximum likelihood estimates  $\hat{A}$  of the log-transformed data with rank=3 principal components obtained by using the varimax rotation. Standard errors are enclosed by parentheses.

response	site	Factor loadings		
		1	2	3
1	15	0.025* (0.014)	0.066* (0.015)	0.119* (0.013)
2	16	-0.134* (0.028)	0.053 (0.039)	0.262* (0.024)
3	17	0.058* (0.020)	-0.007 (0.017)	0.073* (0.025)
4	18	0.210* (0.013)	0.008 (0.010)	0.095* (0.025)
5	19	0.281* (0.028)	-0.035 (0.033)	-0.041 (0.050)
6	20	0.263* (0.017)	-0.013 (0.020)	0.029 (0.037)
7	21	0.073* (0.015)	0.036* (0.013)	0.106* (0.015)
8	22	-0.021* (0.009)	0.149* (0.007)	0.037 (0.021)

Table 3. Maximum likelihood estimate  $\hat{A}$  of the log-transformed data with rank=3 principal component obtained by rotation such that  $\hat{B}X$  is orthonormalized. Standard errors are enclosed by parentheses.

response	site	Factor loadings		
		1	2	3
1	15	0.559* (0.033)	-0.065 (0.057)	0.227* (0.035)
2	16	0.178* (0.052)	-0.087 (0.150)	0.723* (0.036)
3	17	0.333* (0.045)	-0.201* (0.057)	0.056 (0.072)
4	18	1.009* (0.032)	-0.405* (0.036)	-0.106 (0.065)
5	19	0.864* (0.085)	-0.393* (0.133)	-0.491* (0.111)
6	20	1.010* (0.055)	-0.425* (0.084)	-0.317* (0.084)
7	21	0.613* (0.027)	-0.175* (0.041)	0.121* (0.047)
8	22	0.536* (0.025)	0.295* (0.040)	0.169* (0.048)

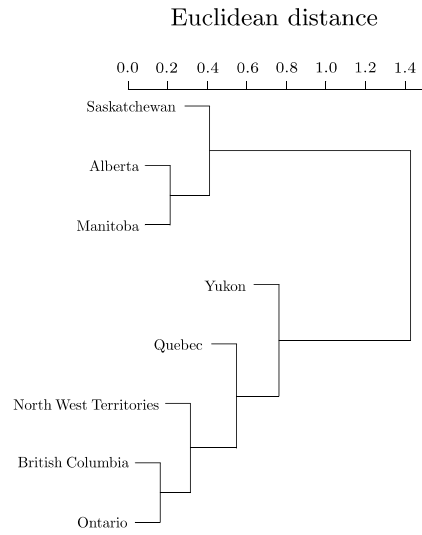


Figure 2. Hierarchical cluster analysis of the lynx data based on the average Euclidean distance and complete linkage.

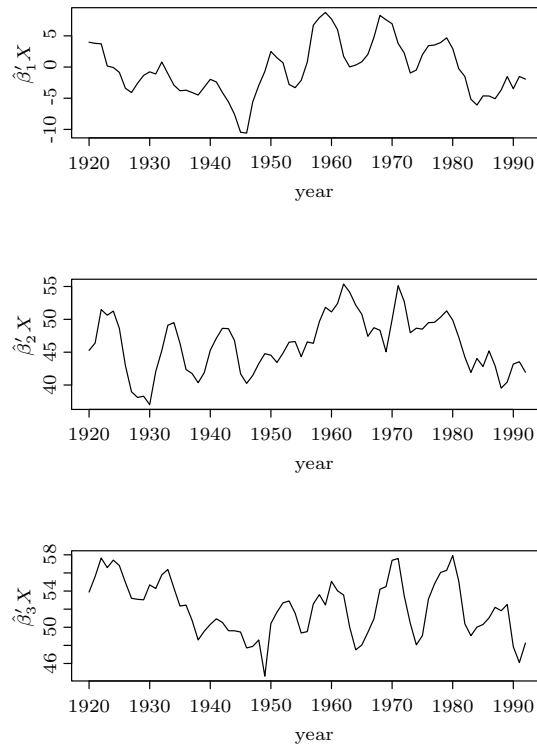


Figure 3. Time series plots of the 3 latent processes  $\hat{\beta}_i^t X_t$ ,  $i = 1, 2, 3$ .

series from individually fitted Threshold Autoregressive models, even after adjusting for contemporaneous correlations; see Li (2000) for an illustration with the lynx data.

## 8. Conclusion

We have demonstrated the usefulness of the RETAR model for exploring the common dynamic structure of a panel of nonlinear time series. The RETAR model assumes that the nonlinear principal components are piecewise linear functions of the past lags of the time series. An interesting future research problem concerns the use of other nonlinear functions of the data to model the nonlinear principal components. A related approach is to model the nonlinear principal components nonparametrically, see Li and Chan (2001).

## Acknowledgement

We thank Howell Tong, Nils Chr. Stenseth, W.K. Li and the referees for their valuable comments. This work is partially supported by the National Science Foundation (DMS-0405267).

## References

- Anderson, T. W. (1971). *The Statistical Analysis of Time Series*. John Wiley, New York.
- Chan, K. S. (1993a). Consistency and limiting distribution of the least squares estimator of a threshold autoregressive model. *Ann. Statist.* **21**, 520-533.
- Chan, K. S. (1993b). A review of some limit theorems of Markov chains and their applications. In *Dimension Estimation and Models* (Edited by H. Tong). World Scientific Publishing, River Edge, NJ.
- Chan, K. S., Petrucci, J. D., Tong, H. and Woolford, S. W. (1985). A multiple-threshold AR(1) model. *J. Appl. Probab.* **22**, 267-279.
- Chan, K. S. and Tong, H. (1985). On the use of the deterministic Lyapunov function for the ergodicity of stochastic difference equations. *Adv. Appl. Probab.* **17**, 666-678.
- Chan, K. S. and Tong, H. (2001). *Chaos: a Statistical Perspective*. Springer Verlag, New York.
- Chan, K. S., Tong, H. and Stenseth, N. C. (2004). Testing for common structures in a panel of threshold models. *Biometrics* **60**, 225-32.
- Cheng, B. and Tong, H. (1992). On consistent nonparametric order determination and chaos. *J. Roy. Statist. Soc. Ser. B* **54**, 427-449.
- Cline, D. B. H. and Pu, H. H. (2001). Stability of nonlinear time series: what does noise have to do with it? In *Selected Proceedings of the Symposium on Inference for Stochastic Processes* (Edited by I. V. Basawa, C. C. Heyde and R. L. Taylor). IMS Lecture Notes - Monograph Series, Volume 37.
- Foster, F. G. (1953). On the stochastic matrices associated with certain queueing processes. *Ann. Math. Statist.* **24**, 355-360.
- Graybill, F. A. (1983). *Matrices with Applications in Statistics*. Wadsworth, Belmont.

- Izenman, A. J. (1980). Assessing dimensionality in multivariate regression. In *Handbook of Statistics* (Edited by P. Krishnaiah), 571-592. North Holland, New York.
- Li, M.-C. (2000). Multivariate non-linear time series modeling. Unpublished Ph.D. thesis, The University of Iowa.
- Li, M.-C. and K. S. Chan (2001). Semiparametric reduced-rank regression. Technical Report, The University of Iowa, Department of Statistics and Actuarial Science.
- Lim, K. S. (1987). A comparative study of various univariate time series models for Canadian lynx data. *J. Time Ser. Anal.* **8**, 161-176.
- Lin, T. C. and Pourahmadi, M. (1998). Nonparametric and non-linear models and data mining in time series: a case-study on the Canadian lynx data. *Appl. Statist.* **47**, 187-201.
- Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press, London.
- Meyn, S. P. and Tweedie, R. L. (1993). *Markov Chains and Stochastic Stability*. Springer-Verlag, New York.
- Nummelin, E. (1984). *General Irreducible Markov Chains and Non-Negative Operators*. Cambridge University Press.
- Petruccioli, J. D. and Woolford, S. W. (1984). A threshold AR(1) model. *J. Appl. Probab.* **21**, 270-286.
- Reinsel, G. C. and Velu, R. (1998). *Multivariate-Reduced Rank Regression*. Springer-Verlag, New York.
- Robinson, P. M. (1973). Generalized canonical analysis for time series. *J. Multivariate Anal.* **3**, 141-60.
- Stenseth, N. C., Chan, K. S., Tong, H., Boonstra, R., Boutin, S., Krebs, C. J., Post, E., O'Donoghue, M., Yoccoz, N. G., Forchhammer, M. C. and Hurrell, J. W. (1999). Common dynamic structure of canadian lynx populations within three geo-climatic zones. *Science* **285**, 1071-1073.
- Tjøstheim, D. (1990). Non-linear time series and markov chains. *Adv. Appl. Probab.* **22**, 587-611.
- Tong, H. (1977). Some comments on the Canadian lynx data. *J. Roy. Statist. Soc. Ser. A* **140**, 432-436.
- Tong, H. (1990). *Non-linear Time Series: A Dynamical System Approach*. Oxford University Press, Oxford.
- Tong, H. and Lim, K. S. (1980). Threshold autoregression, limit cycles and cyclical data (with discussion). *J. Roy. Statist. Soc. Ser. B* **42**, 24-292.
- Tweedie, R. L. (1975). A relation between positive recurrence and mean drift for markov chains. *Austral. J. Statist.* **17**, 9-102.
- Tweedie, R. L. (1976). Criteria for classifying general markov chains. *Adv. Appl. Probab.* **8**, 737-771.
- Velu, R., Reinsel, G. C. and Wichern, D. W. (1986). Reduced rank models for multiple time series. *Biometrika* **73**, 10-18.

The EMMES Corporation, 401 N. Washington Street, Rockville, MD 20850, U.S.A.

E-mail: mli@emmes.com

The University of Iowa, Department of Statistics & Actuarial Science, Iowa City, IA 52242, U.S.A.

E-mail: kchan@stat.uiowa.edu

(Received May 2005; accepted February 2006)

## MULTIVARIATE REDUCED-RANK NONLINEAR TIME SERIES MODELING

Ming-Chung Li and Kung-Sik Chan

*The EMMES Corporation and The University of Iowa*

### Supplementary Material

#### Appendix A. Proof of Theorem 1

We first consider the case that the error  $\epsilon_t$  is suppressed from (4.2). Without loss of generality, assume  $\vec{a} \in R_1$ . If  $X_t = x \in R_k$  for some  $k$ , then  $X_{t+i} = \vec{a} \vec{b}'_k x = \vec{a} (\vec{b}'_k x) \in R_1$  or  $R_3$  for any  $i > 0$ . Therefore the stability of the deterministic system is determined by the first and the third difference equations. A deterministic system composed of these two difference equations is stable if and only if  $\lambda(A_1) < 1$ ,  $\lambda(A_3) < 1$  and  $\lambda(A_1 A_3) < 1$ . This justifies the statement regarding the stability of the skeleton.

Now let us consider the stochastic case of (4.2). We note that the eigenvalue conditions imply that the origin is uniformly asymptotically stable for the skeleton, which implies the geometric ergodicity of  $\{X_t\}$ ; see Theorem 4.5 in Chan and Tong (1985).

Next we show the nonergodicity for the two cases as stated in Remark 1.

**Lemma 1.** *If  $\lambda(C_j) > 1$  or  $\lambda(C_{j+2}) > 1$  then the process  $\{X_t\}$  is not ergodic.*

**Proof.** Without loss of generality we assume  $\vec{a} \in R_1$  and only consider the case  $\lambda(C_1) > 1$ . We first show that  $X_t$  will go to infinity, with positive probability. Let  $M > 0$  be a constant to be determined below. Consider  $X_t = \vec{a} h + \epsilon_t \in R_1$  and that  $X_t = \begin{pmatrix} X_{1t} \\ X_{2t} \end{pmatrix}$  is such that  $X_{it} > M, i = 1, 2$ . For any  $1 < \eta < \lambda(C_1) \equiv \lambda_1$  and conditioned on  $X_t$ , we have,

$$\begin{aligned} & P((1, 1)X_{t+1} > 2^{-1}(\eta + 1)(1, 1)X_t) \\ &= P((1, 1)(\vec{a} \vec{b}'_1 X_t + \epsilon_{t+1}) > 2^{-1}(\eta + 1)(1, 1)X_t) \\ &= P((1, 1)\epsilon_{t+1} > (2^{-1}(\eta + 1)(1, 1) - (1, 1)\vec{a} \vec{b}'_1)X_t) \\ &= P((1, 1)\epsilon_{t+1} > (2^{-1}(\eta + 1)(1, 1)(\vec{a} h + \epsilon_t) - (1, 1)\vec{a} \vec{b}'_1(\vec{a} h + \epsilon_t))) \\ &= P((1, 1)\epsilon_{t+1} > (2^{-1}(\eta + 1) - \lambda_1)(1, 1)(\vec{a} h + \epsilon_t) + \lambda_1(1, 1)\epsilon_t - (1, 1)\vec{a} \vec{b}'_1 \epsilon_t) \end{aligned}$$

$$\begin{aligned}
&= P(\epsilon_{t+1}^* > 2^{-1}(1-\eta)(1,1)X_t) \quad \text{where } \epsilon_{t+1}^* = (1,1)[\epsilon_{t+1} + (\vec{a} \vec{b}'_1 - \lambda_1)\epsilon_t] \\
&= 1 - P(\epsilon_{t+1}^* \leq 2^{-1}(1-\eta)(1,1)X_t) \\
&\geq 1 - P(|\epsilon_{t+1}^*| > 2^{-1}(\eta-1)(1,1)X_t) \\
&\geq 1 - E|\epsilon_{t+1}^*|/((\eta-1)M) \quad \text{whenever } X_{1t} > M, X_{2t} > M \\
&= 1 - c \quad \text{where } c = E|\epsilon_{t+1}^*|/((\eta-1)M) \text{ is chosen such that } c < 1. \quad (\text{A.1})
\end{aligned}$$

Similarly, we can show that given  $X_{11} > M$  and  $X_{21} > M$ ,

$$\begin{aligned}
&P((1,1)X_3 > 2^{-1}(\eta+1)(1,1)X_2, (1,1)X_2 > 2^{-1}(\eta+1)(1,1)X_1 | X_1) \\
&\geq \left(1 - \frac{E|\epsilon_3^*|}{\frac{\eta+1}{2}\frac{\eta-1}{2}2M}\right)(1-c) \\
&= (1-c\beta)(1-c)
\end{aligned}$$

where  $\beta = 2/(\eta+1) < 1$ . Continuing in this manner, we have, whenever  $X_{1,1} > M, X_{2,1} > M$ ,

$$\begin{aligned}
P((1,1)X_{l+1} > 2^{-1}(\eta+1)(1,1)X_l, l = 1, \dots, t | X_1) &\geq \prod_{i=1}^t (1 - c\beta^{i-1}) \\
&\geq (1-c)^{1/(1-\beta)}
\end{aligned}$$

for all  $t$ ; the last inequality follows from routine analysis. Consequently for any  $X_0 \in R_1$

$$P((1,1)X_t \rightarrow \infty | X_0) \geq (1-c)^{1/(1-\beta)} P(X_{1,1} > M, X_{2,1} > M | X_0) > 0.$$

Hence,  $\{X_t\}$  is not ergodic if  $\lambda(C_1) > 1$ .

**Lemma 2.** *If  $\lambda(C_j) < 0, \lambda(C_{j+2}) < 0$  and  $\lambda(C_j)\lambda(C_{j+2}) > 1$  then the process  $\{X_t\}$  is not ergodic.*

**Proof.** Without loss of generality we consider the case  $\vec{a} \in R_1, \lambda(C_1) < -1$  and  $\lambda(C_1)\lambda(C_3) > 1$ . The proof is similar to that for the above lemma except that we show that the Markov chain  $\{X_{2t}; t \geq 0\}$  has the property that, for any  $X_0 \in R$ ,

$$P((1,1)X_{2t} \rightarrow \infty | X_0) > 0.$$

As the proof is similar to the proof of lemma 1, we skip the detail.

## Appendix B. Proof of Theorem 2

The proof is similar to that used in Chan and Tong (1985) except for a minor change. Let  $Z = (z_1, \dots, z_m) = (x_m, x_{m-1}, \dots, x_2, x_1) \in R^m$ . As  $\max_i \sum_j \|A_{ij}\|$

$< 1 \exists p_1, \dots, p_m > 0$  such that  $\max_i \sum_j \|A_{ij}\| p_1/p_j < \theta < 1$  for some  $\theta$ . Moreover,  $\theta$  can be chosen such that  $\theta > p_{i+1}/p_i$ . Define the test function  $g : R^m \rightarrow R$  by  $g(z) = 1 + \max_i |z_i| p_i$ . Then

$$\begin{aligned}
& E[g(Z_{t+1})|Z_t = z] \\
&= 1 + \max\{E|h(X_1, \dots, X_m) + e_{t+1}|p_1, E|X_m|p_2, E|X_{m-1}|p_3, \dots, E|X_2|p_m\} \\
&\leq c + \max\{|h(X_1, \dots, X_m)|p_1, |X_m|p_2, \dots, |X_2|p_m\} \\
&= c + \max\left\{\left|\sum_j^m A_{kj} X_{m+1-j}\right|p_1, |X_m|p_2, \dots, |X_2|p_m\right\} \\
&\leq c + \max\left\{\sum_j^m \|A_{kj}\| |X_{m+1-j}|p_1, |X_m|p_2, \dots, |X_2|p_m\right\} \\
&\leq c + \theta \max\{|X_m|p_1, |X_{m-1}|p_2, \dots, |X_2|p_{m-1}, |X_1|p_m\} \\
&= c' + \theta g(z).
\end{aligned}$$

Then the rest of the proof follows the argument as in Chan and Tong (1985).

### Appendix C. Proof of Theorem 3

We review some theories of uniform convergence of empirical measures which are useful for showing the strong consistency of the weighted least squares estimator of RETAR model; for details, see Chapter 2 of Pollard (1984).

Suppose that we observe a stochastic process  $\{\xi_i\}$  consisting of independent samples taken from  $P$ . Let  $P_n$  represent the empirical measure that puts equal mass at each of the  $n$  observations  $\xi_1, \dots, \xi_n$ , so that an average over the observations can be written as an expectation with respect to  $P_n$ :

$$\frac{1}{n} \sum_{i=1}^n f(\xi_i) \equiv P_n f = \int f dP_n.$$

The following two theorems concerning the uniform law of large numbers (ULLN) taken from Pollard (1984, p.8) will be useful below.

**Theorem 1.** *Suppose that for each  $\epsilon > 0$  there exists a finite class  $\mathcal{F}_\epsilon$  containing lower and upper approximations to each  $f$  in  $\mathcal{F}$ , for which*

$$f_{\epsilon,L} \leq f \leq f_{\epsilon,U} \quad \text{and} \quad P(f_{\epsilon,U} - f_{\epsilon,L}) < \epsilon. \quad (\text{C.1})$$

Then  $\sup_{\mathcal{F}} |P_n f - P f| \rightarrow 0$  a.s..

**Theorem 2.** *Suppose that for each  $\epsilon > 0$  there exists a finite class  $\mathcal{F}_\epsilon$  of functions such that for each  $f$  in  $\mathcal{F}$  there exists an  $f_\epsilon$  in  $\mathcal{F}_\epsilon$  such that  $f_\epsilon \leq f$  and  $P f_\epsilon \geq$*



$Pf - \epsilon$ . Then, as  $n \rightarrow \infty$

$$\liminf_n \inf_{\mathcal{F}} (P_n f - Pf) \geq 0 \quad a.s..$$

**Remark.** The independence condition on the  $\xi$ 's can be replaced by weaker assumptions, such as stationarity and ergodicity; see Pollard (1984, p.9).

If  $f$  depends on an unknown parameter  $\theta$ , i.e.,  $f(\cdot) = f_\theta(\cdot)$ , and  $\theta_n$  minimizes  $P_n f_\theta$ , the above theorem suggests that  $\theta_n$  might converge to the  $\theta_0$  that minimizes  $Pf$ . The preceding strategy is applied to study the convergence properties of the least squares estimator of the SETAR model which has been proved by Chan (1993) but we give an alternative proof here because it can be adapted to proving the consistency of the least squares estimator of the RETAR model.

**Theorem 3.** Let  $\{Y_1, \dots, Y_n\}$  be generated from the TAR model

$$Y_t = (B'_1 Z_t + e_t)I(Y_{t-d} \leq r) + (B'_2 Z_t + e_t)I(Y_{t-d} > r), \quad (\text{C.2})$$

where  $B_1, B_2$  are  $p$ -dimensional coefficients and  $Z_t = (1, Y_{t-1}, \dots, Y_{t-p})$ . Assume that (i) there exists an  $z = (1, y_1, \dots, y_p)$  with  $y_d = r$  such that  $B'_1 z \neq B'_2 z$ , (ii) the errors  $\{e_t\}$  are i.i.d. with absolutely continuous distribution and finite second moment, (iii)  $\{Y_t\}$  is stationary and ergodic, with finite second moments and the stationary pdf of  $(Y_1, \dots, Y_{p+1})$  is positive everywhere, and (iv)  $d$  is less than some known fixed upper integer bound  $D$ . Then the conditional least squares estimator of  $\theta_n = (B_{1n}, B_{2n}, r_n, q_n)$ , which minimizes  $\sum_t (Y_t - B'_1 Z_t)^2 I(Y_{t-q} \leq r) + \sum_t (Y_t - B'_2 Z_t)^2 I(Y_{t-q} > r)$ , converges almost surely to the true parameters  $\theta_0 = (B_{10}, B_{20}, r_0, d_0)$  as  $n \rightarrow \infty$ .

**Proof.** We first consider the case  $p = d = 1$  with  $d$  known; hence  $q_n = d$ . Let  $B_1 = (c_1, b_1)'$  and  $B_2 = (c_2, b_2)'$ . Throughout the proof, the ULLN will be applied a number of times, the validity of which can be routinely checked using Theorem 6 and hence omitted; a prototype of such checking is given at the end of this proof. The model (C.2) becomes

$$Y_t = (c_1 + b_1 Y_{t-1})I(Y_{t-1} \leq r) + (c_2 + b_2 Y_{t-1})I(Y_{t-1} > r) + e_t.$$

Denote by  $P$  the joint distribution of  $Q_t = (Y_t, Y_{t-1})$  and  $P_n$  the empirical measure constructed by sampling from  $P$ . Let

$$W(c_1, c_2, b_1, b_2, r, P_n) = P_n f_{c_1, c_2, b_1, b_2, r},$$

where

$$f_{c_1, c_2, b_1, b_2, r}(y_1, y_2) = (y_1 - c_1 - b_1 y_2)^2 I(y_2 \leq r) + (y_1 - c_2 - b_2 y_2)^2 I(y_2 > r).$$

Note the finiteness of  $W(\cdot, \cdot, \cdot, \cdot, \cdot, P)$ , because  $P|y_1|^2 < \infty$ . Since  $Q_t$  has a density positive everywhere, the true  $\theta_0 = (c_{10}, c_{20}, b_{10}, b_{20}, r_0)$  is the unique argument minimizing  $W(\cdot, \cdot, \cdot, \cdot, \cdot, P)$ .

We first show that with probability 1,  $r_n \not\rightarrow \infty$  and  $r_n \not\rightarrow -\infty$ . Define

$$g_\theta(y) = (c_1 + b_1 y)^2 I(y \leq r) + (c_2 + b_2 y)^2 I(y > r).$$

Then for  $r > r_0$ ,

$$\begin{aligned} & P_n f_{c_1, c_2, b_1, b_2, r} \\ & \geq P_n [(Y_t - c_1 - b_1 Y_{t-1})^2 I(Y_{t-1} \leq r)] \\ & = P_n [(e_t + g_{\theta_0}(Y_{t-1}) - c_1 - b_1 Y_{t-1})^2 I(Y_{t-1} \leq r)] \\ & = P_n e_t^2 I(Y_{t-1} \leq r) + P_n [(g_{\theta_0}(Y_{t-1}) - c_1 - b_1 Y_{t-1})^2 I(Y_{t-1} \leq r)] \\ & \quad + 2P_n [e_t (g_{\theta_0}(Y_{t-1}) - c_1 - b_1 Y_{t-1}) I(Y_{t-1} \leq r)]. \end{aligned} \quad (\text{C.3})$$

However, as  $r$  is large,

$$\begin{aligned} P_n e_t^2 I(Y_{t-1} \leq r) & \rightarrow P e_t^2 I(Y_{t-1} \leq r) \quad \text{by the ULLN} \\ & \rightarrow \sigma_e^2 \quad \text{if } r \rightarrow \infty \text{ by the Dominated Convergence Theorem,} \end{aligned}$$

and for  $r \geq r_0$ ,

$$\begin{aligned} & P_n [(g_{\theta_0}(Y_{t-1}) - c_1 - b_1 Y_{t-1})^2 I(Y_{t-1} \leq r)] \\ & = P_n [(c_{10} + b_{10} Y_{t-1} - c_1 - b_1 Y_{t-1})^2 I(Y_{t-1} \leq r_0)] \\ & \quad + P_n [(c_{20} + b_{20} Y_{t-1} - c_1 - b_1 Y_{t-1})^2 I(r_0 < Y_{t-1} \leq r)] \\ & = ((c_1 - c_{10})^2 + (b_1 - b_{10})^2) P_n \left[ \left( \frac{c_{10} - c_1}{\sqrt{(c_1 - c_{10})^2 + (b_1 - b_{10})^2}} \right. \right. \\ & \quad \left. \left. + \frac{(b_{10} - b_1) Y_{t-1}}{\sqrt{(c_1 - c_{10})^2 + (b_1 - b_{10})^2}} \right)^2 I(Y_{t-1} \leq r_0) \right] \\ & \quad + ((c_1 - c_{20})^2 + (b_1 - b_{20})^2) P_n \left[ \left( \frac{c_{20} - c_1}{\sqrt{(c_1 - c_{20})^2 + (b_1 - b_{20})^2}} \right. \right. \\ & \quad \left. \left. + \frac{(b_{20} - b_1) Y_{t-1}}{\sqrt{(c_1 - c_{20})^2 + (b_1 - b_{20})^2}} \right)^2 I(r_0 < Y_{t-1} \leq r) \right]. \end{aligned}$$

Applying the ULLN to  $\{P_n[(c + bY_{t-1})^2 I(Y_{t-1} \leq r_0)], c^2 + b^2 = 1\}$ , and also to  $\{P_n[(c + bY_{t-1})^2 I(r_0 + \Delta \leq Y_{t-1} < r)], c^2 + b^2 = 1, \Delta > 0, 0 < r < \infty\}$ , it holds almost surely that for  $n$  sufficiently large and  $r \geq r_0 + \Delta$ ,

$$\begin{aligned} & P_n [(g_{\theta_0}(Y_{t-1}) - c_1 - b_1 Y_{t-1})^2 I(Y_{t-1} \leq r)] \\ & > \kappa [(c_1 - c_{10})^2 + (b_1 - b_{10})^2 + (c_1 - c_{20})^2 + (b_1 - b_{20})^2], \end{aligned}$$

where  $\kappa = \inf\{P[(c + bY_{t-1})^2 I(Y_{t-1} \leq r_0)], P[(c + bY_{t-1})^2 I(r_0 < Y_{t-1} < r)]\}/2 > 0$ , and the infimum is taken over  $r \geq r_0 + \Delta$ . Applying the ULLN to  $P_n[e_t I(Y_{t-1} \leq r)]$  and  $P_n[e_t Y_{t-1} I(Y_{t-1} \leq r)]$ , it holds almost surely that for all  $\epsilon > 0$ , for  $n$  sufficiently large,

$$\begin{aligned} & |P_n[e_t(g_{\theta_0}(Y_{t-1}) - c_1 - b_1 Y_{t-1})I(Y_{t-1} \leq r)]| \\ & < \epsilon[(c_1 - c_{10})^2 + (b_1 - b_{10})^2 + (c_1 - c_{20})^2 + (b_1 - b_{20})^2]^{\frac{1}{2}}. \end{aligned}$$

We can then adapt an argument given in Chan (1993, pp.525-526) to show that  $\liminf_n P_n[f_{c_1, c_2, b_1, b_2, r}(Y_1, Y_2)] > \sigma^2$  as  $r$  sufficiently large. This shows that  $r_n \not\rightarrow \infty$  a.s.. A similar argument shows that  $r_n \not\rightarrow -\infty$  a.s..

Since  $r_n \not\rightarrow \pm\infty$ , there exists an  $M_1 > 0$  such that  $-M_1 < r_n < M_1$  for  $n$  sufficiently large a.s.. We will henceforth in this proof assume  $r_n \in [-M_1, M_1]$ . Using a similar argument, we can show that there exists an  $M_2 > 0$  such that  $(c_{1n}, c_{2n}, b_{1n}, b_{2n}) \in [-M_2, M_2]^4$  a.s..

Let  $M = \max(M_1, M_2)$  and  $C = [-M, M]^5$ . Then we have shown that the optimal  $(c_{1n}, c_{2n}, b_{1n}, b_{2n}, r_n)$  lies in  $C$ . Assuming there exists a finite class  $\mathcal{F}_\epsilon$  containing lower and upper approximations to  $\{f_{c_1, c_2, b_1, b_2, r} : (c_1, c_2, b_1, b_2, r) \in C\}$  in the manner as required by (C.1), then from Theorem 7,

$$\liminf_n \inf_C (P_n f_{c_1, c_2, b_1, b_2, r} - P f_{c_1, c_2, b_1, b_2, r}) \geq 0.$$

Thus

$$\liminf_n (W(c_{1n}, c_{2n}, b_{1n}, b_{2n}, r_n, P_n) - W(c_{1n}, c_{2n}, b_{1n}, b_{2n}, r_n, P)) \geq 0 \quad \text{a.s..}$$

Since

$$\begin{aligned} W(c_{1n}, c_{2n}, b_{1n}, b_{2n}, r_n, P_n) &\leq W(c_{10}, c_{20}, b_{10}, b_{20}, r_0, P_n) \\ &\rightarrow W(c_{10}, c_{20}, b_{10}, b_{20}, r_0, P) \quad \text{a.s..} \\ &\leq W(c_{1n}, c_{2n}, b_{1n}, b_{2n}, r_n, P), \end{aligned}$$

it follows that

$$W(c_{1n}, c_{2n}, b_{1n}, b_{2n}, r_n, P) \rightarrow W(c_{10}, c_{20}, b_{10}, b_{20}, r_0, P) \quad \text{a.s..}$$

Because  $W(\theta, P)$  is continuous in  $\theta$  and  $W(\theta, P) > W(\theta_0, P)$  for  $\theta \neq \theta_0$ , we deduce that  $(c_{1n}, c_{2n}, b_{1n}, b_{2n}, r_n)$  converges almost surely to  $(c_{10}, c_{20}, b_{10}, b_{20}, r_0)$ .

To complete the proof we need to construct the finite approximating class alluded to above. We first note that

$$f_{c_1, c_2, b_1, b_2, r}(y_1, y_2) \leq (|y_1| + M + M|y_2|)^2$$

for  $(c_1, c_2, b_1, b_2, r)$  in  $C$ . Write  $F(y_1, y_2)$  for the upper bound. Because  $PF < \infty$ , there exists a constant  $D$ , larger than  $M$ , for which  $PF([-D, D] \times [-D, D])^c < \infty$ . Hence, we shall restrict the domain of  $f_{c_1, c_2, b_1, b_2, r}$  to  $[-D, D]^2$ .

By suitably enlarging  $D$ , we may assume that  $(c_1, c_2, b_1, b_2, r) \in [-3D, 3D]^5$ . Let  $C_\epsilon$  be a finite subset of  $[-3D, 3D]^5$  such that each  $(c_1, c_2, b_1, b_2, r)$  in that subset has an  $(c_1^*, c_2^*, b_1^*, b_2^*, r^*)$  with  $\|c_1 - c_1^*\| < \epsilon/D^3$ ,  $\|c_2 - c_2^*\| < \epsilon/D^3$ ,  $\|b_1 - b_1^*\| < \epsilon/D^3$ ,  $\|b_2 - b_2^*\| < \epsilon/D^3$  and  $|r - r^*| < \epsilon/D$ . Then for each  $(y_1, y_2)$  in  $[-D, D]^2$  and, without loss of generality, assuming  $r < r^*$ ,

$$\begin{aligned}
& |f_{c_1, c_2, b_1, b_2, r, d}(y_1, y_2) - f_{c_1^*, c_2^*, b_1^*, b_2^*, r^*, d}(y_1, y_2)| \\
&= |(y_1 - c_1 - b_1 y_2)^2 I(y_2 \leq r) + (y_1 - c_2 - b_2 y_2)^2 I(y_2 > r) \\
&\quad - (y_1 - c_1^* - b_1^* y_2)^2 I(y_2 \leq r^*) - (y_1 - c_2 - b_2^* y_2)^2 I(y_2 > r^*)| \\
&\leq |(y_1 - c_1 - b_1 y_2)^2 I(y_2 \leq r) + (y_1 - c_2 - b_2 y_2)^2 I(y_2 > r^*)| \\
&\quad + (y_1 - c_2 - b_2 y_2)^2 I(r < y_2 \leq r^*) - (y_1 - c_1^* - b_1^* y_2)^2 I(y_2 \leq r) \\
&\quad - (y_1 - c_1^* - b_1^* y_2)^2 I(r < y_2 \leq r^*) - (y_1 - c_2 - b_2^* y_2)^2 I(y_2 > r^*)| \\
&\leq |2(c_1 - c_1^* + (b_1 - b_1^*)y_2)(y_1 - \frac{c_1 + c_1^*}{2} - \frac{b_1 + b_1^*}{2}y_2)I(y_2 \leq r)| \\
&\quad + |2(c_2 - c_2^* + (b_2 - b_2^*)y_2)(y_1 - \frac{c_2 + c_2^*}{2} - \frac{b_2 + b_2^*}{2}y_2)I(y_2 > r^*)| \\
&\quad + |2(c_1 - c_2 + (b_1^* - b_2)y_2)(y_1 - \frac{c_1^* + c_2}{2} - \frac{b_1^* + b_2}{2}y_2)I(r < y_2 < r^*)| \\
&\leq 2\frac{\epsilon}{D^3}D(4D + 3pD^2) + 2\frac{\epsilon}{D^3}D(4D + 3pD^2) \\
&\quad + 12pD^2(4D + 3pD^3)I(|y_2 - r| < |r^* - r|) \\
&\leq 7p\epsilon + 7p\epsilon \\
&= 14p\epsilon.
\end{aligned}$$

The class  $\mathcal{F}_\epsilon$  consists of all functions  $(f_{c_1^*, c_2^*, b_1^*, b_2^*, r^*}(y_1, y_2) - 14\epsilon)I(|y_1| \leq D, |y_2| \leq D)$  for  $(c_1^*, c_2^*, b_1^*, b_2^*, r^*)$  ranging over  $C_\epsilon$ . This completes the proof.

**Proof of Theorem 3.** Since  $\Gamma$  is known, we shall, without loss of generality, assume that  $\Gamma = I$ . To simplify our proofs for Theorem 3 and 4, we consider the case  $m = 2$  and  $p = 1$  as the proof for the general case is similar and hence omitted. In practice, we may and shall assume  $d$  to be known. (The proof here can be readily generalized to the case of unknown  $d$ .) Let  $n = 6$  and  $r = 1$ . Define

$$f_{(A,B),\gamma}(y, x) = \text{tr}[(y - ABx)(y - ABx)'].$$

The optimal  $\hat{A}_T, \hat{B}_T$  and  $\hat{\gamma}_T$  are chosen to minimize

$$\begin{aligned} W((A, B), \gamma, P_T) &= P_T f_{(A, B), \gamma} \\ &= \frac{1}{T} \sum_t f_{(A, B), \gamma}(y_t, x_t). \end{aligned}$$

where  $P_T$  denotes the empirical measure of  $(Y_t, X_t)$  based on the observations  $(Y_1, X_1), \dots, (Y_T, X_T)$ , with  $A$  and  $B$  satisfying the normalization conditions.

The finiteness of  $W((\cdot, \cdot), \cdot, P)$  follows from the condition that  $\text{Ptr}(y_t y_t') < \infty$ . Because  $(Y_t, X_t)$  has a positive density everywhere, the true parameter  $((\text{vec}(A_0)', \text{vec}(B_0)'), \gamma_0)'$  is the unique argument minimizing  $W((\cdot, \cdot), \cdot, P)$ . Applying a similar argument as in the proof of Theorem 8, we can show that  $((\text{vec}(\hat{A}_T)', \text{vec}(\hat{B}_T)'), \hat{\gamma}_T)'$ , which minimizes  $W((A, B), \gamma, P_T)$ , falls into the region

$$C = [-M, M]^{r(m+n-r)} \times M^2$$

for some suitably large  $M$ .

We claim that the collection of  $f_{(A, B), \gamma}(y_1, y_2)$  over  $C$  admits a finite approximation as stated in Theorem 7. Assuming this claim, we can deduce using arguments similar to those employed in the proof of Theorem 8 that

$$W((\hat{A}_T, \hat{B}_T), \hat{\gamma}_T, P) \rightarrow W((A_0, B_0), \gamma_0, P) \quad \text{a.s.}$$

Because  $W((A, B), \gamma, P)$  is continuous over  $C$  and the true parameter  $((A_0, B_0), \gamma_0)$  is the unique argument minimizing  $W((\cdot, \cdot), \cdot, P)$ , we have the consistency of  $((\text{vec}(\hat{A}_T)', \text{vec}(\hat{B}_T)'), \hat{\gamma}_T)'$ . The weak consistency of  $\hat{\Sigma}$  to  $\Sigma$  follows from the consistency of the other parameter estimators.

It remains to verify the claim on the finite approximating class for  $f_{(A, B), \gamma}(y_1, y_2)$  over  $C$ . To construct the finite approximating class, we first note that

$$f_{(A, B), \gamma}(y_1, y_2) \leq 2\text{tr}[(|y_1| + rM^2|y_2|)(|y_1| + rM^2|y_2|)']$$

for  $((\text{vec}(A)', \text{vec}(B)'), \gamma)'$  in  $C$ . The notation  $|y_i|$  is defined as  $|y_i| = (|y_{1i}|, |y_{2i}|)'$ . Write  $F(y_1, y_2)$  for the preceding upper bound. Because  $PF < \infty$ , there exists a constant  $D$ , larger than  $M$ , such that  $PF([-D, D]^m \times [-D, D]^n)^c < \infty$ . Hence, we need only consider the approximation of  $f_{(A, B), \gamma}$  over  $[-D, D]^m \times [-D, D]^n$ .

Without loss of generality, we assume that  $((\text{vec}(A)', \text{vec}(B)'), \gamma)'$   $\in [-3D, 3D]^{r(m+n-r)} \times [-3D, 3D]^2$ . Let  $C_\epsilon$  be a finite subset of  $S = [-3D, 3D]^{r(m+n-r)} \times [-3D, 3D]^2$  such that for all  $((\text{vec}(A)', \text{vec}(B)'), \gamma)'$   $\in S$  there exists  $((\text{vec}(A^*)', \text{vec}(B^*)'), \gamma^*)' \in C_\epsilon$  such that

$$\begin{aligned} \|\text{vec}(A) - \text{vec}(A^*)\| &< \frac{\epsilon}{rD^5}, \\ \|\text{vec}(B) - \text{vec}(B^*)\| &< \frac{\epsilon}{rD^5}, \\ \|\gamma - \gamma^*\| &< \frac{\epsilon}{rD^5}. \end{aligned}$$

Write

$$f^{(i)} = y - ABx^{(i)}, \quad g^{(i)} = y - A^*B^*x^{(i)}$$

and  $X^{(1)}$  to  $X^{(4)}$  are defined in (5.5) to (5.8). Then for each  $y \in [-D, D]^2$  and  $x = (x_1, x_2, x_3, x_4, x_5, x_6)' \in [-D, D]^6$ ,

$$\begin{aligned} & |f_{A,B,\gamma}(y, x) - f_{A^*,B^*,\gamma^*}(y, x)| \\ &= |\text{tr}[(y - ABx)(y - ABx)'] - \text{tr}[(y - A^*B^*x)(y - A^*B^*x)']| \\ &= |\text{tr}(f^{(1)}f^{(1)'})I(x_1 > \gamma_1, x_2 > \gamma_2) + \text{tr}(f^{(2)}f^{(2)'})I(x_1 \leq \gamma_1, x_2 > \gamma_2) \\ &\quad + \text{tr}(f^{(3)}f^{(3)'})I(x_1 \leq \gamma_1, x_2 \leq \gamma_2) + \text{tr}(f^{(4)}f^{(4)'})I(x_1 > \gamma_1, x_2 \leq \gamma_2) \\ &\quad - \text{tr}(g^{(1)}g^{(1)'})I(x_1 > \gamma_1^*, x_2 > \gamma_2^*) - \text{tr}(g^{(2)}g^{(2)'})I(x_1 \leq \gamma_1^*, x_2 > \gamma_2^*) \\ &\quad - \text{tr}(g^{(3)}g^{(3)'})I(x_1 \leq \gamma_1^*, x_2 \leq \gamma_2^*) - \text{tr}(g^{(4)}g^{(4)'})I(x_1 > \gamma_1^*, x_2 \leq \gamma_2^*)|. \end{aligned} \quad (\text{C.4})$$

Without loss of generality, assume that  $\gamma_1 \leq \gamma_1^*, \gamma_2 \leq \gamma_2^*$ . Below we will show that the RHS of (C.4) is bounded by some multiple of  $\epsilon$ . It is clear that the RHS of (C.4) equals to  $\text{tr}(f^{(i)}f^{(i)'}) - \text{tr}(g^{(j)}g^{(j)'})$  for some  $1 \leq i, j \leq 4$  dependent on which two of the indicator functions equal 1. We provide the proof for the case  $x_1 > \gamma_1^*, x_2 > \gamma_2^*$  and omit the similar proofs for the other cases. Consider

$$\begin{aligned} & |\text{tr}[(y - ABx^{(1)})(y - ABx^{(1)})']I(x_1 > \gamma_1, x_2 > \gamma_2) \\ &\quad - \text{tr}[(y - A^*B^*x^{(1)})(y - A^*B^*x^{(1)})']I(x_1 > \gamma_1^*, x_2 > \gamma_2^*)| \\ &\leq |(y - ABx^{(1)})'(y - ABx^{(1)}) - (y - A^*B^*x^{(1)})'(y - A^*B^*x^{(1)})| \\ &= |(y - ABx^{(1)})'(y - ABx^{(1)}) - (y - A^*B^*x^{(1)})'(y - ABx^{(1)}) \\ &\quad + (y - A^*B^*x^{(1)})'(y - ABx^{(1)}) - (y - A^*B^*x^{(1)})'(y - A^*B^*x^{(1)})| \\ &= |[A^*B^* - AB]x^{(1)'}(y - ABx^{(1)}) + (y - A^*B^*x^{(1)})'(A^*B^* - AB)x^{(1)}| \\ &= |x^{(1)'}[A^*(B^* - B) - (A - A^*)B]'(y - ABx^{(1)}) \\ &\quad + (y - A^*B^*x^{(1)})'[A^*(B^* - B) - (A - A^*)B]x^{(1)}| \\ &\leq 2\sqrt{nD^2}(2\sqrt{rnD^2}\frac{\epsilon}{D^5})(\sqrt{2D^2} + \sqrt{2rD^2}\sqrt{nrD^2}\sqrt{nD^2}) \\ &\leq 6rn\sqrt{p}\epsilon. \end{aligned}$$

To sum up, we have shown that the RHS of (C.4) is  $\leq k\epsilon$  for some  $k > 0$ . The class  $\mathcal{F}_\epsilon$  consists of all functions  $(f_{(A^*,B^*),\gamma^*}(y, x) - k\epsilon)I(y \in [-D, D]^2, x \in [-D, D]^n)$  for  $((\text{vec}(A^*), \text{vec}(B^*)), \gamma^{*\prime})'$  ranging over  $C_\epsilon$ . This completes the proof.

#### Appendix D. Proof of Theorem 4

Since  $\hat{\theta}_T$  is strongly consistent, without loss of generality, the parameter space can be restricted to a neighborhood of  $\theta_0 = (A_0, B_0, \gamma_0)$ , say,

$$w(\Delta) = \{\theta \in \Omega : |A - A_0| < \Delta, |B - B_0| < \Delta \text{ and } |z - \gamma_0| < \Delta\},$$

for some  $0 < \Delta < 1$  to be determined below. For simplicity of proof, we assume  $\gamma_0 = 0$ . Then it suffices to verify the following claims.

**Claim 1.** For all  $\epsilon > 0$ , there exists a constant, say,  $K$  such that with probability greater than  $1 - \epsilon$ ,  $\theta \in w(\Delta)$  and the  $L^2$ -norm  $\|z\| > K/T$  implies that  $L_T(A, B, z) - L_T(A, B, 0) > 0$ .

First, consider the case that  $z_1 > 0, z_2 > 0$ . Define  $f_t^{(i)} = Y_t - ABX_t^{(i)}, i = 1, \dots, 4$ ; see the notation  $X^{(i)}$  defined in (5.5)-(5.8). Then

$$\begin{aligned}
& L_T(A, B, z) - L_T(A, B, 0) \\
&= \sum \text{tr}(f_t^{(1)} f_t^{(1)'}) I(Y_{1,t-1} > z_1, Y_{2,t-1} > z_2) \\
&\quad + \sum \text{tr}(f_t^{(2)} f_t^{(2)'}) I(Y_{1,t-1} \leq z_1, Y_{2,t-1} > z_2) \\
&\quad + \sum \text{tr}(f_t^{(3)} f_t^{(3)'}) I(Y_{1,t-1} \leq z_1, Y_{2,t-1} \leq z_2) \\
&\quad + \sum \text{tr}(f_t^{(4)} f_t^{(4)'}) I(Y_{1,t-1} > z_1, Y_{2,t-1} \leq z_2) \\
&\quad - \sum \text{tr}(f_t^{(1)} f_t^{(1)'}) I(Y_{1,t-1} > 0, Y_{2,t-1} > 0) \\
&\quad - \sum \text{tr}(f_t^{(2)} f_t^{(2)'}) I(Y_{1,t-1} \leq 0, Y_{2,t-1} > 0) \\
&\quad - \sum \text{tr}(f_t^{(3)} f_t^{(3)'}) I(Y_{1,t-1} \leq 0, Y_{2,t-1} \leq 0) \\
&\quad - \sum \text{tr}(f_t^{(4)} f_t^{(4)'}) I(Y_{1,t-1} > 0, Y_{2,t-1} \leq 0) \\
&= \sum \text{tr}(f_t^{(2)} f_t^{(2)' } - f_t^{(1)} f_t^{(1)'}) I(0 < Y_{1,t-1} \leq z_1, Y_{2,t-1} > z_2) \\
&\quad + \sum \text{tr}(f_t^{(4)} f_t^{(4)' } - f_t^{(1)} f_t^{(1)'}) I(Y_{1,t-1} > z_1, 0 < Y_{2,t-1} \leq z_2) \\
&\quad + \sum \text{tr}(f_t^{(3)} f_t^{(3)' } - f_t^{(4)} f_t^{(4)'}) I(0 < Y_{1,t-1} \leq z_1, Y_{2,t-1} \leq 0) \\
&\quad + \sum \text{tr}(f_t^{(3)} f_t^{(3)' } - f_t^{(1)} f_t^{(1)'}) I(0 < Y_{1,t-1} \leq z_1, 0 < Y_{2,t-1} \leq z_2) \\
&\quad + \sum \text{tr}(f_t^{(3)} f_t^{(3)' } - f_t^{(2)} f_t^{(2)'}) I(Y_{1,t-1} \leq 0, 0 < Y_{2,t-1} \leq z_2). \tag{D.1}
\end{aligned}$$

The first term on the RHS of (D.1) can be simplified as follows:

$$\begin{aligned}
& \sum \text{tr}(f_t^{(2)} f_t^{(2)' } - f_t^{(1)} f_t^{(1)'}) I(0 < Y_{1,t-1} \leq z_1, Y_{2,t-1} > z_2) \\
&= \sum \text{tr}[(Y_t - ABX_t^{(2)})(Y_t - ABX_t^{(2)})' - (Y_t - ABX_t^{(1)})(Y_t - ABX_t^{(1)})'] \\
&\quad \times I(0 < Y_{1,t-1} \leq z_1, Y_{2,t-1} > z_2) \\
&= \sum \text{tr}[(e_t + A_0 B_0 X_t^{(1)} - ABX_t^{(2)})(e_t + A_0 B_0 X_t^{(1)} - ABX_t^{(2)})' \\
&\quad - (e_t + (A_0 B_0 - AB)X_t^{(1)})(e_t + (A_0 B_0 - AB)X_t^{(1)})'] \\
&\quad \times I(0 < Y_{1,t-1} \leq z_1, Y_{2,t-1} > z_2)
\end{aligned}$$

$$\begin{aligned}
&= \sum tr[2e_t(A_0B_0X_t^{(1)} - ABX_t^{(2)})' + (A_0B_0X_t^{(1)} - ABX_t^{(2)})(A_0B_0X_t^{(1)} - ABX_t^{(2)})' \\
&\quad - 2e_t((A_0B_0 - AB)X_t^{(1)})' - ((A_0B_0 - AB)X_t^{(1)})((A_0B_0 - AB)X_t^{(1)})'] \\
&\quad \times I(0 < Y_{1,t-1} \leq z_1, Y_{2,t-1} > z_2) \\
&= \sum tr[2e_t(AB(X_t^{(1)} - X_t^{(2)}))']I(0 < Y_{1,t-1} \leq z_1, Y_{2,t-1} > z_2) \\
&\quad - 2 \sum [A_0B_0X_t^{(1)}(ABX_t^{(2)})']I(0 < Y_{1,t-1} \leq z_1, Y_{2,t-1} > z_2) \\
&\quad + \sum tr[ABX_t^{(2)}(ABX_t^{(2)})']I(0 < Y_{1,t-1} \leq z_1, Y_{2,t-1} > z_2) \\
&\quad + 2 \sum tr[A_0B_0X_t^{(1)}(ABX_t^{(1)})']I(0 < Y_{1,t-1} \leq z_1, Y_{2,t-1} > z_2) \\
&\quad - \sum tr[ABX_t^{(1)}(ABX_t^{(1)})']I(0 < Y_{1,t-1} \leq z_1, Y_{2,t-1} > z_2) \\
&= \sum tr[2e_t(AB(X_t^{(1)} - X_t^{(2)}))']I(0 < Y_{1,t-1} \leq z_1, Y_{2,t-1} > z_2) \\
&\quad + 2 \sum tr[A_0B_0X_t^{(1)}(ABX_t^{(1)} - ABX_t^{(2)})']I(0 < Y_{1,t-1} \leq z_1, Y_{2,t-1} > z_2) \\
&\quad - \sum tr[ABX_t^{(1)}(ABX_t^{(1)})' - ABX_t^{(2)}(ABX_t^{(2)})'] \\
&\quad \times I(0 < Y_{1,t-1} \leq z_1, Y_{2,t-1} > z_2). \tag{D.2}
\end{aligned}$$

If  $\Delta$  is sufficiently small, then it follows from Condition 4 that the sum of the second and third terms is greater than or equal to  $\delta^2 \sum I(Y_{1,t-1} \leq 0, 0 < Y_{2,t-1} \leq z_2)$ , because

$$\begin{aligned}
&\sum tr[2A_0B_0X_t^{(1)}(ABX_t^{(1)} - ABX_t^{(2)})']I(0 < Y_{1,t-1} \leq z_1, Y_{2,t-1} > z_2) \\
&\quad - \sum tr[ABX_t^{(1)}(ABX_t^{(1)})' - ABX_t^{(2)}(ABX_t^{(2)})']I(0 < Y_{1,t-1} \leq z_1, Y_{2,t-1} > z_2) \\
&\approx \sum tr[A_0B_0X_t^{(1)}(A_0B_0X_t^{(1)})' - 2A_0B_0X_t^{(1)}(A_0B_0X_t^{(2)})' \\
&\quad + A_0B_0X_t^{(2)}(A_0B_0X_t^{(2)})']I(0 < Y_{1,t-1} \leq z_1, Y_{2,t-1} > z_2) \\
&= \sum tr[A_0B_0(X_t^{(1)}X_t^{(1)'} + X_t^{(2)}X_t^{(2)'} - 2X_t^{(1)}X_t^{(2)'})B_0'A_0'] \\
&\quad \times I(0 < Y_{1,t-1} \leq z_1, Y_{2,t-1} > z_2) \\
&\geq \delta^2 \sum I(0 < Y_{1,t-1} \leq z_1, Y_{2,t-1} > z_2).
\end{aligned}$$

The first term on the RHS of (D.2) is bounded in absolute value by  $\nu |\sum e_t'(X_t^{(1)} - X_t^{(2)})I(0 < Y_{1,t-1} \leq z_1, Y_{2,t-1} > z_2)|$  for some constant  $\nu$  independent of  $T$ . Define

$$Q(z_1) = EI(0 < Y_{1,t-1} \leq z_1) \tag{D.3}$$

and it is clear that  $Q(z_1) = O(z_1)$  for small  $z_1$ .



**Claim 2.** For any  $\epsilon > 0, \eta > 0$ , there exists  $K > 0$  such that, for all  $T$ ,

$$\begin{aligned} & P\left(\sup_{\substack{K/T < z_1 \leq \Delta \\ 0 < z_2 \leq \Delta}} \left| \sum \frac{I(0 < Y_{1,t-1} \leq z_1, Y_{2,t-1} > z_2)}{TQ(z_1)} - 1 \right| < \eta\right) > 1 - \epsilon \\ & P\left(\sup_{\substack{K/T < z_1 \leq \Delta \\ 0 < z_2 \leq \Delta}} \left| \sum \text{tr}[e_t X_t^{(1)'}] \frac{I(0 < Y_{1,t-1} \leq z_1, Y_{2,t-1} > z_2)}{TQ(z_1)} \right| < \eta\right) > 1 - \epsilon \\ & P\left(\sup_{\substack{K/T < z_1 \leq \Delta \\ 0 < z_2 \leq \Delta}} \left| \sum \text{tr}[e_t X_t^{(2)'}] \frac{I(0 < Y_{1,t-1} \leq z_1, Y_{2,t-1} > z_2)}{TQ(z_1)} \right| < \eta\right) > 1 - \epsilon. \end{aligned}$$

Suppose the above claim is valid. Let  $\epsilon > 0$  be given and  $\eta > 0$  be chosen so that  $-2\nu\eta + \delta^2(1 - \eta) > 0$ . It follows from the preceding claim that there exists a  $K(\epsilon, \eta) > 0$  such that with probability greater than  $1 - 3\epsilon$ ,  $K/T < z_1 \leq \Delta$  implies that

$$\begin{aligned} & \sum \text{tr}(f_t^{(2)} f_t^{(2)'} - f_t^{(1)} f_t^{(1)'}) \frac{I(0 < Y_{1,t-1} \leq z_1, Y_{2,t-1} > z_2)}{TQ(z_1)} \\ & \geq -2\nu\eta + \delta^2(1 - \eta) > 0. \end{aligned}$$

which verifies Claim 1 under the condition that  $z_1 > 0, z_2 > 0$ . The other cases, namely the 2nd, 3rd and 5th term of the RHS of (D.1) are similar and hence omitted with the 4th term being negligible compared to the other term.

Finally, note that claim 2 can be proved by making use of Conditions 1 to 4 and employing arguments as in Chan (1993, pp.528-529).

## Appendix E. Proof of Theorem 5

The proof is similar to the case of reduced-rank linear regression, which relies on the use of perturbation expansion of matrices and the limiting behavior of  $T^{1/2}\text{vec}(U_T)$  where  $U_T = T^{-1} \sum_t X_t \epsilon_t'$ ; see Reinsel and Velu (1998, pp.42-44). Since the perturbation expansion of matrices is the same whether or not the threshold parameter is known, we need only consider the limiting distribution of  $T^{1/2}\text{vec}(U_T)$ .

If  $X_t$  is free of unknown parameters, it is known that

$$T^{1/2}\text{vec}(U_T) \xrightarrow{D} N(0, \Sigma_{\epsilon\epsilon} \otimes \Sigma_{xx}) \quad \text{as } T \rightarrow \infty. \quad (\text{E.1})$$

See Anderson (1971, p.200) for a proof for the case when  $(Y_t)$  is a linear process. In the general case, the result follows from the martingale CLT (Hamilton (1994)). We shall show that (E.1) holds even if the unknown parameter  $\gamma$  in  $X_t$  is replaced by super-consistent estimator  $\hat{\gamma}$ . Then we can mimic the proof in

Reinsel and Velu (1998) to prove Theorem 5. The rest of this proof is devoted to verifying the preceding claim (E.1). For simplicity, we assume  $p = d = 1$  and define the following random variables:

$$W_t = \text{vec}(X_t \epsilon'_t) = \text{vec} \left( \begin{pmatrix} Y_{t-1} \\ I(Y_{t-1} > \hat{\gamma}) \\ Y_{t-1} I(Y_{t-1} > \hat{\gamma}) \end{pmatrix} \epsilon'_t \right),$$

$$W_t^* = \text{vec}(X_t^* \epsilon'_t) = \text{vec} \left( \begin{pmatrix} Y_{t-1} \\ I(Y_{t-1} > \gamma) \\ Y_{t-1} I(Y_{t-1} > \gamma) \end{pmatrix} \epsilon'_t \right).$$

So,

$$\begin{aligned} \sqrt{T} \text{vec}(U_T) &= \frac{1}{T^{1/2}} \sum_t W_t \\ &= \frac{1}{T^{1/2}} \sum_t W_t^* + \frac{1}{T^{1/2}} \sum_t \text{vec} \left( \begin{pmatrix} 0 \\ I(Y_{t-1} > \hat{\gamma}) - I(Y_{t-1} > \gamma) \\ Y_{t-1} [I(Y_{t-1} > \hat{\gamma}) - I(Y_{t-1} > \gamma)] \end{pmatrix} \epsilon'_t \right). \end{aligned} \quad (\text{E.2})$$

Clearly,  $(1/T) \sum_t W_t^* \xrightarrow{D} N(0, \Sigma_{\epsilon\epsilon} \otimes \Sigma_{xx})$ . We shall show the second term of RHS of (E.2) converges to 0 in probability; hence the proof is done by appealing to Slutsky's theorem.

Let  $V_t = [I(Y_{i,t-1} > \hat{\gamma}_i) - I(Y_{i,t-1} > \gamma_i)] \epsilon_{j,t}$  where  $Y_{i,t}$  denotes the  $i$ th element of vector  $Y_t$  and so  $\epsilon_{j,t}$  the  $j$ th element of  $\epsilon_t$ . Let  $f_i(\cdot)$  be the stationary density function of  $Y_{i,1}$ . Since  $\hat{\gamma}_i = \gamma_i + O_p(1/T)$ , without loss of generality, we assume that  $|\hat{\gamma}_i - \gamma_i| \leq M/T$  for some  $M > 0$ . Then

$$\begin{aligned} E|T^{-1/2} \sum_t V_t| &\leq \frac{1}{T} \sum_{t=1}^T E|V_t| \\ &\leq T^{-1/2} \sum_{t=1}^T E|I(Y_{i,t-1} > \hat{\gamma}_i) - I(Y_{i,t-1} > \gamma_i)| |\epsilon_{j,t}| \\ &\leq T^{-1/2} \sum_{t=1}^T EI(\gamma_i - \frac{M}{T} < Y_{i,t-1} < \gamma_i + \frac{M}{T}) |\epsilon_{j,t}| \\ &= T^{-1/2} \sum_{t=1}^T EI(\gamma_i - \frac{M}{T} < Y_{i,t-1} < \gamma_i + \frac{M}{T}) E|\epsilon_{j,t}| \\ &= T^{-1/2} \sum_{t=1}^T \int_{\gamma_i - \frac{M}{T}}^{\gamma_i + \frac{M}{T}} f_i(y_i) dy_i E|\epsilon_{j,t}| \\ &= O\left(\frac{1}{T^{1/2}}\right) \rightarrow 0 \quad \text{as } T \rightarrow \infty, \end{aligned}$$

where  $E|\epsilon_{j,t}| < \infty$  by Condition 2. Consequently,  $T^{-1/2} \sum_t \text{vec}((I(Y_{t-1} > \gamma) - I(Y_{t-1} > \hat{\gamma}))\epsilon_t) \xrightarrow{p} 0$ . Similarly, it can be shown that  $T^{-1/2} \sum_t \text{vec}(Y_{t-1}[I(Y_{t-1} > \hat{\gamma}) - I(Y_{t-1} > \gamma)]\epsilon'_t) \xrightarrow{p} 0$ .

## References

- Anderson, T. W. (1971). *The Statistical Analysis of Time Series*. Wiley, New York.
- Chan, K. S. (1993). Consistency and limiting distribution of the least squares estimator of a threshold autoregressive model. *Ann. Statistics* **21**, 520-533.
- Chan, K. S. and Tong, H. (1985). On the use of the deterministic Lyapunov function for the ergodicity of stochastic difference equations. *Adv. Appl. Probab.* **17**, 666-678.
- Hamilton, J. D. (1994). *Time Series Analysis*. Princeton, Princeton University Press, New Jersey.
- Pollard, D. (1984). *Convergence of Stochastic Processes*. Springer-Verlag, New York.
- Reinsel, G. C. and Velu, R. (1998). *Multivariate-Reduced Rank Regression*. Springer-Verlag, New York.