# ANÁLISE BAYESIANA DE MODELOS THRESHOLDS UTILIZANDO O SOFTWARE R

Luciene Resende Gonçalves [1], Thelma Sáfadi [2], Anderson Castro Soares Oliveira [3]

## RESUMO

O modelo "Open loop threshold autoregressive" - TARSO é um tipo de modelo usado para modelar séries não lineares. Esses modelos são úteis na detecção de propriedades observadas em séries temporais como ciclos limites e "jump phenomena". O progresso computacional vem facilitando a utilização desses modelos em larga escala. Este trabalho se propõe, por meio de um estudo de simulação, estabelecer uma rotina computacional em R para análise bayesiana de modelos threshold.

## INTRODUÇÃO

O modelo "Open loop threshold autoregressive" - TARSO é utilizado para modelar séries não lineares. Os modelos não lineares têm sido objeto de muito estudo nos últimos anos em razão dos modelos lineares usuais falharem em detectar certas propriedades observadas em séries temporais tais como ciclos limites e "jump phenomena". O progresso computacional vem facilitando a utilização desses modelos em larga escala.

Dentre os diversos modelos não lineares propostos para séries com tais propriedades podemos citar os modelos "threshold", os bilineares e os auto-regressivos com coeficientes aleatórios. Esses modelos possuem larga aplicação em ecologia (estudos de população), física solar, entomologia (dinâmica de população), economia, finanças, saúde, poluição, geofísica e hidrologia.

Tong (1977) foi pioneiro em modelar uma série temporal discreta usando modelos lineares por partes. Tong e Lim (1980) fizeram diversas aplicações e discussões a respeito desses modelos. O uso de variáveis e valores "threshold"permite representar uma série temporal por meio de modelos lineares por partes dividindo o conjunto de dados em grupos que podem ser chamados de classes ou regimes.

Safadi e Morettin (2001) desenvolveram a análise bayesiana do modelo TARSO e exemplificaram a metodologia nas séries do número de óbitos causados por problemas cardíacos e na série de temperatura mínima diária na cidade de São Paulo. Observaram que, para temperaturas entre $12,90$ e $15,23$ graus, há um acréscimo de 300 óbitos causados por problemas cardíacos.

O objetivo deste trabalho é fazer uma extensão do modelo utilizado em Safadi e Morettin (2001) considerando um número maior de variáveis dependentes e desenvolver uma rotina em R para o ajuste do modelo sob o enfoque Bayesiano.

[1] Doutoranda em Estatística e Experimentação Agropecuária, luciene.goncalves@unifal-mg.edu.br
[2] Professor Associado II do Departamento de Ciências Exatas, safadi@ufla.br
[3] Professor Adjunto da UFMT, ancasli@yahoo.com.br

**Modelo TARSO** $(l; p_l; m_l; n_l)$

Suponha que $\{Y_t, V_t, t = 0, \pm1, \pm2, \ldots\}$ seja um processo estocástico satisfazendo

$$Y_t \quad = \quad \phi_{l_0} + V_t + a_t^{(l)} + \sum_{i=1}^{p_l} \phi_{l_i} Y_{t-i}, \quad se \quad Y_{t-d} \; \in R_l; \;\; l = 1, \ldots, k, \tag{1}$$

indica o número de regimes. Neste modelo $Y_t$ é uma série de dados dependentes e estacionários e

$$V_t \quad = \quad \sum_{\nu=0}^{m_l} \beta_{l\nu} Z_{t-\nu} + \sum_{u=0}^{n_l} \theta_{lu} T_{t-u} \tag{2}$$

é a soma de variáveis independentes estacionárias $Z_t$ e $T_t$. $\phi_{l0}, \phi_{li}, \beta_{l\nu}$ e $\theta_{lu}$ são parâmetros e $p_l$, $m_l$ e $n_l$ são as ordens das variáveis e $l$ é o número de regimes. $\{a_t^{(l)}\}$ é uma sequência de variáveis aleatórias i.i.d. $\sim N(0, \tau_l^{-1})$, em que $\tau_l > 0, \tau_l^{-1} = \sigma_l^2 = Var(a_t^{(l)})$ é desconhecida, $Z_{t-d}$ é uma variável "threshold", independente de $a^{(l)}$, para todo $l$ e todo $t$, $d$ é o parâmetro de retardo. $Z_{t-d}$ determina os regimes do modelo. As $k$ sequências são supostas serem independentes entre si.

O modelo TARSO, considerando 2 regimes ($k = 2$) é descrito por:

$$Y_t = \begin{cases} \phi_{10} + \sum_{i=1}^{p_1} \phi_{1i} Y_{t-i} + \sum_{\nu=0}^{m_1} \beta_{1\nu} Z_{t-\nu} + \sum_{u=0}^{n_1} \theta_{1u} T_{t-u} + a_t^{(1)} & \text{, se } Z_{t-d} \leq r; \\ \phi_{20} + \sum_{i=1}^{p_2} \phi_{2i} Y_{t-i} + \sum_{\nu=0}^{m_2} \beta_{2\nu} Z_{t-\nu} + \sum_{u=0}^{n_2} \theta_{2u} T_{t-u} + a_t^{(2)} & \text{, se } Z_{t-d} > r. \end{cases} \tag{3}$$

Supondo conhecidas as ordens $(p_1, p_2, m_1, m_2, n_1, n_2)$ os parâmetros para o modelo são

$$\gamma_j = (\phi_{j0}, \phi_{j1}, \phi_{j2}, \ldots, \phi_{jpj}, \beta_{jo}, \beta_{j1}, \ldots, \beta_{jm_j}, \theta_{jo}, \theta_{j1}, \ldots, \theta_{jqj},), \quad \tau_j, \;\; j = 1, 2, \;\; r \;\; e \;\; d.$$

$r$ indica o valor threshold.

Fazendo $X_{jt} = (1, Y_{t-1}, \ldots, Y_{t-pj}, Z_t, Z_{t-1}, \ldots, Z_{t-mj}, T_t, T_{t-1}, \ldots, T_{t-nj})'$, $j = 1, 2$, a equação (3) pode ser reescrita como

$$Y_t = \begin{cases} \gamma_1' X_{1t} + a_t^{(1)} & \text{, se } \; Z_{t-d} \leq r; \\ \gamma_2' X_{2t} + a_t^{(2)} & \text{, se } \; Z_{t-d} > r. \end{cases} \tag{4}$$

A função de verossimilhança aproximada condicionada às $p$-primeiras observações é dada por

$$\mathcal{L}(\gamma_1, \gamma_2, \tau_1, \tau_2, r, d|D) \quad \propto \quad \tau_1^{\frac{b_1}{2}} \tau_2^{\frac{b_2}{2}} exp\{-\frac{\tau_1}{2} \sum_1 (Y_t - \gamma_1' X_{1t})^2 - \frac{\tau_2}{2} \sum_2 (Y_t - \gamma_2' X_{2t})^2\}$$

em que $D = \{Y_t, Z_t, T_t, t = p+1, \ldots, b\}$ é o conjunto de todas as observações, $b_1, b_2$ são os números de observações em cada regime, $\sum_1$ é a somatória em $t$ para $\{t = p + 1, \ldots, b; Z_{t-d} \leq r\}$, $\sum_2$ é a soma em $t$ para $\{t = p +$

$1,\ldots,b; Z_{t-d} > r\}$ sendo $p = max\{p_1, p_2, m_1, m_2, n_1, n_2\}$, isto é , o máximo das ordens. Tem-se que

$$\begin{aligned}\sum_j (Y_t - \gamma_j^{'} X_{jt})^2 &= \sum_j (Y_t - \gamma_j^{'} X_{jt})(Y_t - \gamma_j^{'} X_{jt})^{'} = \sum_j Y_t^2 - \sum_j L_t X_{jt}^{'} \gamma_j \\ &\quad -\gamma_j^{'} \sum_j X_{jt} L_t + \gamma_j^{'} \sum_j X_{jt} X_{jt}^{'} \gamma_j, \quad j = 1, 2.\end{aligned}$$

Denotando por

$$\sum_j X_{jt} Y_t = \mathbf{B}_j = \begin{pmatrix} \mathbf{B}_{1j} \\ \mathbf{B}_{2j} \\ \mathbf{B}_{3j} \end{pmatrix}, \quad \sum_j X_{jt} X_{jt}^{'} = \mathbf{A}_j = \begin{pmatrix} \mathbf{A}_{11}^j & \mathbf{A}_{12}^j & \mathbf{A}_{13}^j \\ \mathbf{A}_{21}^j & \mathbf{A}_{22}^j & \mathbf{A}_{23}^j \end{pmatrix}, j = 1, 2,$$

em que

$$\mathbf{B}_{1j} = \begin{pmatrix} \sum_j Y_t \\ \sum_j Y_t Y_{t-1} \\ \vdots \\ \sum_j Y_t Y_{t-pj} \end{pmatrix}, \mathbf{B}_{2j} = \begin{pmatrix} \sum_j Y_t Z_t \\ \sum_j Y_t Z_{t-1} \\ \vdots \\ \sum_j Y_t Z_{t-mj} \end{pmatrix}, \mathbf{B}_{3j} = \begin{pmatrix} \sum_j Y_t T_t \\ \sum_j Y_t T_{t-1} \\ \vdots \\ \sum_j Y_t T_{t-nj} \end{pmatrix}$$

e

$$\mathbf{A}_{11}^j = \begin{pmatrix} \sum_j 1 & \sum_j Y_{t-1} & \ldots & \sum_j Y_{t-pj} \\ \sum_j Y_{t-1} & \sum_j Y_{t-1} Y_{t-1} & \ldots & \sum_j Y_{t-1} Y_{t-pj} \\ \vdots & \vdots & \vdots & \vdots \\ \sum_j Y_{t-pj} & \sum_j Y_{t-1} L_{t-pj} & \ldots & \sum_j Y_{t-pj} Y_{t-pj} \end{pmatrix},$$

$$\mathbf{A}_{12}^j = \begin{pmatrix} \sum_j Z_t & \sum_j Z_{t-1} & \sum_j Z_{t-2} & \ldots & \sum_j Z_{t-mj} \\ \sum_j Y_{t-1} Z_t & \sum_j Y_{t-1} Z_{t-1} & \sum_j Y_{t-1} Z_{t-2} & \ldots & \sum_j Y_{t-1} Z_{t-mj} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum_j Y_{t-pj} Z_t & \sum_j Y_{t-pj} Z_{t-1} & \sum_j Y_{t-pj} Z_{t-2} & \ldots & \sum_j Y_{t-pi} Z_{t-mj} \end{pmatrix},$$

$$\mathbf{A}_{13}^j = \begin{pmatrix} \sum_j T_t & \sum_j T_{t-1} & \sum_j T_{t-2} & \ldots & \sum_j T_{t-nj} \\ \sum_j Y_{t-1} T_t & \sum_j Y_{t-1} T_{t-1} & \sum_j Y_{t-1} T_{t-2} & \ldots & \sum_j Y_{t-1} T_{t-nj} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum_j Y_{t-pj} T_t & \sum_j Y_{t-pj} T_{t-1} & \sum_j Y_{t-pj} T_{t-2} & \ldots & \sum_j Y_{t-pi} T_{t-nj} \end{pmatrix},$$

$$\mathbf{A}_{23}^j = \begin{pmatrix} \sum_j Z_t T_t & \sum_j Z_t T_{t-1} & \sum_j Z_t T_{t-2} & \ldots & \sum_j Z_t T_{t-nj} \\ \sum_j Z_{t-1} T_t & \sum_j Z_{t-1} T_{t-1} & \sum_j Z_{t-1} T_{t-2} & \ldots & \sum_j Z_{t-1} T_{t-nj} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum_j Z_{t-pj} T_t & \sum_j Z_{t-pj} T_{t-1} & \sum_j Z_{t-pj} T_{t-2} & \ldots & \sum_j Z_{t-pi} T_{t-nj} \end{pmatrix},$$

$\mathbf{A}_{22}^j$ é a matriz cujo lk-ésimo elemento é $\sum_j Z_{t-l} Z_{t-k}$ e $\mathbf{A}_{33}^j$ é a matriz cujo lk-ésimo elemento é $\sum_j T_{t-l} T_{t-k}$. $\mathbf{A}_{21}^j$ é simétrica a $\mathbf{A}_{12}^j$, $\mathbf{A}_{31}^j$ é simétrica a $\mathbf{A}_{13}^j$ e $\mathbf{A}_{32}^j$ é simétrica a $\mathbf{A}_{23}^j$. A função de verossimilhança pode ser reescrita como

$$\mathcal{L}(\gamma_1,\gamma_2,\tau_1,\tau_2,r,d|D) \propto \tau_1^{\frac{b_1}{2}}\tau_2^{\frac{b_2}{2}} exp\{\sum_{j=1}^{2} -\frac{\tau_j}{2}[Y_j^{'}Y_j - \mathbf{B}_j^{'}\mathbf{A}_j^{-1}\mathbf{B}_j \quad (5)$$

$$+ \quad (\gamma_j - \mathbf{A}_j^{-1}\mathbf{B}_j)^{'}\mathbf{A}_j(\gamma_j - \mathbf{A}_j^{-1}\mathbf{B}_j)\} \quad (6)$$

em que $Y_j$ é o vetor das observações do j-ésimo regime.

## MATERIAL E MÉTODOS

Para estimar os parâmetros do modelo TARSO, utilizando inferência bayesiana, foi considerado a priori de Jeffreys, $P(\gamma_1,\gamma_2,\tau_1\tau_2)\alpha\tau_1^{-1}\tau_2^{-1}$. Essa priori combinada com a verossimilhança (6) gera uma posteriori normal-gama sendo as distribuições condicionais completas dadas por

$$(\gamma_i|\tau_1,\tau_2,d,r,D) \sim Normal(\mathbf{A}_i^{-1}\mathbf{B}_i(\tau_i\mathbf{A}_i)^{-1}), i=1,2$$

$$(\tau_i|\gamma_1,\tau_2,d,r,D) \sim Gama\frac{n_i-(p_i+m_i+1)}{2}, \frac{1}{2}(Y_i^{'}Y_i - \mathbf{B_i^{'}A_i^{-1}B_i} + (\gamma_i - \mathbf{A_i^{-1}B_i})^{'}\mathbf{A_i}(\gamma_i - \mathbf{A_i^{-1}B_i})), i=1,2.$$

Utilizando as distribuições condicionais completas, foi implementada uma função no software R para estimar os parâmetros do modelo TARSO, via cadeias de Markov.

Para verificar a eficiência da função implementada, foram realizadas 20 simulações de um modelo TARSO $(2;1,1;1,1;1,1)$. Em cada simulação, inicialmente, foram geradas 400 observações considerando aleatoriamente um valor threshold $r=18$ e uma defasagem $d=1$. As observações de $Y_t$ são condicionadas a $Z_{t-d}\leq r$ no regime 1 e $Z_{t-d}>r$ no regime 2.

Para estimar os parâmetros do modelo de cada simulação foram feitas 20000 iterações, utilizando a função implementada. Considerando o método de cadeias múltiplas, a convergência foi verificada através do pacote Boa do software R (R, 2010) pelo diagnóstico de convergência de Geweke que fornece o Z-Score e o valor-p para cada cadeia de parâmetros.

## RESULTADOS E DISCUSSÃO

A Tabela (1) apresenta os valores reais, valores estimados da média, mediana e moda para cada parâmetro e os respectivos intervalos de credibilidade. Nota-se que todos os parâmetros ficaram próximos dos valores reais. A convergência das cadeias pode ser observada na Tabela (2), utilizando um valor-p de $5\%$.

TABELA 1: Parâmetros simulados de um modelo TARSO (2;1,1;1,1;1,1).

| Regimes | Parâmetros | Valor Simulado | Média | Mediana | Moda | Limite Inferior | Limite Superior |
|---|---|---|---|---|---|---|---|
| Regime 1 | $\phi_{11}$ | 0,25 | 0,26 | 0,26 | 0,26 | 0,24 | 0,29 |
| | $\beta_{11}$ | 0,50 | 0,48 | 0,48 | 0,48 | 0,051 | 0,90 |
| | $\theta_{11}$ | 0,60 | 0,59 | 0,59 | 0,59 | 0,52 | 0,67 |
| | $\tau_1$ | 0,60 | 0,55 | 0,54 | 0,75 | 0,33 | 0,79 |
| Regime 2 | $\phi_{21}$ | -0,40 | -0,40 | -0,40 | -0,40 | -0,41 | -0,39 |
| | $\beta_{21}$ | 0,26 | 0,32 | 0,32 | 0,32 | 0,15 | 0,48 |
| | $\theta_{21}$ | 0,60 | 0,59 | 0,59 | 0,59 | 0,55 | 0,63 |
| | $\tau_2$ | 0,80 | 0,76 | 0,76 | 0,52 | 0,59 | 0,95 |

TABELA 2: Valor real e diagnóstico de convergência de Geweke

| Parâmetros | Valor Real | Z-Score | Valor-p |
|---|---|---|---|
| $\phi_{11}$ | 0,25 | 1,01 | 0,31 |
| $\beta_{11}$ | 0,50 | 1,00 | 0,32 |
| $\theta_{11}$ | 0,60 | -1,01 | 0,31 |
| $\tau_1$ | 0,60 | -1,31 | 0,19 |
| $\phi_{21}$ | -0,40 | 1,03 | 0,30 |
| $\beta_{21}$ | 0,26 | -0,90 | 0,37 |
| $\theta_{21}$ | 0,60 | 0,94 | 0,34 |
| $\tau_2$ | 0,80 | 0,19 | 0,84 |

## CONCLUSÕES

As funções desenvolvidas no pacote estatístico R para realizar a inferência bayesiana do modelo TARSO são adequadas e podem ser utilizadas para proceder à análise do mesmo.

## REFERÊNCIAS BIBLIOGRÁFICAS


R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

SÁFADI, T.; MORETTIN, P . A. Análise Bayesiana do Modelo Open Loop Threshold Autoregressive. Revista Brasileira de Estatística, Rio de Janeiro, v. 62, n. 217, p. 91-105, 2001.

TONG, H. Discussion of a paper by A. J. Lawrence and N. T. Kottegoda. Journal of the Royal Statistical Society A, 140, 34-35. 1977.

TONG, H; LIM, K. S. Threshold autoregression limit cycles and cyclical data (with discussion). Journal of the Royal Statistical Society B, 42, 245-292. 1980.