Identificação da Ordem de um Processo Auto-Regressivo Estacionário e Ergódico

CNPq/UFRGS
Apoio

Felipe Lazzari Vargas
Bolsista

Artur Oscar Lopes
Orientador

Introdução

É comum necessitar-se de modelos que possam descrever a característica de dependência entre observações ao longo do tempo. De fato, a correlação entre períodos de tempo podem ser encontradas em diferentes áreas tais como economia, hidrologia e meteorologia. Um dos modelos de séries temporais capazes de descrever esta característica é o auto-regressivo - AR(p).

Este trabalho tem como objetivo a identificação da ordem de um modelo AR(p) com inovações α-estáveis no caso estacionário e ergodicos. Esta classe de distribuições caracteriza-se pela capacidade de modelar várias características. Uma delas é a presença de caudas pesadas, comum em séries temporais das mais variadas áreas.

A identificação da ordem do processo foi feita por diversos critérios de seleção de modelos, alguns conhecidos da literatura, tais como: o Critério de Informação de Akaike(AIC), Critério de Informação Bayesiana(BIC) e o Critério de Hannan-Quin(HQC) comparando-os com outro mais recente, denominado o critério de Determinação Eficiente(EDC).

Método

Um dos maiores obstáculos à utilização da metodologia Box & Jenkins na construção de modelos ARIMA, está na identificação da ordem. Vários pesquisadores usando a mesma série podem identificar modelos diferentes. Outras propostas de identificação têm sido apresentadas na literatura. Entre elas, estão AIC, BIC, HQC e EDC, métodos baseados em uma função penalizadora.

A ideia é escolher a ordem do AR que minimiza a quantidade:

$$\text{critério} = -2\ln(L(.)) + k\, c_n,$$

onde $ln(L(.))$ é o logaritmo da função de verossimilhança, k depende do número de parâmetros do modelo e c_n é uma função penalizadora.

Estamos interessados em verificar se estes conceitos dados para situações onde as inovações são Gaussianas também são válidos, em questão de desempenho e propriedades estatísticas quando as inovações advém de distribuições α-estáveis.

Uma variável aleatória X é dita α-estável se sua função característica é dada por

$$\varphi_X(t) = \begin{cases} \exp\{-\lambda^\alpha |t|^\alpha \left[1 - i\beta(\tan\frac{\pi\alpha}{2})(\text{sign } t)\right] + i\mu t\}, & \alpha \neq 1 \\ \exp\{-\lambda |t| \left[1 - i\beta\frac{2}{\pi}(\text{sign } t)\ln(t)\right] + i\mu t\}, & \alpha = 1. \end{cases}$$

- α é o parâmetro de estabilidade ($0 < \alpha \leq 2$)
- β é o parâmetro de simetria ($|\beta| \leq 1$)
- λ é o parâmetro de escala ($\lambda > 0$)
- μ é o parâmetro de locação ($\mu \in \mathbb{R}$).

Foram escolhidos 4 critérios de seleção de ordem:

- Critério de Informação de Akaike(1976)

$$\text{AIC} = -2\ln(L(.)) + 2k$$

- Critério de Informação Bayesiana(1978)

$$\text{BIC} = -2\ln(L(.)) + 2\frac{1}{2}k\ln(n),$$

- Critério de Hannan-Quinn(1979)

$$\text{HQC} = -2\ln(L(.)) + 2\,c\,k\,\ln(\ln(n)),$$

onde c é uma constante real maior do que 1.

- Critério Eficiente de Informação(2006)

$$\text{EDC} = -2\ln(L(.)) + 2\,k\,\eta\ln(\ln(n)),$$

onde $\eta > 0$.

Como -2ln(L(.)) + 2k é um termo comum aos 4 critérios, podemos ver facilmente as diferenças entre os critérios de penalização através da seguinte tabela.

Tabela 1: Critério de Seleção de Modelos e sua respectiva Função Penalizadora

Critério	Função de Penalização
AIC	1
BIC	$\frac{1}{2}\ln(n)$
HQC	$c\ln(\ln(n))$
EDC	$\eta\ln(\ln(n))$

Para dar igual chance a todos os critérios, consideraremos que $\eta \in \{\frac{1}{2}, 1, 2\}$. De fato, quando $\eta = \frac{1}{2}$, estaremos comparando o critério EDC com o BIC; quando $\eta = 1$, estaremos comparando o EDC com o critério AIC e quando $\eta = 2$, estaremos comparando o EDC com o critério HQC, com o valor da constante c igual à 2.

Na literatura, estes critérios também apresentam uma formulação que ao invés de depender da função de máxima verossimilhança, dependem da estimativa de máxima verossimilhança residual. Como modelos AR(p) com inovações advindas de distribuições α-estáveis não apresentam variância residual finita, usamos no lugar a estimativa de máxima verossimilhança do parâmetro de escala de uma α-estável.

A fórmula geral, neste caso, é dada por:

$$\text{critério} = -2\ln\sigma^2 + k\,c_n,$$

Desta forma, as fórmulas para os critérios de seleção ficam:

- Critério de Informação de Akaike

$$\text{AIC} = n\ln(\hat{\sigma^2}) + 2k,$$

- Critério de Informação Bayesiana

$$\text{BIC} = n\ln(\hat{\sigma^2}) + 2\frac{1}{2}k\ln(n),$$

- Critério de Hannan-Quinn

$$\text{HQC} = n\ln(\hat{\sigma^2}) + 2\,c\,k\,\ln(\ln(n)),$$

onde c é uma constante real maior do que 1.

- Critério Eficiente de Informação

$$\text{EDC} = n\ln(\hat{\sigma^2}) + +2\,k\,\eta\ln(\ln(n)),$$

onde $\eta > 0$.

Onde n é o tamanho da amostra, $\ln(\sigma^2)$ é o logaritmo da estimativa do parâmetro de escala feito através do método de máxima verossimilhança ao quadrado, k é número de parâmetros do modelo e c_n é uma função penalizadora.

Como $n\ln(\hat{\sigma^2}) + 2\,k$ é uma expressão comum aos quatro critérios, as diferenças nas funções de penalização são dadas pela Tabela 1.

Simulações

Foram geradas séries temporais a partir de um processo AR(3), no software R, com inovações advindas de uma distribuição α-estável. Em que os parâmetros de escala e simetria do processo de inovação foram considerados iguais à zero enquanto que o índice de estabilidade (α) variou no conjunto de $\{0.8, 1, 1.2, 1.5, 1.8, 2\}$ e o parâmetro de escala (σ) no conjunto $\{1, 2, 3\}$. Foram utilizados diferentes tamanhos de amostras, burn-in, e replicações.

Define-se um processo AR(p) com inovação advinda de uma α-estável como:

$$X_t = c + \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \varepsilon_t, \quad \text{para todo} \quad t \in \mathbb{Z},$$

onde $\varepsilon_t \sim \alpha$-estável (α, σ^2) são variáveis aleatórias independentes e identicamente distribuídas (i.i.d.) onde $\boldsymbol{\eta} = (c, \phi_1, \phi_2, \cdots, \phi_p, \alpha, \sigma^2)$ é o vetor de parâmetros.

No momento de fazer o "goodness of fit", ou seja, de ajustar o modelo para as séries temporais, ajustou-se um AR(p), com p variando em $\{1, 2, 3\}$. Para cada um dos valores de p, foram solicitados os valores dos critérios de seleção de modelo estudados. Com a finalidade de verificar qual critério selecionava corretamente e com maior precisão. Além disso, comparamos o desempenho do critério EDC com os demais.

Tabela 2: Cenários para o processo de geração de dados

Cenário	n	burn-in	re
(i)	1000	100	500
(ii)	1000	500	500
(iii)	3000	100	500
(iv)	3000	500	500
(v)	1000	100	1000

Tabela 3: Estimativas e suas estatísticas de simulação com $n = 1000$, *burn-in* $= 100$, $re = 500$.

AR(p)	α	$\bar{\alpha}$	σ	$\bar{\sigma}$	ϕ_1	$\bar{\phi}_1$	ϕ_2	$\hat{\phi}_2$	ϕ_3	$\bar{\phi}_3$	loglik
AR(3)	1.0	1.843	2.0	4.078	0.8	0.782	0.4	0.299	-0.3	-0.198	-3732.397
AR(2)	1.0	1.659	2.0	5.110	-	0.773	-	0.084	-	-	-3974.152
AR(1)	1.0	1.127	2.0	4.333	-	0.857	-	-	-	-	-3809.292
AR(3)	1.5	1.906	1.0	1.248	0.8	0.793	0.4	0.407	-0.3	-0.300	-2167.901
AR(2)	1.5	1.811	1.0	1.345	-	0.749	-	0.144	-	-	-2279.095
AR(1)	1.5	1.517	1.0	1.201	-	0.879	-	-	-	-	-2244.004
AR(3)	2.0	1.950	3.0	2.960	0.8	0.804	0.4	0.389	-0.3	-0.293	-2868.105
AR(2)	2.0	1.970	3.0	3.118	-	0.749	-	0.174	-	-	-2913.246
AR(1)	2.0	1.990	3.0	3.188	-	0.903	-	-	-	-	-2929.049

Tabela 4: Valores para os critérios nas simulações quando $n = 1000$, *burn-in* $= 100$, $re = 500$.

AR(p)	AIC_L	$EDC(1)_L$	$AIC_{\hat{\sigma}}$	$EDC(1)_{\hat{\sigma}}$	BIC_L	$EDC(\frac{1}{2})_L$	$BIC_{\hat{\sigma}}$	$EDC(\frac{1}{2})_{\hat{\sigma}}$	HQC_L	$HQC_{\hat{\sigma}}$
AR(3)	7474.8	7484.1	2821.0	2830.4	7499.3	7474.5	2845.6	2820.7	7503.4	2849.7
AR(2)	7956.3	7963.8	3270.3	3277.8	7975.9	7956.0	3289.9	3270.0	7979.2	3293.2
AR(1)	7624.6	7630.2	2938.7	2944.3	7639.3	7624.4	2953.4	2938.5	7641.8	2955.9
AR(3)	4345.8	4355.1	452.6	462.0	4370.3	4345.5	477.2	452.3	4374.5	481.3
AR(2)	4566.2	4573.7	600.1	607.5	4585.8	4565.9	619.7	599.8	4589.1	623.0
AR(1)	4494.0	4499.6	372.1	377.6	4508.7	4493.8	386.8	371.8	4511.2	389.2
AR(3)	5746.2	5755.5	2180.5	2189.8	5770.7	5745.9	2205.0	2180.2	5774.9	2209.2
AR(2)	5834.5	5842.0	2282.2	2289.7	5854.1	5834.2	2301.9	2282.0	5857.4	2305.2
AR(1)	5864.1	5869.7	2325.0	2330.6	5878.8	5863.9	2339.7	2324.8	5881.3	2342.2

Com base, nos resultados obtidos nas simulações, observamos que os critérios de seleção de modelo quando utilizados na formulação que envolve o logaritmo da função de máxima verossimilhança obtiveram ótimos resultados selecionando corretamente o modelo AR(3). Enquanto que a formulação que utiliza a estimativa do parâmetro de escala não obteve resultando semelhante, como podemos ver, na Tabela 4 no segundo caso, em que seria selecionado o modelo AR(1) de forma incorreta.

Na comparação do critério EDC com os demais, ele saiu-se melhor que o BIC. Entretanto apresentou resultados inferiores ao AIC e por conta da fórmula utilizada apresentou resultados idênticos ao critério HQC.

Referências

Dorea, C.C.Y. e L.C. Zhao (2006). "Exponential bounds for the probability of wrong determination of the order of a Markov chain by using the EDC criterion". *Journal of Statistical Planning and Inference*, Vol.136, 3689-3697.

Morettin, P.A. e C.M.C. Toloi (2004). *Análise de Séries Temporais*. ABE - Projeto Fisher. São Paulo: Edgard Blücher.

Nolan, J.P. (2009). *Stable Distributions: Models for Heavy Tailed Data*.