Selecção de ordem em modelos INAR

Isabel da Silva[1]
Fac. de Engenharia da U.P., Dep. de Eng. Civil e Fac. de Ciências da U.P., Dep. de Mat. Aplic.
Maria Eduarda Silva
Faculdade de Ciências da Universidade do Porto, Departamento de Matemática Aplicada

Resumo: O modelo AutoRegressivo de valor INteiro, INAR, foi proposto na literatura para modelar séries de contagem. Neste trabalho, propõe-se e avalia-se um critério automático de selecção de ordem para modelos INAR, baseado no AICC, um dos critérios usados para determinar a ordem em modelos AutoRegressivos, AR.

Palavras–chave: Modelos INAR, Selecção de ordem, AICC.

Abstract: The INteger-valued AutoRegressive models, INAR, have been proposed in the literature to model count series. Here, an automatic criterion for selecting the order of INAR models is proposed and evaluated. This criterion is based in the AICC, one of the existing automatic criteria for selecting the order of AR models.

Keywords: INAR models, order selection, AICC.

1 Introdução

Muitas das séries temporais observadas são séries de valores inteiros não negativos e, em particular, séries de contagens. Os modelos usuais, quer lineares quer não lineares, para séries temporais não são neste caso adequados pois o produto de uma constante real por uma variável aleatória de valor inteiro produz uma variável aleatória real. Assim, McKenzie (1986, 1988) e Al-Osh & Alzaid (1987) recorreram à operação *thinning* binomial definida por Steutel & van Harn (1979) para substituir a operação de multiplicação usual e propuseram os modelos INAR(1), definidos a seguir.

Um processo estocástico discreto de valor inteiro não negativo, $\{X_t\}$, diz-se um processo INAR(1) se satisfaz a seguinte equação

$$X_t = \alpha * X_{t-1} + e_t$$

onde $\alpha \in]0,1]$, $\{e_t\} \in \mathbb{N}_0$ é uma sequência de variáveis aleatórias independentes e identicamente distribuídas (i.i.d.), com média μ_e e variância σ_e^2, e $*$ é a operação *thinning* binomial definida por (Steutel & van Harn (1979))

$$\alpha * X = \sum_{k=1}^{X} Y_k$$

[1]Isabel da Silva agradece ao PRODEP III pelo apoio financeiro.

onde $\alpha \in [0,1]$, $X \in \mathbb{N}_0$ é uma variável aleatória e $\{Y_k\}$, dita série de contagem, é uma sequência de variáveis aleatórias i.i.d., independentes de X, tais que $P(Y_k = 1) = 1 - P(Y_k = 0) = \alpha$.

Uma extensão natural para a ordem p, INAR(p), proposta por Du & Li (1991), é

$$X_t = \alpha_1 * X_{t-1} + \ldots + \alpha_p * X_{t-p} + e_t \tag{1}$$

onde $\{e_t\} \in \mathbb{N}_0$ são variáveis aleatórias i.i.d., com média μ_e e variância σ_e^2 finitas, $\alpha_i \in [0,1], i = 1, \ldots, p$, $\alpha_p \neq 0$, e as séries de contagem de $\alpha_k * X_{t-k}$, $k = 1, \ldots, p$, são mutuamente independentes e independentes de $\{e_t\}$. Sob estas condições, os momentos de segunda ordem do processo INAR são análogos aos de um processo AR.

Posteriormente, Gauthier & Latour (1994) generalizaram o conceito de operação *thinning*, permitindo que as séries de contagem de $\alpha_k * X_{t-k}$, $k = 1, \ldots, p$, sigam qualquer distribuição discreta, com média α_k e variância β_k, finitas.

Du & Li (1991) mostraram que a condição de estacionaridade do processo INAR(p) definido por (1) é que as raízes da equação $z^p - \alpha_1 z^{p-1} - \ldots - \alpha_{p-1} z - \alpha_p = 0$ estejam no interior do círculo unitário. Posteriormente, Latour (1998) mostrou que esta condição é equivalente a $\sum_{k=1}^{p} \alpha_k < 1$.

Utilizando as propriedades da operação *thinning*, Silva & Oliveira (2000b) obtiveram as expressões dos momentos e cumulantes de segunda e terceira ordem dos modelos definidos por (1). Em particular o valor esperado, μ_x, pode escrever-se como $\mu_x = \mu_e \left(1 - \sum_{i=1}^{p} \alpha_i\right)^{-1}$ e a função de autocovariância, $R(k)$, satisfaz um conjunto de equações do tipo Yule-Walker que podem ser escritas na forma escalar por

$$\begin{cases} R(0) &= V_p + \sum_{i=1}^{p} \alpha_i R(i) \\ R(k) &= \sum_{i=1}^{p} \alpha_i R(k-i), \; k \geq 1 \end{cases}$$

e na forma vectorial por

$$\mathbf{R}_p \, \boldsymbol{\alpha} = \begin{bmatrix} R(0) & R(1) & \ldots & R(p) \\ R(1) & R(0) & \ldots & R(p-1) \\ \vdots & \vdots & \ddots & \vdots \\ R(p) & R(p-1) & \ldots & R(0) \end{bmatrix} \begin{bmatrix} -1 \\ \alpha_1 \\ \vdots \\ \alpha_p \end{bmatrix} = \begin{bmatrix} -V_p \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{2}$$

com

$$V_p = \sigma_e^2 + \mu_x \; \sum_{k=1}^{p} \beta_k \tag{3}$$

onde β_k é a variância da série de contagem envolvida no *k*-ésimo operador *thinning*, $\alpha_k * X_{t-k}$, $k = 1, \ldots, p$.

Du & Li (1991) e Gauthier & Latour (1994) mostraram que os estimadores usuais da média e das funções de autocovariância e autocorrelação, $\bar{X} = \frac{1}{n} \sum_{t=1}^{n} X_t$, $\hat{R}(k) = \frac{1}{n} \sum_{t=1}^{n-k} (X_t - \bar{X})(X_{t+k} - \bar{X})$, e $\hat{\rho}(k) = \frac{\hat{R}(k)}{\hat{R}(0)}, 0 \leq k \leq n-1$, respectivamente, são fortemente consistentes.

No domínio da frequência, Silva & Oliveira (2000a, 2000b) obtiveram as expressões das funções de densidade espectral, $f(\omega)$, e biespectral, $f(\omega_1, \omega_2)$. Em particular, $f(\omega)$ pode escrever-se como

$$f(\omega) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} R(k)\, e^{-i\omega k} = \frac{1}{2\pi} \frac{V_p}{|1 - \sum_{k=1}^{p} \alpha_k e^{-i\omega k}|^2}, \ -\pi \leq \omega \leq \pi, \quad (4)$$

onde V_p está definido em (3).

O problema da estimação dos parâmetros $\alpha_i, i = 1, \ldots, p,\ \mu_e$ e σ_e^2 do modelo, tem sido considerado por diversos autores. Du & Li (1991) e Latour (1998) propuseram estimadores do tipo Yule-Walker e o método dos mínimos quadrados condicionais e demonstraram que este último método fornece estimativas assimptoticamente normais. No domínio da frequência, Silva & Oliveira (2000a) e Oliveira (2000) propuseram dois métodos de estimação baseados na minimização do critério de Whittle e do critério de Taniguchi.

No entanto, a modelação efectiva de uma série de observações por modelos INAR depende da determinação da ordem do modelo a usar.

Devido à semelhança da estrutura de correlação entre os INAR(p) e os AR(p), utilizaram-se, a título experimental e sem qualquer modificação, alguns dos critérios de selecção de ordem propostos para os modelos AR (FPE, AIC, AICC) nos modelos INAR, fazendo-se um estudo de simulação. Os resultados não foram satisfatórios no sentido dos critérios utilizados apresentarem, na maioria dos casos, uma tendência para seleccionar uma ordem maior que a verdadeira.

Neste trabalho propõe-se um critério automático para a selecção de ordem em modelos INAR, baseado num dos critérios existentes para modelos AR Gaussianos, o AICC. Apresentam-se, também, os resultados de um estudo de simulação onde se verifica a qualidade do critério proposto e procede-se, ainda, à aplicação deste critério na análise de observações provenientes de uma série real.

2 Selecção de ordem

2.1 Introdução

O problema da selecção da ordem em modelos de regressão e modelos de séries temporais tem sido considerado por vários autores e diversos critérios automáticos para selecção de ordem têm sido propostos. Os critérios automáticos têm como objectivo equilibrar o risco da escolha de uma ordem menor que a verdadeira, o que provoca inconsistência na estimação dos parâmetros, e o da escolha de uma ordem superior, que conduz ao incremento da variância desses estimadores. Este equilíbrio é feito através da atribuição de um custo ou penalização pela introdução de variáveis adicionais.

A ideia é, então, escolher a ordem k que minimiza um critério que pode ser escrito como uma função das observações (em geral, o erro quadrático médio da previsão a 1-passo ou a soma dos quadrados dos resíduos) mais um termo de penalização que depende do número de observações e da ordem do modelo a ajustar (Zhang (1992)).

De um modo geral, estes critérios baseiam-se na informação de Kullback-Leibler que Akaike (1973, 1974) propôs como critério de discriminação entre modelos concorrentes e que pode ser definido da seguinte maneira. Suponha-se que se dispõe de n observações de uma distribuição $g(y)$ (distribuição verdadeira) e que o modelo estatístico aproxima $g(y)$ através de $f(y)$. Então o índice de Kullback-Leibler pode definir-se por

$$\begin{aligned} I(g,f) &= E_y[\log(\, g(Y)/f(Y)\,)] = \int_{-\infty}^{+\infty} g(y)\, \log(\, g(Y)/f(Y)\,)\, dy \\ &= \int_{-\infty}^{+\infty} g(y)\, \log(g(y))dy - \int_{-\infty}^{+\infty} g(y)\, \log(f(y))dy \end{aligned} \tag{5}$$

onde E_y representa o valor esperado sob a distribuição verdadeira. Note-se que o primeiro integral de (5) tem o mesmo valor para qualquer aproximação $f(y)$ escolhida, pelo que o índice de Kullback-Leibler pode ser aproximado por $\int_{-\infty}^{+\infty} g(y) \log(f(y))dy$ e é estimado através de $\sum_{k=1}^{n} \log(f(y_k))$. Note-se, ainda, que se $f(y)$ representa a função de verosimilhança do modelo aproximante então, a menos da constante, o primeiro integral de (5) não é mais que o valor esperado de menos duas vezes a log-verosimilhança,

$$E_y[-2\log(f(y))], \tag{6}$$

que é outra aproximação do índice de Kullback-Leibler encontrada na literatura.

Dadas n observações independentes de um processo com vector de parâmetros $\boldsymbol{\theta}$, a proposta de Akaike (1974) consiste em considerar a verosimilhança $f(x|\boldsymbol{\theta})$ para várias dimensões de $\boldsymbol{\theta}$. Akaike (1974) mostra que, desde que a distribuição verdadeira pertença à família das distribuições aproximantes, i.e., $g(x) = f(x|\boldsymbol{\theta}_0)$, e sendo $\hat{\boldsymbol{\theta}}$ (estimador de máxima verosimilhança de $\boldsymbol{\theta}$) suficientemente próximo de $\boldsymbol{\theta}_0$, tem-se que

$$E[I(f(x|\boldsymbol{\theta}_0), f(x|\boldsymbol{\theta}))] \simeq -2\log(\hat{\sigma}_\epsilon^2) + 2(p+1) = AIC(p) \tag{7}$$

onde $\hat{\sigma}_\epsilon^2$ é o estimador de máxima verosimilhança da variância do ruído e p é a ordem do modelo aproximante. A ordem do modelo que melhor se ajusta às observações é o valor de p que minimiza AIC.

Uma outra solução para o problema de selecção de ordem do modelo baseia-se no erro de previsão. Akaike (1969, 1970) propôs o critério FPE, Erro de Previsão Final, que é um estimador do erro quadrático médio de previsão a 1-passo para uma realização independente da realização observada e utilizada para estimar os parâmetros do processo. Sejam $\{X_1, \ldots, X_n\}$ e $\{Y_1, \ldots, Y_n\}$ duas realizações independentes de um processo AR(p) estacionário com coeficientes $\alpha_1, \ldots, \alpha_p$, média nula e σ_ϵ^2 variância das inovações. Se $\hat{\alpha_1}, \ldots, \hat{\alpha_p}$ e $\hat{\sigma}_\epsilon^2$ são os estimadores de máxima verosimilhança dos parâmetros do processo obtidos a partir de $\{Y_1, \ldots, Y_n\}$ e o preditor linear a 1-passo de X_{n+1} é dado por $\hat{X}_{n+1} = \hat{\alpha}_1\, X_n + \ldots + \hat{\alpha}_p\, X_{n+1-p}$, então o erro quadrático médio de previsão é (Brockwell & Davis (1991, §9.3))

$$E[(X_{n+1} - \hat{X}_{n+1})^2] = \sigma_\epsilon^2 + E[(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})^T\, \mathbf{R}_p (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})],$$

onde $\mathbf{R}_p = E[X_i X_j]_{i,j=1}^p$, $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_p]^T$ e $\hat{\boldsymbol{\alpha}} = [\hat{\alpha_1}, \ldots, \hat{\alpha_p}]^T$.

Uma vez que os estimadores de máxima verosimilhança dos coeficientes são assimptoticamente normais e $n\,\hat{\sigma}_\epsilon^2/\sigma_\epsilon^2 \sim \chi^2_{n-p}$, então o erro quadrático médio de previsão pode ser aproximado por

$$FPE(p) = \hat{\sigma}_\epsilon^2 \left(1 + \tfrac{2(p+1)}{n}\right), \tag{8}$$

que define o critério proposto por Akaike (1969, 1970). A ordem seleccionada é o valor de p que minimiza o critério FPE.

Em geral, os critérios automáticos podem ser classificados como *assimptoticamente eficientes* (Shibata (1976) , Hurvich & Tsai (1989)) no sentido de que são procedimentos que escolhem o processo AR que atinge uma razão óptima de convergência para o erro quadrático médio de previsão ou como *consistentes*, i.e., se as observações são realmente geradas por um processo AR(k), então a ordem $\hat{k}$ escolhida pela minimização deste tipo de critérios é tal que $\hat{k} \to k$, com probabilidade 1 quando $N \to \infty$ (ver Brockwell & Davis (1991, p.305)). Entre estas duas propriedades dos estimadores de ordem, é preferida a eficiência assimptótica uma vez que quando se modelam dados reais raramente a ordem "verdadeira" é finita.

Os critérios AIC e FPE, já referidos, e o AICC (Hurvich & Tsai (1989)), versão corrigida do AIC, definido por $AICC(k) = n\,\log(\hat{\sigma}_\epsilon^2) + n\frac{1+k/n}{1-(k+2)/n}$, são assimptoticamente eficientes. Entre os critérios consistentes tem-se o HQ (Hannan & Quinn (1979)) definido por $HQ(k) = \log(\hat{\sigma}_\epsilon^2) + \frac{2kc}{n}\log(\log(n))$, $c > 1$, e os critérios construídos no contexto Bayesiano, SIC (Schwarz (1978)) e BIC (Akaike (1978)), dados respectivamente por $SIC(k) = n\log(\hat{\sigma}_\epsilon^2) + k\log(n)$ e $BIC(k) = (n-k)\log\left(\frac{n\hat{\sigma}_\epsilon^2}{n-k}\right) + n(1 + \log(\sqrt{2}\,\pi)) + k\log\left(\frac{\sum_{t=1}^n (X_t^2 - n\hat{\sigma}_\epsilon^2)}{k}\right)$.

2.2 Selecção de ordem nos modelos INAR

Considere-se uma série temporal de valores inteiros não negativos que se pretende modelar ajustando um modelo INAR. O uso dos critérios expostos anteriormente para determinar a *melhor* ordem do modelo INAR a considerar não é correcto, uma vez que a interpretação de σ_ϵ^2 como variância do erro de previsão a 1-passo ou, ainda, como variância dos resíduos para os modelos AR não tem analogia para os modelos INAR.

À semelhança de Hurvich & Tsai (1989), vai-se deduzir um critério do tipo AICC, considerando uma aproximação da função de verosimilhança através da função de densidade espectral, proposta por Whittle (1953) e usualmente designada por critério de Whittle.

Sejam $X_1, \ldots, X_n$ observações de um processo estacionário, com função de autocovariância $R(k)$ e função de densidade espectral $f(\omega)$. Suponha-se que $g(\omega)$ é uma função par, não negativa e integrável em $[-\pi,\, \pi]$. Segundo Whittle (1953), uma aproximação para a log-verosimilhança, $\ell(g)$, é tal que (Hurvich & Tsai (1989))

$$-2\ell(g) \simeq n\log(2\pi) + \tfrac{n}{2\pi}\int_{-\pi}^{\pi} \log(g(\omega)) + \tfrac{I_n(\omega)}{g(\omega)}\,d\omega$$

onde $I_n(\omega)$ é o periodograma definido por $I_n(\omega) = \frac{1}{2\pi n}|\sum_{t=1}^{n} X_t\, e^{-i\omega t}|^2$. Como o periodograma é um estimador assimptoticamente cêntrico da função de densidade espectral, $f(\omega)$, tem-se que

$$E[-2\ell(g)] \simeq d(f,g) = n\log(2\pi) + \tfrac{n}{2\pi}\int_{-\pi}^{\pi}\left(\log(g(\omega)) + \tfrac{f(\omega)}{g(\omega)}\right)d\omega,$$

que é uma aproximação do índice de Kullback-Leibler (ver (6)).

Suponha-se que o modelo aproximante é um modelo INAR(p), com vector de parâmetros $\hat{\boldsymbol{\alpha}} = [-1, \hat{\alpha}_1, \ldots, \hat{\alpha}_p]^T$, $\hat{\mu}_e$ e $\hat{\sigma}_e^2$ estimados, por exemplo, através do método dos mínimos quadrados condicionais, e espectro estimado por

$$\hat{f}(\omega) = \frac{1}{2\pi}\frac{\hat{V}_p}{|1-\sum_{k=1}^{p}\hat{\alpha}_k e^{-i\omega k}|^2}, \text{ com } \hat{V}_p = \hat{\sigma}_e^2 + \bar{X}\left(\sum_{k=1}^{p}\hat{\beta}_k\right), \qquad (9)$$

onde $\hat{\beta}_k$ é o estimador da variância da série de contagem envolvida no *k*-ésimo operador *thinning*, $\alpha_k * X_{t-k}$, $k = 1, \ldots, p$ e $\bar{X}$ é a média amostral.

Assumindo que a família de modelos aproximante inclui o modelo verdadeiro, i.e., o modelo verdadeiro é um INAR(p) com coeficientes $\boldsymbol{\alpha} = [-1, \alpha_1, \ldots, \alpha_p]^T$, sabe-se por (2) que são satisfeitas equações do tipo Yule-Walker, $\mathbf{R}_p\,\boldsymbol{\alpha} = [-V_p\ 0\ \cdots\ 0]^T$, com $\mathbf{R}_p$ definido em (2) e V_p definido em (3).

Então, utilizando a fórmula de Kolmogorov adaptada aos processos INAR(p), $V_p = 2\pi\exp[\int_{-\pi}^{\pi}\log(f(\omega))d\omega]$ (ver Brockwell & Davis (1991, p.191) para os modelos AR), e as propriedades das equações de Yule-Walker, tem-se que

$$\begin{aligned}
E\left[\tfrac{d(\hat{f},f)}{n}\right] &= E[\log(2\pi)] + \tfrac{1}{2\pi}E\left[\int_{-\pi}^{\pi}\left(\log\hat{f}(\omega) + \tfrac{f(\omega)}{\hat{f}(\omega)}\right)d\omega\right] \\
&= E[\log(\hat{V}_p)] + E\left[\int_{-\pi}^{\pi}\tfrac{f(\omega)}{\hat{V}_p}\left|1-\textstyle\sum_{k=1}^{p}\hat{\alpha}_k\, e^{i\omega k}\right|^2 d\omega\right] \\
&= E[\log(\hat{V}_p)] + E\left[\tfrac{1}{\hat{V}_p}\,\hat{\boldsymbol{\alpha}}^T\boldsymbol{R}_p\,\hat{\boldsymbol{\alpha}}\right] \\
&= E[\log(\hat{V}_p)] + E\left[\tfrac{V_p + (\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha})^T\,\boldsymbol{R}_p\,(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha})}{\hat{V}_p}\right] \qquad (10)
\end{aligned}$$

Tentou-se determinar, ou pelo menos aproximar, o valor esperado de $[(V_p + (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})^T\boldsymbol{R}_p(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}))/\hat{V}_p]$ através do método Delta (van der Vaart (1998, §3.1)) ou do desenvolvimento em série de Taylor dessa expressão, utilizando a distribuição normal assimptótica dos estimadores dos mínimos quadrados condicionais dos parâmetros de um modelo INAR(p) (Du & Li (1991)). No entanto, como a matriz de covariância é bastante complexa, não foi possível obter um termo de penalização satisfatório no sentido de que as aproximações obtidas dependem dos parâmetros $(\alpha_i, \mu_e, \sigma_e^2)$, da ordem do modelo a ajustar e do número de observações disponíveis.

Assim, optou-se por utilizar o termo de penalização correspondente ao critério existente para os modelos AR e, por conseguinte, o critério a considerar é:

$$\text{AICC}_{inar}(k) \quad = \quad n\,\log(\hat{V}_k) + n\frac{1+k/n}{1-(k+2)/n}. \qquad (11)$$

2.3 Estudo de Simulação

Para verificar o desempenho do critério proposto na secção anterior, foi calculada a frequência de selecção de ordem em 100 realizações de modelos INAR(p) com inovações de Poisson e operação *thinning* binomial, para diferentes ordens e valores dos parâmetros. Note-se que são apresentados, unicamente, os casos mais representativos das características encontradas para este critério.

Consideraram-se três dimensões de amostras: $n = 50$, $n = 100$ e $n = 200$ observações. Foram utilizados quatro métodos de estimação para obter os parâmetros dos modelos INAR a ajustar: o método dos mínimos quadrados condicionais, sem e com restrições e a estimação através do critério de Whittle, sem e com restrições. As restrições consideradas foram $0 < \hat{\alpha}_i < 1$, $i = 1, \ldots, p$ e $0 < \hat{\sigma}_e^2 < 60$.

Em cada realização foram estimados os parâmetros, pelos quatro métodos de estimação referidos, do modelo INAR(k) que se pretende ajustar, $k = 0, \ldots, 5$, foi calculado o valor do critério AICC$_{inar}$, dado em (11), para cada uma das ordens candidatas e a ordem escolhida é o valor de k onde o critério é mínimo. Posteriormente é calculada a frequência de selecção de ordem para todas as realizações. Também foi calculada a frequência de selecção de ordem para o critério usual, dado por AICC$_{ar}(k) = n \log(\hat{\sigma}_\epsilon^2) + n\frac{1+k/n}{1-(k+2)/n}$, para as mesmas realizações.

Uma primeira observação importante é que o método de estimação utilizado condiciona os resultados obtidos. Embora não sejam aqui apresentados resultados, por razões de espaço, o método de estimação que fornece as maiores frequências de selecção para a ordem verdadeira, na maioria dos casos, é a minimização do critério de Whittle sem restrições, pelo que a partir deste ponto, serão apresentados unicamente as frequências de selecção de ordem do AICC$_{inar}$ relativos a este método.

Na Tabela 1, apresentam-se as frequências de selecção de ordem dos critérios AICC$_{inar}$ e AICC$_{ar}$ para 100 realizações dos seguintes modelos INAR(p), p = 1, 2, 3, com inovações de Poisson de média 3, $\mu_e = 3$, quando os parâmetros são estimados através do método de Whittle, sem restrições. Estão indicadas, a **negrito**, as frequências máximas obtidas pelos critérios para cada um dos modelos.

- Modelo **I**: $X_t = 0.1 * X_{t-1} + e_t$, $n = 200$,
- Modelo **II**: $X_t = 0.4 * X_{t-1} + e_t$, $n = 50$,
- Modelo **III**: $X_t = 0.1 * X_{t-1} + 0.4 * X_{t-2} + e_t$, $n = 100$,
- Modelo **IV**: $X_t = 0.4 * X_{t-1} + 0.1 * X_{t-2} + e_t$, $n = 100$,
- Modelo **V**: $X_t = 0.1 * X_{t-1} + 0.1 * X_{t-2} + e_t$, $n = 200$,
- Modelo **VI**: $X_t = 0.3 * X_{t-1} + 0.3 * X_{t-2} + e_t$, $n = 200$,
- Modelo **VII**: $X_t = 0.1 * X_{t-1} + 0.2 * X_{t-2} + 0.3 * X_{t-3} + e_t$, $n = 200$,
- Modelo **VIII**: $X_t = 0.3 * X_{t-1} + 0.1 * X_{t-2} + 0.2 * X_{t-3} + e_t$, $n = 50$,
- Modelo **IX**: $X_t = 0.3 * X_{t-1} + 0.1 * X_{t-2} + 0.2 * X_{t-3} + e_t$, $n = 100$,

	AICC$_{inar}$						AICC$_{ar}$					
Mod.	**0**	**1**	**2**	**3**	**4**	**5**	**0**	**1**	**2**	**3**	**4**	**5**
I	**55**	26	8	2	2	7	0	0	0	1	16	**83**
II	16	**61**	14	7	2	0	0	**31**	30	26	9	4
III	2	0	**76**	13	6	3	0	0	**65**	17	11	7
IV	1	**60**	26	8	3	2	0	4	30	**34**	21	11
V	27	**29**	22	10	8	4	0	0	0	5	23	**72**
VI	0	0	**76**	11	9	4	0	0	**39**	15	17	29
VII	1	0	1	**85**	6	7	0	0	0	**77**	9	14
VIII	21	**50**	19	9	1	0	0	8	42	**43**	4	3
IX	1	26	23	**40**	7	3	0	1	11	**71**	12	5
X	0	2	7	**69**	15	7	0	0	0	**50**	31	19

Tabela 1: Frequência de selecção de ordem dos critérios AICC$_{inar}$ e AICC$_{ar}$ para 100 realizações de modelos INAR(1), INAR(2) e INAR(3).

- Modelo **X**: $X_t = 0.3 * X_{t-1} + 0.1 * X_{t-2} + 0.2 * X_{t-3} + e_t$, $n = 200$.

Para a primeira ordem, no modelo **I** ($\alpha_1 = 0.1$), a ordem mais frequentemente seleccionada pelo AICC$_{inar}$ (55%) é $k = 0$, que é incorrecta. Por outro lado, para o modelo **II** ($\alpha_1 = 0.4$), mesmo com uma amostra de dimensão reduzida (50 observações), o critério AICC$_{inar}$ apresenta a frequência máxima para a ordem verdadeira do modelo (61%). Isto poderá dever-se ao facto de que no primeiro caso o valor do coeficiente está "próximo" da região de não estacionaridade dos modelos INAR(1).

No caso da segunda ordem, para o modelo **III** ($\alpha_1 = 0.1$, $\alpha_2 = 0.4$,) o critério AICC$_{inar}$ apresenta maior frequência para a ordem verdadeira do processo (76%). Para o modelo **IV** ($\alpha_1 = 0.4$, $\alpha_2 = 0.1$), o AICC$_{inar}$ escolhe uma ordem incorrecta ($k = 1$) mais frequentemente (60%); isto pode dever-se ao facto do valor do primeiro coeficiente (α_1) ser comparativamente maior que o valor do segundo coeficiente (α_2), pelo que o critério escolhe mais frequentemente a primeira ordem. Quando $\alpha_1 = \alpha_2 = 0.1$ (modelo **V**), possivelmente devido à proximidade da região de não estacionaridade, o critério AICC$_{inar}$ escolhe mais frequentemente uma ordem incorrecta ($k = 1$). Para $\alpha_1 = \alpha_2 = 0.3$ (modelo **VI**), o AICC$_{inar}$ apresenta a maior frequência para a ordem correcta (76%).

Finalmente para a terceira ordem, no modelo **VII** ($\alpha_1 = 0.1, \alpha_2 = 0.2, \alpha_3 = 0.3$) a ordem verdadeira apresenta a maior frequência (85%). Os modelos **VIII**, **IX** e **X** são os mesmos, $\alpha_1 = 0.3, \alpha_2 = 0.1, \alpha_3 = 0.2$, mas a dimensão da amostra é diferente. O critério AICC$_{inar}$ escolhe a ordem verdadeira um maior número de vezes para as amostras com 100 e 200 observações. Acredita-se que as frequências de selecção de ordem não são muito altas porque o coeficiente destes processos de terceira ordem que apresenta um maior valor é α_1.

Analisando a segunda parte da tabela, relacionada com o critério AICC$_{ar}$, verifica-se que quando este critério selecciona a ordem correcta, em geral, a frequência de selecção é inferior à obtida pelo AICC$_{inar}$ e quando a frequência de selecção de ordem

é maior para uma ordem errada em ambos os critérios, o AICC$_{ar}$ favorece modelos com um maior número de parâmetros, o que desobedece o princípio da parcimónia.

3 Aplicação

A contagem diária dos ataques epilépticos é uma importante ferramenta para o estudo da doença. Estes dados consistem em valores inteiros não negativos, pelo que foram analisados como séries temporais de contagem por Franke & Seligmann (1993). Para ilustrar a técnica apresentada, nesta secção considera-se uma série de 121 observações correspondentes à contagem do número diário de ataques epilépticos de um dado paciente (Figura 22.3 de Franke & Seligmann (1993)). Os dados são apresentados na Figura 1.

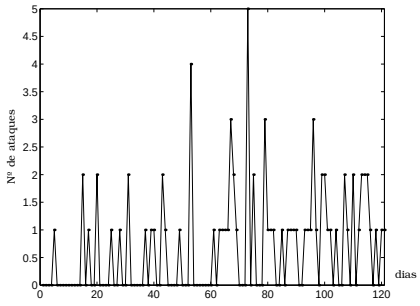

Figura 1: Número diário de ataques epilépticos de um dado paciente.

A função de autocorrelação amostral e a função de autocorrelação parcial amostral estão representadas na Figura 2. Pela análise destas funções, Latour (1998) propõe o seguinte processo INAR(14) generalizado para modelar estes dados $X_t = a_6 * X_{t-6} + a_{14} * X_{t-14} + \epsilon_t$,com $\alpha_6 = 0.28, \alpha_{14} = 0.24$ e $\sigma_\epsilon^2 = 0.75$ estimados através do método dos mínimos quadrados condicionais.

Ao aplicar o critério AICC$_{inar}$ para seleccionar a ordem do modelo INAR que melhor se adapta a este conjunto de dados, com uma ordem máxima possível de 20, o valor mínimo atingido pelo critério é de 105.67 para uma ordem $p = 6$, i.e., o critério selecciona um modelo INAR(6). Os estimadores dos parâmetros do modelo INAR(6) ajustado aos dados, obtidos através do critério de Whittle considerando as restrições $0 < \alpha_i < 1, i = 1, \ldots, p$, $\sum_{i=1}^{p} \alpha_i < 1$ e $0 < \mu_e < 60$, são $\hat{\alpha}_1 = 0.0232$, $\hat{\alpha}_2 = 0.0575$, $\hat{\alpha}_3 = 0.0239$, $\hat{\alpha}_4 = 0$, $\hat{\alpha}_5 = 0.0356$, $\hat{\alpha}_6 = 0.3149$, $\hat{\mu}_e = 0.4747$.

Para comparar entre os dois modelos propostos, o INAR(14) e o INAR(6), calculou-se o valor do critério para o modelo proposto por Latour, obtendo-se 153.18, que é maior que o valor obtido pelo modelo INAR(6) (105.67), pelo que este último deve ser preferido.

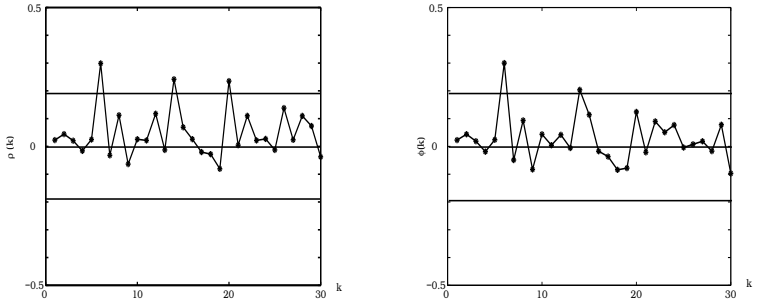

Figura 2: Funções de autocorrelação e autocorrelação parcial amostrais.

4 Conclusões

A análise de todos os resultados obtidos no estudo de simulação permite concluir que o critério

$$\text{AICC}_{inar}(k) = n\,\log(\hat{V}_k) + n\frac{1+k/n}{1-(k+2)/n}$$

apresenta um bom desempenho na maioria dos casos simulados, mesmo no caso de amostras pequenas (50 observações).

Vários factores influenciam o resultado obtido pela aplicação do critério proposto, como também acontece para os critérios usuais. Entre estes factores, observados no estudo de simulação, encontram-se a dimensão da amostra, o valor dos parâmetros (proximidade da região de não estacionaridade ou disparidade relativa entre os diversos coeficientes) e o método de estimação utilizado para obter os parâmetros a partir da amostra e calcular o valor do critério.

Na aplicação à série do número diário de ataques epilépticos de um dado paciente, o modelo que melhor se ajusta aos dados, do ponto de vista do AICC_{inar}, é um INAR(6), o que está de acordo com outros estudos realizados ao mesmo conjunto de dados.

Referências

[1] Akaike, H.(1969) Fitting autoregressive models for prediction. *Ann. Inst. Statist. Math*, Vol. 21, pp. 243-7.

[2] Akaike, H.(1970) Statistical predictor identification. *Ann. Inst. Statist. Math*, Vol. 22, pp. 203-17.

[3] Akaike, H.(1973) Information theory and an extension of the maximum likelihood principle. Em *2nd International Symposium on Information Theory*, Ed. B.N. Petrov and F. Csaki, pp. 267-81. Budapest: Akademia Kiado.

[4] Akaike, H.(1974) A new look at the statistical model identification. *IEEE Trans. Auto. Control*, Vol. AC-19, pp. 716-23.

[5] Akaike, H.(1978) A bayesian analysis of the minimum AIC procedure. *Ann. Inst. Statist. Math*, Vol. 30A pp. 9-14.

[6] Al-Osh, M.A. & Alzaid, A.A.(1987) First-order integer-valued autoregressive (INAR(1)) process. *J. time ser. anal.*, Vol. 8, pp. 261-75.

[7] Brockwell, P.J. & Davis, R.A. (1991) *Time Series: Theory and Methods*, 2nd Ed., Springer-Verlag, New York.

[8] Du, Jin-Guan & Li, Yuan (1991) The integer-valued autoregressive ((p)) model. *J. time ser. anal.*, Vol. 12, pp. 129-42.

[9] Franke, J. & Seligmann, T. (1993) Conditional maximum likelihood estimates for I NAR(1) processes and their application to modelling epileptic seizure counts. Em *Developments in Time Series Analysis*, ed. T. Subba Rao, Chapmann & Hall, London, pp. 310-30

[10] Gauthier, G. & Latour, A. (1994) Convergence forte des estimateurs des paramèters d'un processus GENAR(p). *Ann. Sci. Math. Québec*, Vol. 18, pp. 37-59.

[11] Hannan, E.J. & Quinn, B.G. (1979) The determination of the order of an autoregression. *J. R. Stat. Soc., B*, Vol. 41, pp. 190-5.

[12] Hurvich, C. & Tsai, C.L. (1989) Regression and time series model selection in small samples. *Biometrika*, Vol. 76, pp. 297-307.

[13] Latour, A. (1998) Existence and stochastic structure of a non-negative integer-valued autoregressive process. *J. time ser. anal.*, Vol. 19, pp. 439-55.

[14] McKenzie, E. (1986) Autoregressive moving-average process with negative-binomial and geometric marginal distributions. *Adv. Appl. Probab.*, Vol. 18, pp. 679-705.

[15] McKenzie, E. (1988) Some ARMA models for dependent sequences of Poisson counts. *Adv. Appl. Probab.*, Vol. 20, pp. 822-35.

[16] Oliveira, V.L. (2000) Modelos Autoregressivos para sucessões cronológicas de contagem: caracterização e modelação. *Tese de Doutoramento*, Universidade do Porto.

[17] Schwarz, G. (1978) Estimating the dimension of a model. *Ann. stat.*, Vol. 6, pp. 461-4.

[18] Shibata, R. (1976) Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika*, Vol. 63, pp. 117-26.

[19] Silva, M.E. & Oliveira, V.L. (2000a) Difference equations for the higher order moments and cumulants of the I NAR(1) model. *Technical Report CMA/6/00*, Universidade do Porto.

[20] Silva, M.E. & Oliveira, V.L. (2000b) Difference equations for the higher order moments and cumulants of the I NAR(p) model. *Technical Report CMA/9/00*, Universidade do Porto.

[21] Steutel, F.W. & van Harn, K. (1979) Discrete analogues of self-decomposability and stability. *Ann. probab.*, Vol. 7, pp. 893-99.

[22] van der Vaart, A.W. (1998) *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

[23] Whittle, P. (1953) The analysis of multiple stationary time series. *J. R. Stat. Soc., B*, Vol. 15, pp. 125-39.

[24] Zhang, P. (1992) On the distributional properties of model selection criteria. *J. Am. Stat. Assoc.*, Vol. 87, pp. 732-7.