Estimação em processos com longa dependência sazonais na presença de outliers

G.H.C. Laureano[1], C. Bisognin e S.R.C. Lopes

Instituto de Matemática - UFRGS
Porto Alegre, RS, Brasil

Resumo. Frequentemente são encontradas em séries temporais observações que são discordantes comparadas às restantes. Algumas se devem a erros grosseiros de medição, outras podem ser resultantes de influências externas, tais como greves, alterações súbitas na estrutura de mercado, entre outras. Como resultado destas influências externas surgem observações discordantes que são classificadas como *outliers*.

Neste trabalho, apresentamos os modelos de contaminação por mistura e estimação dos parâmetros dos processos SARFIMA$(0, d, 0) \times (0, D, 0)_s$ através de amostras geradas por processos contaminados, isto é, séries temporais contaminadas.

Introdução

Existem várias definições para *outliers*, entre elas citamos duas: "*outliers* are observations that do not follow the pattern of the majority of the data" [Rousseeuw e Zomeren (1990), pág. 633]; "We shall define an *outlier* in a set of data to be an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data" [Barnett e Lewis (1994), pág. 7]. Quando estudamos os *outliers*, a primeira questão que surge é sobre a sua classificação. Fox (1972) introduziu os conceitos de *outliers* do tipo I e tipo II, conhecidos na literatura, respectivamente, como *outliers* aditivos e inovadores e denotados, respectivamente, por *AO* e *IO*. Os *outliers* aditivos correspondem a um erro grosseiro de medição ou gravação afetando uma única observação. No caso dos *outliers* inovadores, ocorre um choque em um determinado período e o efeito se propaga para as observações subseqüentes. Uma segunda questão refere-se ao tipo de modelo gerador dos *outliers*. Denby e Martin (1979) e Bustos e Yohai (1986), entre outros, consideram o modelo de contaminação por mistura, isto é, o *outlier* (*AO*) Aditivo é gerado por uma dada distribuição de probabilidade.

Hotta e Neves (1992) e Chan e Palma (1998) consideram o modelo de contaminação paramétrico, isto é, os *outliers* são alocados em uma posição pré-determinada na série temporal.

Neste trabalho será abordada, apenas, a contaminação por *outliers* aditivos.

1. Processos de Longa Dependência e Sazonalidade

Para podermos definir os processos SARFIMA$(p, d, q) \times (P, D, Q)_s$ precisamos, primeiramente, introduzir os operadores de diferença e de diferença sazonal e definir o que é um processo de inovação ruído branco.

Definição 1.1. Quando os processos de inovação $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ tem $\varepsilon_t \sim Normal(0, \sigma^2)$, para todo $t \in \mathbb{Z}$. Considera-se que $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ trata-se de processo de inovação do tipo ruído branco.

Definição 1.2. Para todo $D > -1$, definimos o *operador diferença sazonal* $\nabla_s^D := (1 - \mathcal{B}^s)^D$, onde $s \in \mathbb{N}$ é a sazonalidade, através da expansão binomial

$$\nabla_s^D := (1 - \mathcal{B}^s)^D = \sum_{j\in\mathbb{Z}_{\geqslant}} \binom{D}{j} (-\mathcal{B}^s)^j = 1 - D\mathcal{B}^s - \frac{D(1-D)}{2!}\mathcal{B}^{2s} - \cdots , \tag{1}$$

onde

$$\binom{D}{j} \equiv \frac{\Gamma(D+1)}{\Gamma(j+1)\Gamma(D-j+1)},$$

na qual $\Gamma(\cdot)$ é a função Gama.

[1] E-mail: 00158953@ufrgs.br

Na Definição 1.2, se $s = 1$ e $D = d$, temos o operador diferença, denotado por $\nabla^d := (1 - \mathcal{B})^d$.

Os operadores diferença e diferença sazonal são muito importantes para a definição dos processos SARFIMA$(p, d, q)\times(P, D, Q)_s$ a seguir.

Definição 1.3. Seja $\{X_t\}_{t\in\mathbb{Z}}$ um processo estocástico satisfazendo a equação

$$\phi(\mathcal{B})\Phi(\mathcal{B}^s)\nabla^d\nabla_s^D(X_t - \mu) = \theta(\mathcal{B})\Theta(\mathcal{B}^s)\varepsilon_t, \tag{2}$$

onde μ é a média do processo, $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ é um processo ruído branco (ver Definição 1.1), $s \in \mathbb{N}$ é a sazonalidade, $\mathcal{B}$ é o operador de *defasagem ou* de *retardo*, isto é, $\mathcal{B}^j(X_t) = X_{t-j}$ e $\mathcal{B}^{sj}(X_t) = X_{t-sj}$, para $j, s \in \mathbb{N}$, ∇^d e ∇_s^D são os operadores, respectivamente, diferença e diferença sazonal, $\phi(\cdot)$ e $\theta(\cdot)$, $\Phi(\cdot)$ e $\Theta(\cdot)$ são os polinômios de ordem P, p, q e Q, respectivamente, definidos por

$$\begin{aligned} \phi(z) &= \sum_{\ell=0}^{p}(-\phi_\ell)\, z^\ell, \quad \theta(z) = \sum_{m=0}^{q}(-\theta_m)\, z^m, \\ \Phi(z) &= \sum_{r=0}^{P}(-\Phi_r)\, z^r, \quad \Theta(z) = \sum_{l=0}^{Q}(-\Theta_l)\, z^l, \end{aligned} \tag{3}$$

onde ϕ_ℓ, $1 \le \ell \le p$, θ_m, $1 \le m \le q$, Φ_r, $1 \le r \le P$, e Θ_l, $1 \le l \le Q$, são constantes reais e $\phi_0 = \Phi_0 = -1 = \theta_0 = \Theta_0$. Então, $\{X_t\}_{t\in\mathbb{Z}}$ é um *processo sazonal auto-regressivo fracionariamente integrado de média móvel de ordem* $(p, d, q) \times (P, D, Q)_s$ com sazonalidade s, denotado por SARFIMA$(p, d, q) \times (P, D, Q)_s$, onde d e D são, respectivamente, *o grau de diferenciação e o grau de diferenciação sazonal*.

Alguns casos particulares dos processos SARFIMA(p, d, q) $\times(P, D, Q)_s$ são ressaltados a seguir:

1) Quando $P = p = 0 = q = Q$ e $d = 0$ temos o chamado *processo sazonal fracionariamente integrado com sazonalidade* s, denotado por SARFIMA$(0, D, 0)_s$, e é representado por

$$\nabla_s^D(X_t - \mu) \equiv (1 - \mathcal{B}^s)^D(X_t - \mu) = \varepsilon_t, \text{ para todo } t \in \mathbb{Z}. \tag{4}$$

Para maiores detalhes sobre estes processos ver Brietzke et al. (2005) e Bisognin e Lopes (2007).

2) Quando $P = 0 = Q$, $D = 0$ e $s = 1$ o processo SARFIMA$(p, d, q) \times (P, D, Q)_s$ se reduz ao processo ARFIMA(p, d, q) (Lopes 2007).

Propriedades de *longa dependência*, os processos ARFIMA(p, d, q) e SARFIMA $(p, d, q) \times (Q, D, P)_s$.

A propriedade de longa dependência para um processo estocástico $\{X_t\}_{t\in\mathbb{Z}}$ pode ser definida no domínio do tempo ou no domínio da frequência. Como é apresentado na Definição 1.4 a seguir:

Definição 1.4. Seja $\{X_t\}_{t\in\mathbb{Z}}$ um processo estocástico estacionário. No domínio do tempo, se existe um número real $u \in (0, 1)$ tal que

$$\rho_{_X}(k) \sim C_1 k^{-u}, \text{ quando } k \to \infty,$$

onde $C_1 \neq 0$ e $\rho_{_X}(\cdot)$ é a função de autocorrelação do processo, então $\{X_t\}_{t\in\mathbb{Z}}$ possui *longa dependência*. Equivalentemente, no domínio da freqüência, se existe um número real $b \in (0, 1)$ tal que

$$f_{_X}(w) \sim C_2|w|^{-b}, \text{ quando } w \to 0,$$

onde $C_2 > 0$ e $f_{_X}(\cdot)$ é a função densidade espectral do processo, então $\{X_t\}_{t\in\mathbb{Z}}$ possui *longa dependência* (ver Bisognin e Lopes (2007)).

Notação: Na Definição 1.4, a notação $f(w) \sim g(w)$, quando $w \to 0$, significa $\lim_{w\to 0}\dfrac{f(w)}{g(w)} = 1$.

2. Contaminação por Mistura

Nesta seção, apresentamos o modelo de contaminação por mistura, proposto por Denby e Martin (1979), Bustos e Yohai (1986) e Beran (1994). Este modelo de contaminação será utilizado para contaminar os processos SARFIMA$(0, d, 0) \times (0, D, 0)_s$.

2.1. Outlier Aditivo (AO)

Um primeiro modelo de contaminação por mistura para os processos SARFIMA$(0,d,0)\times(0,D,0)_s$ é aquele contendo *outliers* aditivos, definidos a seguir.

Definição 2.1. Seja $\{Z_t\}_{t\in\mathbb{Z}}$ um processo estocástico satisfazendo a equação

$$Z_t := X_t + V_t, \tag{5}$$

onde $\{X_t\}_{t\in\mathbb{Z}}$ é um processo SARFIMA$(p,d,Q)\times(p,D,q)_s$, com processo de inovação $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ (ver Definição em Bisognin e Lopes (2007)). O processo $\{V_t\}_{t\in\mathbb{Z}}$ é constituído de variáveis aleatórias dadas por $V_t = \chi_t^c V_t^*$, onde $\{\chi_t\}_{t\in\mathbb{Z}}$ é um processo de Bernoulli com variáveis aleatórias independentes com probabilidade de sucesso c e $\{V_t^*\}_{t\in\mathbb{Z}}$ é um processo estocástico arbitrário com distribuição G.

Então, $\{Z_t\}_{t\in\mathbb{Z}}$ é dito ser um *processo com contaminação por mistura por* outliers *aditivos.*

Este modelo de contaminação é o mais geral onde podemos contaminar os processos SARFIMA$(p,d,q)\times(P,d,Q)_s$ com um processo estocástico $\{V_t^*\}_{t\in\mathbb{Z}}$ arbitrário. Para maiores detalhes sobre este modelo de contaminação ver Beran (1994). Neste trabalho utilizamos um modelo de contaminação semelhante o qual é definido a seguir.

Definição 2.2. Seja $\{Z_t\}_{t\in\mathbb{Z}}$ um processo estocástico satisfazendo a equação

$$Z_t := X_t + V_t, \tag{6}$$

onde $\{X_t\}_{t\in\mathbb{Z}}$ é um processo SARFIMA$(p,d,q)\times(P,d,Q)_s$, com processo de inovação $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ (ver Definição em Bisognin e Lopes (2007)). O processo $\{V_t\}_{t\in\mathbb{Z}}$ é constituído de variáveis aleatórias independentes e identicamente distribuídas com distribuição dada por

$$H_{\mathrm{V}} = (1-c)\delta_0 + cG, \tag{7}$$

onde $0 \leqslant c \leqslant 1$, δ_0 uma distribuição degenerada na origem, isto é, a esperança e a variância de uma variável aleatória com esta distribuição são ambas nulas e G uma distribuição arbitrária. Além disso, o processo $\{V_t\}_{t\in\mathbb{Z}}$ é um processo independente do processo $\{X_t\}_{t\in\mathbb{Z}}$. Então, $\{Z_t\}_{t\in\mathbb{Z}}$ é dito ser um *processo com contaminação por mistura por* outliers *aditivos.*

Neste caso, o processo $\{Z_t\}_{t\in\mathbb{Z}}$, que é um processo contaminado com *AO*, é igual ao processo $\{X_t\}_{t\in\mathbb{Z}}$ com probabilidade $1-c$, e tem probabilidade c de ser igual ao processo $\{X_t\}_{t\in\mathbb{Z}}$ adicionado de um erro $\{V_t\}_{t\in\mathbb{Z}}$. Então, no caso *AO* uma componente aleatória adicional sobrepõe-se ocasionalmente ao processo $\{X_t\}_{t\in\mathbb{Z}}$. Esta observação exibe um comportamento de *outlier*. Se $c=0$, não temos contaminação, isto é, o processo $\{Z_t\}_{t\in\mathbb{Z}}$ é idêntico ao processo $\{X_t\}_{t\in\mathbb{Z}}$.

Neste trabalho, vamos considerar que a função de distribuição de probabilidade G como sendo uma distribuição $N(0,\tau_V^2)$. Assim, podemos calcular a variância e a esperança das variáveis aleatórias V_t, para todo $t\in\mathbb{Z}$.

Sejam υ_t e ν_t variáveis aleatórias com distribuição, respectivamente, $F_\upsilon = \delta_0$ e $G_\nu = N(0,\tau_V^2)$. Então,

$$\mathbb{E}(V_t) \;=\; \int_{-\infty}^{\infty} x dH(x) = \int_{-\infty}^{\infty} x d((1-c)F_\upsilon(x) + cG_\nu(x)) = 0.$$

Da mesma forma, podemos calcular a variância. Então,

$$\mathrm{Var}(V_t) \;=\; \mathbb{E}(V_t^2) = \int_{-\infty}^{\infty} x^2 dH(x) = \int_{-\infty}^{\infty} x^2 d((1-c)F_\upsilon(x) + cG_\nu(x)) = c\tau_V^2.$$

Como os processos $\{V_t\}_{t\in\mathbb{Z}}$ e $\{X_t\}_{t\in\mathbb{Z}}$ são independentes, a função densidade espectral do processo $\{Z_t\}_{t\in\mathbb{Z}}$ é dada por

$$f_Z(w) = f_X(w) + f_V(w), \;\text{ para }\; w\in(0,\pi],$$

onde $f_X(\cdot)$ é a função densidade espectral do processo SARFIMA$(p,d,q)\times(P,D,Q)_s$ (ver em Bisognin e Lopes (2007)) e $f_V(w) = \frac{c\tau_V^2}{2\pi}$ é a função densidade espectral do processo $\{V_t\}_{t\in\mathbb{Z}}$. Logo,

$$f_Z(w) = \frac{\sigma_\varepsilon^2}{2\pi}\frac{|\theta(e^{-iw})|^2|\Theta(e^{-isw})|^2}{|\phi(e^{-iw})|^2|\Phi(e^{-isw})|^2}\left|2\ \mathrm{sen}\left(\frac{w}{2}\right)\right|^{-2d}\left|2\ \mathrm{sen}\left(\frac{sw}{2}\right)\right|^{-2D} + \frac{c\tau_V^2}{2\pi}. \tag{8}$$

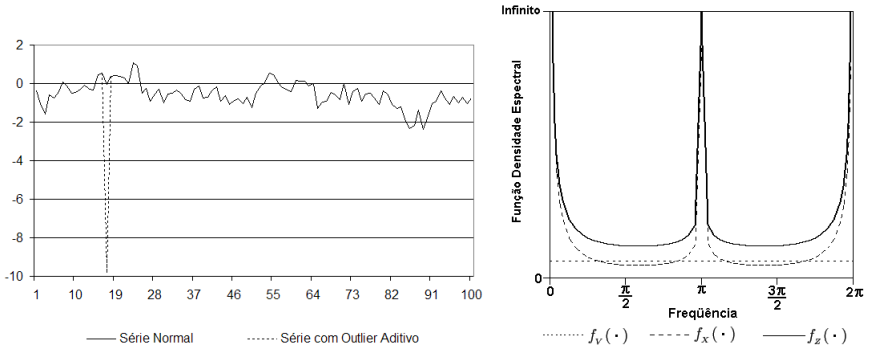
Por definição, a função de autocovariância de ordem h, $h\in\mathbb{Z}_{\geqslant}$, do processo $\{V_t\}_{t\in\mathbb{Z}}$, é dada por

$$\gamma_V(h) = 2\int_0^{\pi} f_V(w)\cos(wh)dw = \begin{cases} c\tau_V^2, & \text{se } h = 0, \\ 0, & \text{se } h \neq 0. \end{cases} \tag{9}$$

Desta forma, a função de autocovariância de ordem h, $h \in \mathbb{Z}_{\geqslant}$, do processo $\{Z_t\}_{t\in\mathbb{Z}}$, é dada por

$$\gamma_Z(h) = \begin{cases} \gamma_X(0) + \gamma_V(0), & \text{se } h = 0, \\ \gamma_X(h), & \text{se } h \neq 0, \end{cases} \tag{10}$$

onde $\gamma_X(\cdot)$ é a função de autocovariância do processo $\{X_t\}_{t\in\mathbb{Z}}$ que é um processo SARFIMA$(p,d,q)\times(P,D,Q)_s$ (ver em Bisognin e Lopes (2007)).



(a) (b)

FIGURA 1. (a)Série gerada a partir de um processo$\{Z_t\}_{t\in\mathbb{Z}}$, dado pela expressão (6), com $n = 100$, $d = 0.2$, $D = 0.25$, $P = p = 0 = q = Q$ e $s = 2$ e G dada por uma distribuição $N(0, 10)$. (b)Função densidade espectral de um processo $\{Z_t\}_{t\in\mathbb{Z}}$ ver em Bisognin e Lopes (2007), denotada por $f_Z(\cdot)$ (linha contínua), com $d = 0.2$, $D = 0.25$, $P = p = 0 = q = Q$, $s = 2$. O processo $\{X_t\}_{t\in\mathbb{Z}}$ possui função densidade espectral denotada por $f_X(\cdot)$ (em linha tracejada), ver em Bisognin e Lopes (2009). O processo $\{V_t\}_{t\in\mathbb{Z}}$ possui função densidade espectral dada por $f_V(\cdot)$ (em linha pontilhada), onde $c = 0.2$ e G é a distribuição $N(0, 10)$.

3. Estimação

Nesta seção apresentamos três estimadores utilizados na estimação dos parâmetros de longa dependência do processo SARFIMA$(0, d, 0) \times (0, D, 0)$.

Seja $\{X_t\}_{t\in\mathbb{Z}}$ um processo SARFIMA$(0, d, 0) \times (0, D, 0)_s$, com d e $D \in$ (-0.5,0.5). Cuja densidade espectral é dada pela expressão (8). Podemos reescrever a expressão (reffdesarfimaao) da seguinte forma:

$$f_X(\omega) = f_u(w)\left|2\sin\left(\frac{\omega}{2}\right)\right|^{-2d}\left|\sin\left(\frac{s\omega}{2}\right)\right|^{-2D} \tag{11}$$

para todo $0 < w < \pi$, onde $f_u(w) = \frac{\sigma_\varepsilon^2}{2\pi}$. Aplicando a função logarítmica nos dois lados da função densidade espectral e utilizando as propriedades logarítmicas e algumas transformações obtidas em Bisognin e Lopes (2007). Obtemos

$$\ln[I(\omega_j)] \simeq \ln[f_u(0)] - d\ln\left[2\,\text{sen}\left(\frac{\omega_j}{2}\right)\right]^2 - D\ln\left[2\,\text{sen}\left(\frac{s\omega_j}{2}\right)\right]^2 + \ln\left[\frac{I(\omega_j)}{f_X(\omega_j)}\right]. \tag{12}$$

Podemos observar que a equação (12) é uma forma aproximada da equação de regressão múltipla dada por

$$y_i \simeq \beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2} + \epsilon_t \tag{13}$$

para todo $j = 1, 2, \cdots, g(n)$, onde

$$y_i = \ln[I(\omega_j)], \quad x_{j1} = \ln\left[2\,\text{sen}\left(\frac{\omega_j}{2}\right)\right]^2, \quad x_{j2} = \ln\left[2\,\text{sen}\left(\frac{s\omega_j}{2}\right)\right]^2, \tag{14}$$

$$\epsilon_t = \ln\left[\frac{I(\omega_j)}{f_X(\omega_j)}\right], \quad \beta_0 = \ln[f_u(0)] - c, \quad c = \mathbb{E}\left(\ln\left[\frac{I(\omega_j)}{f_X(\omega_j)}\right]\right), \tag{15}$$

$$\beta_1 = -d, \quad \beta_2 = -D, \tag{16}$$

Onde, a função periodograma $I(\omega)$ é dada por:

$$I(\omega) = (2\pi)^{-1} \left| \sum_{t=1}^{n} X_t e^{-i\omega h} \right|^2 \tag{17}$$

Os estimadores obtidos através do procedimento MQ, sob a hipótese de normalidade dos erros, são consistentes e tem mínima variância entre todos os estimadores não viciados. A presença de *outliers* e pontos de alavanca, ou mesmo a perda da hipótese de normalidade dos erros são responsáveis por um considerável vício e ineficiência dos estimadores MQ (ver Huber, 1981 e Rousseeuw e Leroy, 2003).

Outro procedimento utilizado para estimar o vetor de parâmetros $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_l)$ são os procedimentos de estimação robusta, os quais apresentam estimadores que não são fortemente afetados por *outliers*.

Esse procedimento consiste em robustificar o estimador de MQ, isto é, ao invés de minimizarmos a soma dos quadrados dos resíduos minimizamos uma versão robusta da dispersão dos resíduos. Um estimador robusto é o dos "mínimos quadrados podados" (MQP), proposto por Rousseeuw (1984). Este estimador é dado pelo vetor $(\widehat{\beta}_0, \widehat{\beta}_1) \in \mathbb{R}^2$ que minimiza a função perda

$$\mathbf{P}(g(n)) = \sum_{j=1}^{m^*} (r^2)_{j:m}, \tag{18}$$

onde $(r^2)_{1:m} \leqslant \cdots \leqslant (r^2)_{m^*:m}$ são os resíduos ao quadrado e ordenados e m^* é o número de pontos utilizados no procedimento de otimização. A constante m^* é responsável pelo ponto de ruptura e eficiência.

Consideramos também os estimadores MM, propostos por Yohai (1987) e definidos como a solução $(\widehat{\beta}_0, \widehat{\beta}_1) \in \mathbb{R}^2$ que minimiza a função perda

$$\mathbf{P}(g(n)) = \sum_{j=1}^{g(n)} \rho_2 \left(\frac{r_j}{\kappa}\right)^2, \quad \text{sujeita a restrição} \quad \frac{1}{g(n)} \sum_{j=1}^{g(n)} \rho_1 \left(\frac{r_j}{\kappa}\right) \leqslant b, \tag{19}$$

onde ρ_1 e ρ_2 são funções simétricas, limitadas e não decrescentes em $[0, \infty)$, com $\rho_\ell(0) = 0$, $\lim_{u\to\infty} \rho_\ell(u) = 1$, para $\ell = 1, 2$, κ é um parâmetro de escala, b pode ser definido por $\mathbb{E}_\phi(\rho_1(u)) = b$, onde ϕ simboliza a distribuição normal padrão e r_j são os resíduos. O ponto de ruptura dos estimadores MM somente depende de ρ_1 e são dados por $\min\{b, 1-b\}$. Os estimadores MM são consistentes e assintoticamente normais.

Proposto por Geweke e Porter-Hudak (1983), denotado por GPHMQ, este estimador baseia-se no método de regressão utilizando a função periodograma (ver Brockwell e Davis, 1991). Para estimar os parâmetros de diferenciação d e diferenciação sazonal D aplica-se o método dos mínimos quadrados equação à equação (17). Para obtermos os estimadores GPHMQ, GPHMQP e GPHMM, aplicamos as metodologias MQ, MQP e MM, respectivamente, à equação (17), com $g(n) = n^\alpha$, para valores de α apropriados.

A função periodograma suavizado de covariâncias é utilizada no estimador BA em vez da função periodograma no estimador proposto por Geweke e Porter-Hudak (1983). Essa alteração decorre do fato de que a função periodograma é um estimador não viciado, mas inconsistente para a função densidade espectral de um processo. No entanto, a função periodograma suavizado de covariâncias é um estimador não viciado e consistente para ela, cuja expressão é dada por

$$I_{smooth}(w) = \frac{1}{2\pi} \sum_{|k| \leqslant m_p} \Lambda\left(\frac{k}{m_p}\right) \widehat{\gamma}_X(k) e^{-ikw}, \quad \text{para todo} \quad w \in (0, \pi], \tag{20}$$

onde $\widehat{\gamma}_X(\cdot)$ denota as funções de autocovariância amostral e $\Lambda(\cdot)$ é a função peso ou núcleo (maiores detalhes em Bisognin e Lopes 2007).

O estimador BA é obtido substituindo-se a função periodograma pela função periodograma suavizado de covariâncias, dado em (20). Assim, obtemos

$$\log\left[I_{smooth}(w_j)\right] \simeq -d \log\left[2\,\mathrm{sen}\left(\frac{w_j}{2}\right)\right]^2 - D \log\left[2\,\mathrm{sen}\left(\frac{s w_j}{2}\right)\right]^2 + \log\left[\frac{I_{smooth}(w_j)}{f_X(w_j)}\right], \tag{21}$$

onde $w_j = \frac{2\pi j}{n}$, $j \in \{0, 1, \cdots, g(n) | j \neq \frac{\nu n}{s}\}$ são as frequências de Fourier.

Neste trabalho, utilizamos a função peso ou núcleo de Bartlett, cuja expressão é dada por

$$\Lambda(x) = \begin{cases} (1 - |x|), & \text{se } |x| \leqslant 1, \\ 0, & \text{se } |x| > 1. \end{cases}$$

O estimador BA é baseado no uso da função periodograma suavizado de covariâncias (ver 20), utilizando a janela de Bartlett em vez da função periodograma no estimador proposto por Geweke e Porter-Hudak (1983). Da mesma forma que o estimador GPHMQ, utilizamos as metodologias MQ, MQP e MM para obtermos os estimadores BAMQ, BAMQP e BAMM.

Maiores detalhes sobre a função periodograma suavizado de covariâncias e janelas espectrais podem ser encontrados em Brockwell e Davis (1991), Morettin e Toloi (2004) e Koopmans (1974).

Na classe dos estimadores paramétricos utilizamos o estimador de máxima verossimilhança aproximado (W) proposto por Fox e Taqqu (1986). O estimador W utiliza uma aproximação para a matriz de autocovariância sugerida por Whittle (1951). Fox e Taqqu (1986) apresentam condições que permitem que este estimador, para uma seqüência com forte dependência, seja consistente e tenha distribuição assintoticamente normal. Sowell (1992) propõe o estimador de máxima verossimilhança exata. As condições necessárias para a consistência e normalidade assintótica do estimador de máxima verossimilhança exato foram apresentadas por Dahlhaus (1989). Ooms (1995) apresenta alguns resultados para modelos com longa dependência sazonais onde compara os estimadores de máxima verossimilhança exata, proposto por Sowell (1992), com os de máxima verossimilhança aproximada, proposto por Fox e Taqqu (1986) e ainda com o estimador proposto por Geweke e Porter-Hudak (1983).

Para maiores detalhes sobre os estimadores acima citados ver Bisognin e Lopes (2007).

Conclusões

Comparamos o desempenho dos estimadores semiparamétricos GPH, BA com o estimador paramétrico W proposto por Fox e Taqqu (1986), analisando seus erros quadráticos médios (eqm), vício e média. Esta comparação foi feita para processos SARFIMA$(0, d, 0) \times (0, D, 0)_s$ quando não são contaminados por mistura por *outlier* e quando são. Os resultados são baseados simulações feitas com 1000 replicações, com séries sazonais geradas utilizando o algoritmo de Durbin-Levinson e contaminadas pelo modelo de contaminação por mistura com *outlier* do tipo *AO*. Foi utilizado os seguintes valores para gerar as séries: tamanho amostral (n) igual a 500 e 1000, $d = 0, 1$, $D = 0, 3$, $l = 5$ e sazonalidade (s) igual a 4 e 7.

Com base nos resultados obtidos por simulação para processos SARFIMA$(0, d, 0) \times (0, D, 0)_s$ não contaminados, chegamos a conclusão de que:

O estimador com menor eqm é o BA, tanto para estimar o parâmetro D quanto para estimar o parâmetro d.

Para o estimador BA, a metodologia MQ apresenta menor eqm, seguida pela metodologia MQP e por MM.

O estimador GPH, utilizando as metodologias MQ e MM, é competitivo frente ao estimador BA, observando-se o eqm independentemente do tamanho amostral e dos valores de sazonalidade s.

Já para processos SARFIMA$(0, d, 0) \times (0, D, 0)_s$ contaminados por mistura por *outlier* os resultados obtidos por simulação mostraram que:

O eqm dos estimadores semiparamétricos aumenta em processos contaminados para todos tamanhos amostrais e valores de sazonalidade s analisados. Esse aumento é maior para as estimativas do parâmetro D do que para as do d.

As estimativas do parâmetro D e d são subestimadas, quanto ao seu verdadeiro valor, por todos os estimadores analisados.

O estimador com menor vício, ao estimar o parâmetro D é o estimador BA, independentemente do tamanho amostral e dos valores de sazonalidade s analisados.

O estimador paramétrico W proposto por Fox e Taqqu (1986) teve estimações semelhantes aos estimadores semiparamétricos BA e GPH. Entretanto, como ele requer alto custo computacional consideramos ele o menos atrativo para fazer as estimações dos parâmentros d e D, uma vez que o BA e o GPH necessitam de um custo computacional significativamente inferior ao estimador W e produzem estimativas boas.

Os resultados das estimãções estão dispostos em tabelas em anexo.

Referências

1. Abadir, K.M., W. Distaso e L. Giraitis (2007). "Nonstationarity- extended local Whittle estimation". *Journal of Econometrics*, Vol. **141**, pp. 1353-1384.
2. Barnett, V. e T. Lewis (1994) *Outliers in statistical data*, terceira edição.
3. Beran, J. (1994). *Statistics for Long-Memory Processes*. New York: Chapman & Hall.

4. Bisognin, C. e S.R.C. Lopes (2007). "Estimating and Forecasting the Long Memory Parameter in the Presence of Periodicity". *Journal of Forecasting* Vol. **26**(6), pp. 405-427.
5. Bisognin, C. (2007). "Estimação e Previsão em processo SARFIMA$(p, d, q) \times (P, D, Q)_s$ na presença de outliers". Tese de Doutorado no Programa de Pós-Graduação em Matemática da UFRGS, Porto Alegre.
6. Brietzke, E.H.M, S.R.C. Lopes, e C. Bisognin (2005) "A closed formula for the durbin-levinson's algorithm in seasonal fractionally integrated processes". *Mathematical and Computer Modelling* Vol. **42**(11-12), pp. 1191-1206.
7. Brockwell, P.J. e R.A. Davis (1991). *Time Series: Theory and Methods*. New York: Springer-Verlag.
8. Bustos, O. H, V e V.J. Yohai (1986). "Robust estimates for ARMA modeles" *Journal Am. Stat. Assoc.*,, Vol. **81**, pp. 155-168.
9. Chan,N.H. e W. Palma (1998), "State space modeling of long-memory processes", *The Annals of Statistics*, Vol. **26** pp. 719 - 740.
10. Dahlhaus, R. e B:M Pötscher (1989). "Convergence results for maximum likelihood type estimators in multivariable ARMA models II". *J. Multiv. Anal.* **30**, pp 241-244.
11. Denby, L. e R. D. Martin (1979). "Robust estimation of the first order autoregressive parameter". *J. Amer. Statist. Assoc.* Vol. **74**, pp. 140-146.
12. Doukhan, P., G. Oppenheim e M.S. Taqqu (2003). *Theory and Applications of Long-Range Dependence.* Boston: Birkhäuser.
13. Fox, A.J. (1972). "Outliers in Time Series". *Journal of the Royal Statistical Society*, Vol. **B-43**, pp. 350-363.
14. Fox, R. e M.S. Taqqu (1986). "Large-sample Properties of Parameter Estimates for Strongly Dependent Stationary Gaussian Time Series". *The Annals of Statistics*, Vol. **14**, pp. 517-532.
15. Geweke, J. e S. Porter-Hudak (1983). "The Estimation and Application of Long Memory Time Series Model". *Journal of Time Series Analysis*, Vol. **4**(4), pp. 221-238.
16. Hosking, J.R.M. (1981). "Fractional Differencing". *Biometrika*, Vol. **68**(1), pp. 165-176.
17. Hosking, J.R.M. (1984). "Modelling Persistence in Hydrological Time Series Using Fractional Differencing". *Water Resources Research*, Vol. **20**(12), pp. 1898-1908.
18. Hotta, L. K. e M. M. C. Neves (1992). "A Brief Review on Test for Detection of Ouliers of Time Series Models". *Revista Colombiana de Estadística*, Vol. **44**, pp. 103-148
19. Huber, P. (1981). Robust Statistics. Wiley, New York.
20. Kim, C.S. e P.C.B. Phillips (2006). "Log Periodogram Regres- sion: The Nonstationary Case". Cowles Foundation Discussion Paper N° 1587. Yale University, New Haven.
21. Koopmans, L.H. (1974). The spectral analysis of time series. Academic Press Inc., New York.
22. Lopes, S.R.C. (2008). "Long-range Dependence in Mean and Volatility: Models, Estimation and Forecasting, In *In and Out of Equilibrium 2 (Progress in Probability)*, (Edited by M.E. Vares and V. Sidoravicius), Birkhäuser, Rio de Janeiro.
23. Lopes, S.R.C. e B.V.M. Mendes (2006). "Bandwidth Selection in Classical and Robust Estimation of Long Memory". *International Journal of Statistics and Systems*, Vol. **1**(2), pp. 167-190.
24. Lopes, S.R.C., B.P. Olbermann e V.A. Reisen (2004). "A Comparison of Estimation in Non-Stationary ARFIMA Processes". *Journal of Statistical Computation and Simulation*, Vol. **74**(5), pp. 339-347.
25. Morettin, P. e C.C Toloi (2004). Análise de Séries Temporais. Editora: Edgard Blücher.
26. Ooms, M (1995). "Flexible Seasonal Long Memory and Economic Time Series". *Preprint of the Econometric Institute*, Erasmus University, Rotterdam.
27. Porter-Hudak, S. (1990). "An Application of the Seasonal Fractionally Differenced Model to the Monetary Aggregates". *Journal of the American Statistical Association*, Vol. **85**(410), pp. 338-344.
28. Rousseeuw, P.J. (1984). "Least Median of Square Regression". *Journal of the American Statistical Association*, Vol. **79**, pp. 871-880.
29. Rousseeuw, P.J. e B. C. Zomeren (1990). "Unmasking multivariate outliers and leverage points". *Journal of the American Statistical Association*, Vol. **85**, pp. 633-651.
30. Rousseeuw, P.J. e A. M. Leroy (2003). Robust Regression and Outlier Detection. Wiley.
31. Sowell, F. (1992). "Maximum Likelihood Estimation of Stationary Univariate Fractionally Integrated Time Series Models". *Journal of Econometrics*, Vol. **53**, pp. 165-188.
32. Whittle, P. (1951). *Hypothesis Testing in Time Series Analysis*. New York: Hafner.
33. Yohai, V. J. (1987). "High breakdown point and high efficiency robust estimates for regression". *Annals of Statistics*, Vol. **15**, pp. 642-656.

4. Anexos

TABELA 1. Resultado da estimação semiparamétrica para o processo SARFIMA$(0,d,0)\times(0,D,0)_s$, quando $d=0.1$, $D=0.3$, $s=4$, $n=500$ e $\ell=5$.

	$\widehat{D}_{GPH.MQ}$	$\widehat{d}_{GPH.MQ}$	$\widehat{D}_{GPH.MM}$	$\widehat{d}_{GPH.MM}$	$\widehat{D}_{GPH.MQP}$	$\widehat{d}_{GPH.MQP}$
média	0.3042	0.1012	0.2955	0.0970	0.3008	0.1020
vício	0.0042	0.0012	-0.0045	-0.0030	0.0008	0.0020
eqm	0.0029	0.0023	0.0058	0.0048	0.0034	0.0025
	$\widehat{D}_{BA.MQ}$	$\widehat{d}_{BA.MQ}$	$\widehat{D}_{BA.MM}$	$\widehat{d}_{BA.MM}$	$\widehat{D}_{BA.MQP}$	$\widehat{d}_{BA.MQP}$
média	0.3100	0.0923	0.3080	0.0940	0.3110	0.0940
vício	0.0100	-0.0077	0.0080	-0.0060	0.0110	-0.0060
eqm	0.0020	0.0015	0.0035	0.0027	0.0029	0.0021

TABELA 2. Resultado da estimação semiparamétrica para o processo SARFIMA$(0,d,0)\times(0,D,0)_s$, quando $d=0.1$, $D=0.3$, $s=4$, $n=1000$ e $\ell=5$.

	$\widehat{D}_{GPH.MQ}$	$\widehat{d}_{GPH.MQ}$	$\widehat{D}_{GPH.MM}$	$\widehat{d}_{GPH.MM}$	$\widehat{D}_{GPH.MQP}$	$\widehat{d}_{GPH.MQP}$
média	0.3030	0.1008	0.2969	0.0974	0.3011	0.0996
vício	0.0030	0.0008	-0.0031	-0.0026	0.0011	-0.0004
eqm	0.0013	0.0011	0.0027	0.0027	0.0015	0.0014
	$\widehat{D}_{BA.MQ}$	$\widehat{d}_{BA.MQ}$	$\widehat{D}_{BA.MM}$	$\widehat{d}_{BA.MM}$	$\widehat{D}_{BA.MQP}$	$\widehat{d}_{BA.MQP}$
média	0.3071	0.0951	0.3053	0.0974	0.3069	0.0964
vício	0.0071	-0.0049	0.0053	-0.0026	0.0069	-0.0036
eqm	0.0009	0.0007	0.0017	0.0014	0.0013	0.0010

TABELA 3. Resultado da estimação paramétrica para o processo SARFIMA$(0,d,0)\times(0,D,0)_s$, quando $d=0.1$, $D=0.3$, $s=4$ e $n\in\{500,1000\}$.

	$n=500$		$n=1000$	
	$\widehat{D}_W$	$\widehat{d}_W$	$\widehat{D}_W$	$\widehat{d}_W$
média	0.2731	0.0957	0.2882	0.0970
vício	-0.0269	-0.0043	-0.0118	-0.0030
eqm	0.0040	0.0015	0.0017	0.0007

TABELA 4. Resultado da estimação semiparamétrica para o processo SARFIMA$(0,d,0)\times(0,D,0)_s$, quando $d=0.1$, $D=0.3$, $s=7$, $n=500$ e $\ell=5$.

	$\widehat{D}_{GPH.MQ}$	$\widehat{d}_{GPH.MQ}$	$\widehat{D}_{GPH.MM}$	$\widehat{d}_{GPH.MM}$	$\widehat{D}_{GPH.MQP}$	$\widehat{d}_{GPH.MQP}$
média	0.3019	0.0985	0.2972	0.0946	0.3020	0.0958
vício	0.0019	-0.0015	-0.0028	-0.0054	0.0020	-0.0042
eqm	0.0031	0.0023	0.0065	0.0047	0.0034	0.0029
	$\widehat{D}_{BA.MQ}$	$\widehat{d}_{BA.MQ}$	$\widehat{D}_{BA.MM}$	$\widehat{d}_{BA.MM}$	$\widehat{D}_{BA.MQP}$	$\widehat{d}_{BA.MQP}$
média	0.3177	0.0880	0.3202	0.0927	0.3213	0.0931
vício	0.0177	-0.0120	0.0202	-0.0073	0.0213	-0.0069
eqm	0.0022	0.0018	0.0041	0.0031	0.0033	0.0024

TABELA 5. Resultado da estimação semiparamétrica para o processo SARFIMA$(0,d,0)\times(0,D,0)_s$, quando $d=0.1$, $D=0.3$, $s=7$, $n=1000$ e $\ell=5$.

	$\widehat{D}_{GPH.MQ}$	$\widehat{d}_{GPH.MQ}$	$\widehat{D}_{GPH.MM}$	$\widehat{d}_{GPH.MM}$	$\widehat{D}_{GPH.MQP}$	$\widehat{d}_{GPH.MQP}$
média	0.3013	0.1007	0.2988	0.0966	0.3015	0.0996
vício	0.0013	0.0007	-0.0012	-0.0034	0.0015	-0.0004
eqm	0.0014	0.0012	0.0029	0.0026	0.0016	0.0014
	$\widehat{D}_{BA.MQ}$	$\widehat{d}_{BA.MQ}$	$\widehat{D}_{BA.MM}$	$\widehat{d}_{BA.MM}$	$\widehat{D}_{BA.MQP}$	$\widehat{d}_{BA.MQP}$
média	0.3124	0.0945	0.3129	0.0958	0.3140	0.0963
vício	0.0124	-0.0055	0.0129	-0.0042	0.0140	-0.0037
eqm	0.0010	0.0008	0.0018	0.0014	0.0014	0.0011

TABELA 6. Resultado da estimação paramétrica para o processo SARFIMA$(0,d,0)\times(0,D,0)_s$, quando $d=0.1$, $D=0.3$, $s=7$ e $n\in\{500,1000\}$.

	$n=500$		$n=1000$	
	$\widehat{D}_W$	$\widehat{d}_W$	$\widehat{D}_W$	$\widehat{d}_W$
média	0.2518	0.0922	0.2892	0.0952
vício	-0.0482	-0.0078	-0.0108	-0.0048
eqm	0.0050	0.0015	0.0011	0.0007

TABELA 7. Resultado da estimação semiparamétrica para o processo SARFIMA$(0,d,0)\times(0,D,0)_s$, quando $d=0.1$, $D=0.3$, $s=4$, $n=500$ e $\ell=5$, com contaminação por *outlier AO*.

	$\widehat{D}_{GPH.MQ}$	$\widehat{d}_{GPH.MQ}$	$\widehat{D}_{GPH.MM}$	$\widehat{d}_{GPH.MM}$	$\widehat{D}_{GPH.MQP}$	$\widehat{d}_{GPH.MQP}$
média	0.2655	0.0925	0.2545	0.0900	0.2658	0.0918
vício	-0.0345	-0.0075	-0.0455	-0.0100	-0.0342	-0.0082
eqm	0.0048	0.0025	0.0092	0.0049	0.0051	0.0029
	$\widehat{D}_{BA.MQ}$	$\widehat{d}_{BA.MQ}$	$\widehat{D}_{BA.MM}$	$\widehat{d}_{BA.MM}$	$\widehat{D}_{BA.MQP}$	$\widehat{d}_{BA.MQP}$
média	0.2768	0.0828	0.2715	0.0882	0.2745	0.0883
vício	-0.0232	-0.0172	-0.0285	-0.0118	-0.0255	-0.0117
eqm	0.0032	0.0019	0.0060	0.0031	0.0042	0.0024

TABELA 8. Resultado da estimação semiparamétrica para o processo SARFIMA$(0,d,0)\times(0,D,0)_s$, quando $d=0.1$, $D=0.3$, $s=4$, $n=1000$ e $\ell=5$, com contaminação por *outlier AO*.

	$\widehat{D}_{GPH.MQ}$	$\widehat{d}_{GPH.MQ}$	$\widehat{D}_{GPH.MM}$	$\widehat{d}_{GPH.MM}$	$\widehat{D}_{GPH.MQP}$	$\widehat{d}_{GPH.MQP}$
média	0.2817	0.0973	0.2758	0.0925	0.2804	0.0971
vício	-0.0183	-0.0027	-0.0242	-0.0075	-0.0196	-0.0029
eqm	0.0020	0.0011	0.0041	0.0026	0.0023	0.0014
	$\widehat{D}_{BA.MQ}$	$\widehat{d}_{BA.MQ}$	$\widehat{D}_{BA.MM}$	$\widehat{d}_{BA.MM}$	$\widehat{D}_{BA.MQP}$	$\widehat{d}_{BA.MQP}$
média	0.2889	0.0909	0.2882	0.0931	0.2897	0.0921
vício	-0.0111	-0.0091	-0.0118	-0.0069	-0.0103	-0.0079
eqm	0.0012	0.0008	0.0021	0.0015	0.0016	0.0011

TABELA 9. Resultado da estimação paramétrica para o processo SARFIMA$(0,d,0)\times(0,D,0)_s$, quando $d=0.1$, $D=0.3$, $s=4$ e $n\in\{500,1000\}$ com contaminação por *outlier AO*.

	$n=500$		$n=1000$	
	$\widehat{D}_W$	$\widehat{d}_W$	$\widehat{D}_W$	$\widehat{d}_W$
média	0.2377	0.0888	0.2669	0.0939
vício	-0.0623	-0.0112	-0.0331	-0.0061
eqm	0.0082	0.0016	0.0029	0.0008

TABELA 10. Resultado da estimação semiparamétrica para o processo SARFIMA$(0,d,0)\times(0,D,0)_s$, quando $d=0.1$, $D=0.3$, $s=7$, $n=500$ e $\ell=5$, com contaminação por *outlier AO*.

	$\widehat{D}_{GPH.MQ}$	$\widehat{d}_{GPH.MQ}$	$\widehat{D}_{GPH.MM}$	$\widehat{d}_{GPH.MM}$	$\widehat{D}_{GPH.MQP}$	$\widehat{d}_{GPH.MQP}$
média	0.2600	0.0902	0.2526	0.0809	0.2628	0.0867
vício	-0.0400	-0.0098	-0.0474	-0.0191	-0.0372	-0.0133
eqm	0.0056	0.0024	0.0083	0.0048	0.0053	0.0025
	$\widehat{D}_{BA.MQ}$	$\widehat{d}_{BA.MQ}$	$\widehat{D}_{BA.MM}$	$\widehat{d}_{BA.MM}$	$\widehat{D}_{BA.MQP}$	$\widehat{d}_{BA.MQP}$
média	0.2791	0.0795	0.2793	0.0838	0.2799	0.0830
vício	-0.0209	-0.0205	-0.0207	-0.0162	-0.0201	-0.0170
eqm	0.0031	0.0019	0.0046	0.0033	0.0038	0.0024

TABELA 11. Resultado da estimação semiparamétrica para o processo SARFIMA$(0,d,0)\times(0,D,0)_s$, quando $d=0.1$, $D=0.3$, $s=7$, $n=1000$ e $\ell=5$, com contaminação por *outlier AO*.

	$\widehat{D}_{GPH.MQ}$	$\widehat{d}_{GPH.MQ}$	$\widehat{D}_{GPH.MM}$	$\widehat{d}_{GPH.MM}$	$\widehat{D}_{GPH.MQP}$	$\widehat{d}_{GPH.MQP}$
média	0.2814	0.0950	0.2775	0.0900	0.2807	0.0944
vício	-0.0186	-0.0050	-0.0225	-0.0100	-0.0193	-0.0056
eqm	0.0019	0.0012	0.0038	0.0029	0.0023	0.0014
	$\widehat{D}_{BA.MQ}$	$\widehat{d}_{BA.MQ}$	$\widehat{D}_{BA.MM}$	$\widehat{d}_{BA.MM}$	$\widehat{D}_{BA.MQP}$	$\widehat{d}_{BA.MQP}$
média	0.2930	0.0891	0.2914	0.0907	0.2925	0.0917
vício	-0.0070	-0.0109	-0.0086	-0.0093	-0.0075	-0.0083
eqm	0.0012	0.0009	0.0021	0.0015	0.0016	0.0011

TABELA 12. Resultados da estimação paramétrica para o processo SARFIMA$(0,d,0)\times(0,D,0)_s$, quando $d=0.1$, $D=0.3$, $s=7$ e $n\in\{500,1000\}$ com contaminação por *outlier AO*.

	$n=500$		$n=1000$	
	$\widehat{D}_W$	$\widehat{d}_W$	$\widehat{D}_W$	$\widehat{d}_W$
média	0.2097	0.0844	0.2595	0.0919
vício	-0.0903	-0.0156	0.0008	-0.0081
eqm	0.0117	0.0017	0.0030	0.0008