ESTIMAÇÃO ROBUSTA EM PROCESSOS DE MEMÓRIA LONGA NA PRESENÇA DE OUTLIERS ADITIVOS

FABIO ALEXANDER FAJARDO MOLINARES

Orientador: Prof. Francisco Cribari-Neto
Co-orientador: Prof. Valdério A. Reisen

Área de Concentração: Estatística Aplicada

Dissertação submetida como requerimento parcial para obtenção do grau de Mestre em Estatística pela Universidade Federal de Pernambuco

Recife, fevereiro de 2007

Molinares, Fábio Alexander Fajardo
Estimação robusta em processos de memória longa na presença de outliers aditivos / Fábio Alexander Fajardo Molinares – Recife : O autor, 2007.
xi, 42 folhas: il., fig., tab.

Dissertação (mestrado) – Universidade Federal de Pernambuco. CCEN. Estatística, 2007.

Inclui bibliografia.

1.Estatística Matemática. 2. Memória longa 3. Outliers. 4. Robustez. Título.

519.5 CDD (22.ed.) MEI2007-008

Universidade Federal de Pernambuco
Pós-Graduação em Estatística

14 de fevereiro de 2007
(data)

Nós recomendamos que a dissertação de mestrado de autoria de

Fábio Alexander Fajardo Molinares

intitulada

"Estimação robusta em processos de memória longa na presença de outliers aditivos"

seja aceita como cumprimento parcial dos requerimentos para o grau de Mestre em Estatística.

Klaus Leite Pinto Vasconcellos
Coordenador da Pós-Graduação em Estatística

Banca Examinadora:

Francisco Cribari Neto
Francisco Cribari Neto orientador

Ronaldo Dias
Ronaldo Dias (UNICAMP)

Klaus Leite Pinto Vasconcellos
Klaus Leite Pinto Vasconcellos

Este documento será anexado à versão final da dissertação.

A Dios.

A mis padres Carmen y Jorge, a mis hermanos Olga, Edgar, Marlon y Jorge.

AGRADECIMENTOS

Ao professor Valdério Anselmo Reisen, pela amizade, confiança, paciência e magnífica orientação, assim como pelas intermináveis discussões que tornaram possível o desenvolvimento deste trabalho, contribuindo para o meu amadurecimento acadêmico e profissional.

Ao professor Francisco Cribari-Neto, pelo estímulo, dedicação e leitura cuidadosa deste trabalho.

A Nátaly, meu grande amor, por todos os momentos de alegria, pelo constante apoio e pelo incentivo para que todas as metas traçadas fossem alcançadas.

A César Fajardo, Zuleima, César Felipe, John Sebastian, Clementina e Nieves Camacho, pelo carinho e pelo constante apoio.

Aos meus amigos da Colômbia, em especial a Camilo e Pacho, que sempre estiveram presentes nos momentos de alegria e tristeza, agradeço as intermináveis horas de papo e de desconcentração que tornaram minha permanência no Recife mais amena.

Aos professores e amigos Rafael E. Ahumada Barrios e Francisco J. Cepeda Coronado, peles ótimos conselhos e pela orientação segura na minha vida académica e profissional.

Aos professores do Departamento de Estatística da Universidad Nacional de Colombia, muito especialmente ao professor Luis A. López, pela amizade e pela contribuição à formação profissional, e ao professor Fabio H. Nieto, por ser sempre um exemplo para mim e pela sua orientação que conduziu à minha paixão por séries temporais.

Aos professores do Departamento de Estatística da UFPE, especialmente ao professor Klaus Vasconcellos pelo excelente curso de Inferência Estatística, ao professor Francisco Cribari-Neto pelo valioso material lecionado nas disciplinas de Estatística Computacional e Estatística Aplicada, ao professor Andrei Toom pelo valioso ensino na área de Probabilidade e Métodos Matemáticos.

Aos meus amigos da colônia colombiana no Brasil, em especial a Barba, John, Patricia e Ricardo.

Aos meus amigos do Brasil, em especial a Lia e Lúcia em Vitória, Gilvan e Márcia em Brasília, Alexandre e Andrea em Recife, pela amizade, apoio emocional e momentos de diversão.

A todos do Departamento de Estatística da Universidade Federal do Espírito Santo que em tão pouco tempo se tornaram meus amigos, especialmente a Bartolomeu Zamprogno, Alyne Neves, Jacqueline, Miriam, Geovane, Giovanni e Alessandro, pelo apoio e momentos de diversão.

Aos meus colegas do Mestrado em Estatística da Universidade Federal de Pernambuco, pelas horas compartilhadas.

A Valéria Bittercourt, pela presteza e amabilidade com que sempre me atendeu.

À CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), pelo apoio financeiro.

RESUMO

O objetivo deste trabalho é propor uma metodologia para estimar os parâmetros que indexam o processo ARFIMA(p, d, q) (Hosking 1981) na presença de *outliers* aditivos. Para estimar d, é proposto um estimador robusto que é uma variante do popular estimador sugerido por Geweke & Porter-Hudak (1983) (GPH). A metodologia proposta faz uso da função de autocovariância amostral robusta, considerada por Ma & Genton (2000), para obtenção do estimador da função espectral do processo. Resultados numéricos evidenciam a robustez do estimador proposto na presença de *outliers* do tipo aditivo.

Palavras-chaves: Memória longa, outliers, robustez.

ABSTRACT

In this thesis, we introduce an alternative semiparametric estimator of the fractional differencing parameter in ARFIMA models. The proposed estimator is a variant of the well-known GPH estimator and is robust against additive outliers. We use the robust sample autocorrelations considered by Ma & Genton (2000) to obtain a robustified estimator for the spectral density of the process. Numerical results show that the estimator we propose for the differencing parameter is robust when the data contain additive outliers.

Key words: Long-memory, outliers, robustness.

Índice

Lista de Figuras

Lista de Tabelas

CAPÍTULO 1

Introdução

Séries econômicas e financeiras comumente contêm observações influenciadas por eventos externos que podem provocar mudanças em suas dinâmicas, algumas vezes de forma transitória e outras vezes de forma permanente. Essas observações são conhecidas na literatura como dados atípicos ou *outliers* e, dependendo de sua natureza, seus efeitos sobre os processos inferenciais podem ser substanciais.

Estudos baseados na suposição de que as séries observadas são geradas por um processo auto-regressivo integrado e de médias móveis (ARIMA) (Box, Jenkins & Reinsel (1994)) mostram a influência de *outliers* sobre as estimativas dos parâmetros do modelo e sobre as previsões obtidas a partir dos modelos ajustados. Por exemplo, Ledolter (1989) mostrou que intervalos de previsão são muito sensíveis a *outliers* aditivos, mas previsões pontuais não são significativamente afetadas por tais dados atípicos, a não ser quando os *outliers* encontram-se próximos da origem da previsão; Chang, Tiao & Chen (1988) e Chen & Liu (1993*b*) mostraram que a presença de *outliers* nos dados provoca viés nas estimativas dos parâmetros do modelo ARMA; Deutsch, Richards & Swain (1990) e Chan (1992, 1995) derivaram resultados sobre o viés ocasionado na função de autocorrelação amostral pela presença de observações atípicas.

No início da década de 80, Granger & Joyeux (1980) e Hosking (1981) propuseram uma extensão dos processos ARIMA em que o parâmetro de integração assume valores fracionários: o processo ARFIMA. Hosking (1981) provou que séries com representação ARFIMA(p, d, q), para valores $d \in (0, 0.5)$, apresentam estacionariedade e memória longa, sendo esta caracterizada por correlações estatisticamente significativas entre observações distantes; equivalentemente, a função de densidade espectral possui singularidade na freqüência zero.

Existem diferentes propostas para estimação dos parâmetros do modelo ARFIMA, tanto de caráter paramétrico quanto de semi-paramétrico. Nos métodos paramétricos procede-se à estimação simultânea dos parâmetros do modelo, em geral por máxima verossimilhança; ver, e.g., Beran (1995), Dahlhaus (1989), Fox & Taqqu (1986), Hauser (1999) e Sowell (1992). No procedimento semi-paramétrico, a estimação dos parâmetros do modelo é feita em dois passos: primeiro estima-se d através, por exemplo, de um modelo de regressão linear do logaritmo da função periodograma e, posteriormente, estimam-se os parâmetros auto-regressivos e de médias móveis. O estimador mais conhecido dentro dessa classe foi proposto por Geweke & Porter-Hudak (1983); variantes desse estimador foram desenvolvidas por Lobato & Robinson (1996), Reisen (1994), Robinson (1995a, 1995b), Velasco (2000), entre outros. Estudos comparativos de simulação sobre diferentes técnicas de estimação em diversos cenários podem ser encontrados, por exemplo, em Bisaglia & Guégan (1998), Haldrup & Nielsen (2007), Reisen, Abraham & Lopes (2001), Reisen, Abraham & Toscano (2000, 2002), Reisen, Rodrigues & Palma (2006) e Smith, Taylor & Yadav (1997).

Aplicações empíricas que empregam o modelo ARFIMA em economia e finanças podem ser encontradas em Baillie (1996), Barkoulas & Baum (1998), Bhardwaj & Swanson (2006), Cunado, Gil-Alana & Péres de Gracia (2004), Franses & Ooms (1997), Gil-Alana (2004), Reisen, Cribari-Neto & Jensen (2003); em climatologia pode-se citar Baillie & Chung (2002), entre outros. A recente publicação de Doukhan, Oppenheim & Taqqu (2003) apresenta uma revisão bibliográfica da teoria e aplicações de processos com longa dependência.

O estudo de modelos de memória longa na presença de *outliers* tem sido, recentemente, um assunto de muito interesse para pesquisadores da área. En especial, evidências

empíricas mostram que as estimativas do parâmetro de memória longa são significativamente alteradas pela presença de *outliers* do tipo aditivo na série temporal.

Haldrup & Nielsen (2007) mostraram, através de simulações de séries temporais com tamanhos de amostras relativamente pequenos, algumas conseqüências da presença de erros de medição, *outliers* e mudanças estruturais sobre as estimativas obtidas para o parâmetro de memória longa. Os resultados revelaram que os diferentes tipos de observações atípicas podem afetar seriamente as estimativas do grau de longa depedência. Por exemplo, a presença de *outliers* do tipo aditivo conduz a viés significativo nas estimativas do parâmetro de integração fracionária. O viés pode ser explicado por uma translação da função de densidade espectral do processo observado. Os autores concluíram que os estimadores semi-paramétricos obtidos por regressão apresentam viés relativamente menor quando o *bandwidth*, que corresponde ao número de freqüências utilizadas para o cálculo das estimativas, é reduzido. Para minimizar o efeito do *outlier* sobre a estimativa de d, os autores sugerem o uso da metodologia proposta por Sun & Phillips (2003), que se fundamenta na inclusão de um termo não-linear na regressão do logaritmo do periodograma. No mesmo contexto, Agostinelli & Bisaglia (2003) sugerem um método alternativo baseado em verossimilhança ponderada como uma modificação do estimador proposto por Beran (1994).

Mudanças na dinâmica de séries temporais afetam as estruturas de correlação e, conseqüentemente, causam viés nas estimativas dos parâmetros. Neste sentido, o presente trabalho propõe uma metodologia para estimar os parâmetros do modelo ARFIMA na presença de *outliers* do tipo aditivo. O método proposto para a estimação de d utiliza a estimativa robusta da função de autocovariância sugerida por Ma & Genton (2000) para obter a função periodograma. Nosso estimador é uma variante do estimador apresentado por Geweke & Porter-Hudak (1983) (GPH). Estudos numéricos evidenciam a robustez do estimador proposto na presença de *outliers* do tipo aditivo.

A presente dissertação está dividida em seis capítulos como descrito a seguir. No Capítulo 2 são apresentados conceitos básicos usados no estudo de séries temporais e processos estacionários; adicionalmente, são apresentados alguns resultados relativos a efeitos

de *outliers* na estimação de modelos estacionários. No Capítulo 3 apresentamos nosso estimador robusto para d. Resultados de simulação e aplicações encontram-se nos Capítulos 4 e 5. Finalmente, o Capítulo 6 contém as principais conclusões deste trabalho.

CAPÍTULO 2

Conceitos Básicos em Séries Temporais

Neste capítulo são introduzidos conceitos básicos utilizados na análise de séries temporais e processos estacionários. Em particular, é importante destacar o conceito de *estacionariedade*, no qual se encontram baseadas todas as técnicas de estimação e modelagem de séries temporais no domínio do tempo, através da função de autocovariância, e no domínio da freqüência, através da função de densidade espectral. Para detalhes, ver Brockwell & Davis (2006), Priestley (1983) e Wei (2005).

2.1 Processos estacionários

A seguir são apresentadas as condições de estacionariedade para um processo estocástico linear geral. Adicionalmente, são definidas as funções que caracterizam a dinâmica do processo nos domínios do tempo e da freqüência.

Definição 2.1.1. (*Processo estocástico*) Um processo estocástico é uma família de variáveis aleatórias $\{X_t(\omega)\}_{t\in T}$, definidas no mesmo espaço de probabilidade $(\Omega, \Im, P)$, onde $\omega \in \Omega$ e T é um conjunto arbitrário. Aqui, Ω é o espaço amostral, $\Im$ é uma σ-algebra de Ω e P é uma medida de probabilidade em $\Im$.

O conjunto T é comumente tomado como o conjunto dos números inteiros $\mathbb{Z} = \{0, \pm 1, \pm 2, \ldots\}$. Seguindo a definição anterior, uma *série temporal* é uma realização de um certo processo estocástico. Os dois primeiros momentos de $\{X_t(\omega)\}_{t\in\mathbb{Z}}$ (ou $\{X_t\}$) são definidos como

$$E[X_t] = \mu_t \text{ e } E(X_t - \mu_t)^2 = \sigma_t^2,$$

enquanto que a covariância entre X_t e X_{t+h} é

$$Cov(X_t, X_{t+h}) = E[(X_t - \mu_t)(X_{t+h} - \mu_{t+h})] \text{ para } h \in \mathbb{Z},$$

e a correlação é dada por

$$\frac{Cov(X_t, X_{t+h})}{\sqrt{\sigma_t^2 \sigma_{t+h}^2}} \text{ para } h \in \mathbb{Z}.$$

Definição 2.1.2. (estacionariedade) Um processo estocástico $\{X_t\}$ é dito ser (fracamente) estacionário se e somente se:

1. $E[X_t] = \mu$, para todo $t \in \mathbb{Z}$,
2. $E(X_t - \mu)^2 = \sigma^2$, $0 < \sigma^2 < \infty$, para todo $t \in \mathbb{Z}$,
3. $R(h) = Cov(X_t, X_{t+h})$ depende apenas de h, para todo $t \in \mathbb{Z}$.

As autocorrelações $\rho(h)$ são obtidas normalizando as autocovariâncias através da sua divisão pelo produto dos respectivos desvios padrão, i.e., $\rho(h) = \frac{R(h)}{R(0)}$. O exemplo mais simples de um processo estacionário é o processo de ruído branco (RB), definido como uma seqüência de variáveis aleatórias não-correlacionadas com média constante e variância constante (estritamente positiva e finita) ao longo do tempo.

Definição 2.1.3. (Processo linear geral) $\{X_t\}$ é um processo linear se pode ser representado como

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j \epsilon_{t-j}, \quad t \in \mathbb{Z},$$

onde $\{\epsilon_t\} \sim RB(0, \sigma_\epsilon^2)$ e $\{\psi_j\}$ é uma seqüência de constantes com $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$.

Definição 2.1.4. (Função geratriz de autocovariâncias) Seja $\{X_t\}$ um processo estacionário com função de autocovariâncias $R(h)$ que satisfaz $\sum_{h=-\infty}^{\infty} |R(h)| < \infty$. A função geratriz de autocovariâncias de $\{X_t\}$ é definida como

$$g(z) = \sum_{h=-\infty}^{\infty} R(h) z^h,$$

onde z é um escalar complexo.

Em particular, a função de densidade espectral (ou espectro) de $\{X_t\}$ é a função dada por

$$\begin{aligned} f(\lambda) &= \frac{1}{2\pi} g(e^{-i\lambda}) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} e^{-ih\lambda} R(h) \\ &= \frac{1}{2\pi} \left[R(0) + 2 \sum_{h=1}^{\infty} R(h) \cos(\lambda h) \right], \quad \lambda \in [-\pi, \pi], \end{aligned}$$

onde $e^{-i\lambda} = \cos(\lambda) - i \sin(\lambda)$ e $i = \sqrt{-1}$. Neste caso, note que a somabilidade de $|R(\cdot)|$ implica que $f(\lambda)$ converge absolutamente.

Estimação da média, autocovariâncias e espectro de um processo estacionário

Sejam $x_1, x_2, \ldots, x_n$ observações de um processo $\{X_t\}$ estacionário. Os estimadores usuais para $E[X_t] = \mu$ e $E(X_t - \mu)^2 = \sigma_X^2$ são $\bar{x} = \frac{1}{n} \sum_{t=1}^{n} x_t$ e $\widehat{R}(0) = \frac{1}{n} \sum_{t=1}^{n} (x_t - \bar{x})^2$, respectivamente. Um estimador razoável da função de autocovariâncias é

$$\widehat{R}(h) = \frac{1}{n} \sum_{t=1}^{n-h} (x_t - \bar{x})(x_{t+h} - \bar{x}), \qquad h = 0, \pm 1, \pm 2, \ldots, \pm(n-1),$$

e um estimador natural para $\rho(h)$ é $\widehat{\rho}(h) = \frac{\widehat{R}(h)}{\widehat{R}(0)}$.

No domínio da freqüência, um estimador assintoticamente não-viesado para a função de densidade espectral $f(\lambda)$ é o *periodograma*, dado por

$$I(\lambda) = \frac{1}{2\pi} \left[\widehat{R}(0) + 2 \sum_{h=1}^{n-1} \widehat{R}(h) \cos(\lambda h) \right]. \tag{2.1}$$

Um estimador consistente para o espectro de um processo estacionário é o *periodograma suavizado*, dado por

$$I_s(\lambda) = \frac{1}{2\pi} \sum_{h=-(n-1)}^{n-1} \kappa(h) \widehat{R}(h) \cos(\lambda h), \quad \lambda \in [-\pi, \pi], \tag{2.2}$$

onde $\kappa(\cdot)$ é uma função contínua e par. Na literatura, essa função é conhecida como "janela" e é útil para reduzir a contribuição de covariâncias provenientes de defasagens (h) elevadas. A "janela" mais simples é a chamada *janela periodograma truncado*:

$$\kappa(u) = \begin{cases} 1, & |u| \leq M, \\ 0, & |u| > M, \end{cases}$$

onde M ($< n-1$) é o parâmetro de truncamento. Existem outras propostas para a função $\kappa(\cdot)$ considerando diferentes ponderações; para detalhes ver Priestley (1983, p. 437).

2.2 Modelos de séries temporais

O estudo das séries temporais pode ser motivado pelo interesse em investigar o mecanismo gerador de um conjunto de dados observados ao longo do tempo para descrever sua dinâmica com o objetivo de gerar previsões acerca do seu comportamento futuro. Para tanto, são construídos modelos probabilísticos que pertencem a um domínio temporal previamente estabelecido. Tais modelos devem respeitar o princípio da parcimônia, ou seja, devem envolver o menor número possível de parâmetros.

A seguir, são descritos de forma geral alguns desses modelos e algumas de suas propriedades são apresentadas.

2.2.1 Processos auto-regressivos e de médias móveis

Seja $\{X_t\}$ um processo que satisfaz a equação em diferenças dada por

$$\Phi(B)X_t = \Theta(B)\epsilon_t, \tag{2.3}$$

onde $\{\epsilon_t\}$ é ruído branco, i.e., $\{\epsilon_t\} \sim RB(0, \sigma_\epsilon^2)$, B é o operador de defasagem definido como $B^m X_t = X_{t-m}$, $m = 1, \ldots, p$, $\Phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p$ e $\Theta(z) = 1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q$. O processo $\{X_t\}$ definido em (2.3) é chamado de processo auto-regressivo e de médias móveis, ARMA(p, q).

Definição 2.2.1. (Invertibilidade) Um processo $\{X_t\}$ com representação ARMA(p, q) é *invertível* se existem constantes $\{\pi_j\}$ tais que $\sum_{j=0}^{\infty} |\pi_j| < \infty$ e $\epsilon_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}$, para todo $t \in \mathbb{Z}$.

Seguindo as Definições 2.1.2 e 2.2.1 o processo (2.3) é estacionário e invertível se as raízes de $\Phi(z) = 0$ e $\Theta(z) = 0$ são não comuns e encontram-se fora do círculo unitário.

Definição 2.2.2. (Causalidade) Um processo $\{X_t\}$ com representação ARMA(p, q) é *causal*, ou função causal de $\{\epsilon_t\}$, se existem constantes $\{\psi_j\}$ tais que $\sum_{j=0}^{\infty} |\psi_j| < \infty$ e $X_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}$, para todo $t \in \mathbb{Z}$.

Note que as propriedades de invertibilidade e causalidade não são apenas do processo $\{X_t\}$, mas também da relação entre os processos $\{X_t\}$ e $\{\epsilon_t\}$ da definição da equação ARMA apresentada em (2.3). Invertibilidade e causalidade garantem que há uma solução única estacionária, com probabilidade um, para a equação ARMA.

Função de autocovariâncias e densidade espectral de um processo ARMA(p, q)

O cálculo da função de autocovariâncias para um processo $\{X_t\}$ com representação ARMA(p, q) causal é realizado através das equações

$$R(k) - \phi R(k-1) - \cdots - \phi_p R(k-p) = \sigma_\epsilon^2 \sum_{j=0}^{\infty} \theta_{k+j}\psi_j, \quad 0 \neq k < m,$$

$$R(k) - \phi R(k-1) - \cdots - \phi_p R(k-p) = 0, \quad k \geq m,$$

onde $m = \max(p, q+1)$, $\psi_j - \sum_{k=1}^{p} \phi_k \psi_{j-k} = \theta_j$, $j = 0, 1, 2, \ldots$. $\psi_j = 0$ para $j < 0$, $\theta_0 = 1$ e $\theta_j = 0$ para $j \notin \{0, 1, \ldots, q\}$; ver, e.g., Brockwell & Davis (2002, p. 88).

O espectro de $\{X_t\}$ é dado por

$$f_{ARMA}(\lambda) = \frac{\sigma_\epsilon^2}{2\pi} \frac{|\Theta(e^{-i\lambda})|^2}{|\Phi(e^{-i\lambda})|^2}, \quad \lambda \in [-\pi, \pi]. \tag{2.4}$$

2.2.2 Processos ARIMA(p, d, q)

Seja d um inteiro não-negativo. $\{X_t\}$ é um processo auto-regressivo integrado e de médias móveis ARIMA(p, d, q) se $Y_t = (1 - B)^d X_t$ é um processo ARMA(p, q) causal. Esta definição sugere que $\{X_t\}$ satisfaz a equação em diferenças da forma

$$\Phi(B)(1 - B)^d X_t = \Theta(B)\epsilon_t, \quad \{\epsilon_t\} \sim RB(0, \sigma_\epsilon^2).$$

2.2.3 Processos ARIMA(p, d, q) fracionários

No início da década de 80, Granger & Joyeux (1980) e Hosking (1981) propuseram uma extensão dos modelos ARIMA em que o parâmetro de integração assume valores fracionários. Esses modelos são conhecidos na literatura como ARFIMA e são utilizados na modelagem de séries que possuem memória longa ou longa dependência. A propriedade de memória longa ocorre em séries que apresentam correlações estatisticamente significativas mesmo para observações distantes, i.e., $\sum_{h=-\infty}^{\infty} |\rho(h)| = \infty$; equivalentemente, o espectro apresenta singularidade para freqüências próximas de 0, i.e., $f(\lambda) \to \infty$ quando $\lambda \to 0$. De maneira mais formal, o processo ARFIMA(p, d, q) é definido como a seguir:

Seja $d \in \mathbb{R}$. $\{X_t\}$ segue um processo ARFIMA(p, d, q) se satisfaz a equação em diferenças da forma

$$\Phi(B)(1 - B)^d X_t = \Theta(B)\epsilon_t, \tag{2.5}$$

com $\Phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$ e $\Theta(z) = 1 - \theta_1 z - \cdots - \theta_p z^q$, $\{\epsilon_t\}$ sendo um processo ruído branco com média 0 e variância σ_ϵ^2. O filtro de diferenciação fracionária $(1 - B)^d$ é definido pela expansão binomial

$$(1 - B)^d = \sum_{j=0}^{\infty} \pi_j B^j,$$

onde $\pi_j = \frac{\Gamma(j-d)}{\Gamma(j+1)\Gamma(-d)}$, $j = 0, 1, 2, \ldots$, e $\Gamma(\cdot)$ é a função gama definida em $\mathbb{R} - \mathbb{Z}^-$:

$$\Gamma(x) = \begin{cases} \int_0^\infty t^{x-1} e^{-t} dt, & x > 0, \\ \infty, & x = 0, \\ x^{-1}\Gamma(1 + x), & x < 0. \end{cases}$$

Quando $d \in (-0.5, 0.5)$ e as raízes dos polinômios $\Phi(z) = 0$ e $\Theta(z) = 0$ são não-comuns e estão fora do círculo unitário, o processo definido em (2.5) é estacionário e invertível e com função de densidade espectral dada por

$$f_{ARFIMA}(\lambda) = f_{ARMA}(\lambda) \left\{ 2 \sin\left(\frac{\lambda}{2}\right) \right\}^{-2d}, \quad \lambda \in [-\pi, \pi], \tag{2.6}$$

onde $f_{ARMA}(\lambda)$ está definida em (2.4).

Hosking (1981) mostrou que para valores $d \geq 0.5$ $\{X_t\}$ é não estacionário e invertível, e ainda que séries com representação ARFIMA(p, d, q) com $d \in (0, 0.5)$ apresentam estacionariedade e memória longa. Assim, no que se segue nós consideraremos o processo ARFIMA(p, d, q) com $d \in (0, 0.5)$.

Métodos de estimação do parâmetro d em modelos ARFIMA(p, d, q)

Existem vários estimadores do parâmetro de diferenciação fracionária d propostos na literatura, que podem ser classificados em paramétricos e semi-paramétricos. Os primeiros envolvem a estimação simultânea dos parâmetros do modelo, em geral utilizando o método de máxima verossimilhança; ver, e.g., Dahlhaus (1989), Fox & Taqqu (1986), Sowell (1992). Nos procedimentos semi-paramétricos, a estimação dos parâmetros do modelo é realizada em dois passos: primeiro estima-se o parâmetro de memória longa d, por exemplo, através de um modelo de regressão do logaritmo da função periodograma e, posteriormente, estimam-se os parâmetros auto-regressivos e de médias móveis. O estimador mais popular dentro dessa classe é o estimador proposto por Geweke & Porter-Hudak (1983) (GPH); variantes foram desenvolvidas por Reisen (1994), Robinson (1995a, 1995b), entre outros.

Estimador GPH

Seja $f(\lambda_j)$ a função definida em (2.6), para $\lambda_j = \frac{2\pi j}{n}$, $j = 0, 1, \ldots, \lfloor \frac{n}{2} \rfloor$, onde n é o tamanho amostral. O logaritmo de $f(\lambda_j)$ pode ser escrito como:

$$\ln f(\lambda_j) = \ln f_u(0) - d \ln \left\{ 2 \sin \left(\frac{\lambda_j}{2} \right) \right\}^2 + \ln \frac{f_u(\lambda_j)}{f_u(0)}, \tag{2.7}$$

onde $f_u(\lambda)$ é a densidade espectral de $U_t = (1 - B)^d X_t$. Aqui, $\lfloor . \rfloor$ denota a função parte inteira.

Geweke & Porter-Hudak (1983) sugerem um estimador semi-paramétrico de d, adicionando $\ln I(\lambda)$ em ambos os lados da equação (2.7) e considerando as freqüências próximas de zero, obtendo a aproximação:

$$\ln I(\lambda_j) \approx \ln f_u(0) - d \ln \left\{ 2 \sin \left(\frac{\lambda_j}{2} \right) \right\}^2 + \ln \frac{I(\lambda_j)}{f(\lambda_j)}, \tag{2.8}$$

que sugere a equação de regressão dada por

$$\ln I(\lambda_j) \approx \beta_0 + \beta_1 \ln \left\{ 2 \sin \left(\frac{\lambda_j}{2} \right) \right\}^2 + e_j, \quad j = 1, 2, \ldots, g(n),$$

onde $\beta_0 = \ln f_u(0)$, $\beta_1 = -d$ e $g(n)$ é o *bandwidth*, que corresponde ao número de freqüências utilizadas na regressão. Os erros $\{e_j\}$ são assintoticamente independentes com distribuição Gumbel de média 0 e variância $\frac{\pi^2}{6}$ (ver Geweke & Porter-Hudak (1983)). O estimador GPH é dado por

$$d_{GPH} = -\frac{\sum_{i=1}^{g(n)} (x_i - \bar{x}) \ln I(\lambda_i)}{\sum_{i=1}^{g(n)} (x_i - \bar{x})^2}, \tag{2.9}$$

onde $x_i = \ln \left\{ 2 \sin \left(\frac{\lambda_j}{2} \right) \right\}^2$. Geweke & Porter-Hudak (1983) sugerem considerar $g(n) = n^\alpha$, $0 < \alpha < 1$. Algumas propriedades assintóticas do estimador dado em (2.9) foram derivadas por Hurvich, Deo & Brodsky (1998) e Velasco (2000).

2.3 Outliers em séries temporais

Na análise de séries temporais é comum encontrar observações influenciadas por eventos externos que podem facilmente afetar os procedimentos convencionais de análise, nomeadamente podem enviesar significativamente as estimativas dos parâmetros do modelo. O efeito dessas observações, conhecidas como dados atípicos ou *outliers*, é no entanto, muitas vezes omitido pela falta do conhecimento de métodos que podem ser usados para detectá-los e para acomodá-los ao processo subjacente à série.

Fox (1972) introduziu o conceito de *outliers* no contexto de séries temporais, tendo considerado dois tipos de observações atípicas a saber: aditivo (AO – "Additive Outlier") e inovador (IO – "Innovational Outlier"). Como uma extensão do trabalho de Fox (1972), Chang et al. (1988), Chen & Liu (1993*a*, 1993*b*) e Tsay (1986) consideraram as alterações na estrutura da série, nomeadamente, alteração de nível permanente (LS – "Level Shift") e alterações temporárias (TC – "Temporary Change"). Os mesmos autores, adotando a formulação de Fox (1972), consideraram tais alterações como casos particulares do modelo geral de intervenção de Box & Tiao (1975), a partir de um processo linear estacionário de

segunda ordem $\{y_t\}$, escrevendo o processo contaminado por *outliers* $\{z_t\}$ como

$$z_t = y_t + \sum_{i=1}^{m} \xi_i(B)\omega_i I_t^{(T_i)}, \tag{2.10}$$

onde m é o número total de outliers, o parâmetro desconhecido ω_i representa a magnitude do i-ésimo *outlier* no tempo T_i, $I_t^{(T_i)}$ é uma variável aleatória satisfazendo

$$I_t^{(T_i)} = \begin{cases} \pm 1, & \text{se } t = T_i, \\ 0, & \text{se } t \neq T_i, \end{cases}$$

e $\xi_i(B)$ determina a dinâmica do *outlier* no tempo T_i, de acordo com o seguinte esquema:

$$\begin{aligned} AO &: \xi_i(B) = 1, \\ IO &: \xi_i(B) = \frac{\Theta(B)}{\Phi(B)}, \\ LS &: \xi_i(B) = \frac{1}{1-B}, \\ TC &: \xi_i(B) = \frac{1}{1-\delta B}, \quad 0 < \delta < 1, \end{aligned}$$

onde $\Phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p$ e $\Theta(z) = 1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q$. As variáveis y_t e $I_t^{(T_i)}$ são independentes para cada valor de t.

Neste trabalho serão considerados unicamente *outliers* do tipo aditivo por serem os mais comuns e por afetarem significativamente análises de séries temporais.

2.3.1 Efeitos de outliers em processos estacionários

A seguir, são apresentados alguns resultados referentes aos efeitos de *outliers* aditivos sobre as funções de densidade espectral e de correlação do processo $\{z_t\}$.

Proposição 1. *Seja $\{z_t\}$ representado pelo modelo (2.10) com $\xi_i(B) = 1$. Se $I_t^{(T_i)}$ é uma variável Bernoulli tal que $Pr\left(I_t^{(T_i)} = -1\right) = Pr\left(I_t^{(T_i)} = 1\right) = \frac{p_i}{2}$ e $Pr\left(I_t^{(T_i)} = 0\right) = 1 - p_i$ e supondo que $T_j \neq 2T_i$ para cada $i, j = 1, 2, \ldots, m$. Então,*

i. A função de autocovariância (FACOV) do processo $\{z_t\}$ é dada por

$$R_z(h) = \begin{cases} R_y(0) + \sum_{i=1}^{m} \omega_i^2 p_i, & se \;\; h = 0, \\ R_y(h), & se \;\; h \neq 0. \end{cases} \tag{2.11}$$

ii. A função de densidade espectral do processo $\{z_t\}$ *é dada por*

$$f_z(\lambda) = f_y(\lambda) + \frac{1}{2\pi}\sum_{i=1}^{m}\omega_i^2 p_i. \tag{2.12}$$

Prova. *i.* Seja $E[y_t] = \mu$. Dado que $R_z(h) = E[z_t z_{t+h}] - E[z_t]E[z_{t+h}]$, então

$$\begin{aligned}
R_z(h) &= E\left[y_t + \sum_{i=1}^{m}\omega_i I_t^{(T_i)}\right]\left[y_{t+h} + \sum_{i=1}^{m}\omega_i I_{t+h}^{(T_i)}\right] - \mu^2 \\
&= R_y(h) + \sum_{i=1}^{m}\omega_i E\left[I_{t+h}^{(T_i)} y_t\right] + \sum_{i=1}^{m}\omega_i E\left[I_t^{(T_i)} y_{t+h}\right] + \sum_{i=1}^{m}\omega_i^2 E\left[I_t^{(T_i)} I_{t+h}^{(T_i)}\right] \\
&= \begin{cases} R_y(0) + \sum_{i=1}^{m}\omega_i^2 E\left[I_t^{(T_i)}\right]^2, & \text{se } \ h = 0, \\ R_y(h), & \text{se } \ h \neq 0. \end{cases}
\end{aligned}$$

ii. Como $f_z(\lambda) = \dfrac{1}{2\pi}\sum\limits_{h=-\infty}^{\infty} R_z(h)e^{-ih\lambda}$, então

$$\begin{aligned}
f_z(\lambda) &= \frac{1}{2\pi}\left[\sum_{h=-\infty}^{-1} R_z(h)e^{-ih\lambda} + R_z(0) + \sum_{h=1}^{\infty} R_z(h)e^{-ih\lambda}\right] \\
&= \frac{1}{2\pi}\left[R_z(0) + 2\sum_{h=1}^{\infty} R_z(h)e^{-ih\lambda}\right].
\end{aligned}$$

Assim, por (i),

$$\begin{aligned}
f_z(\lambda) &= \frac{1}{2\pi}\left[R_y(0) + \sum_{i=1}^{m}\omega_i^2 p_i + 2\sum_{h=1}^{\infty} R_y(h)e^{-ih\lambda}\right] \\
&= \frac{1}{2\pi}\left[R_y(0) + 2\sum_{h=1}^{\infty} R_y(h)e^{-ih\lambda}\right] + \frac{1}{2\pi}\sum_{i=1}^{m}\omega_i^2 p_i.
\end{aligned}$$

□

Os resultados na Proposição 1 revelam que há um aumento na variância de $\{z_t\}$, o que implica diminuição nos valores das autocorrelações e perda de informação sobre a estrutura de autocorrelação do processo. A densidade espectral de $\{z_t\}$ é caracterizada por uma translação provocada em função dos $\omega_1, \omega_2, \ldots, \omega_m$. A propriedade de perda de memória apresentada pela FACOV é abordada por Chan (1992), através do limite da função de autocorrelação amostral (ver Proposição 3).

O resultado a seguir evidencia o efeito de um *outlier* sobre a função de autocovariância amostral e sobre o periodograma.

Proposição 2. *Seja $z_1, z_2, \ldots, z_n$ um conjunto de observações geradas pelo modelo (2.10) com $\xi_i(B) = 1$. Tomando $m = 1$, temos que:*

i. A FACOV amostral é dada por

$$\widehat{R}_z(h) = \widehat{R}_y(h) \pm \frac{\omega}{n}(y_{T-h} + y_{T+h} - 2\bar{y}) - \frac{\omega^2}{n^2} + \frac{\omega^2}{n}\delta(h) + o_p\left(\frac{1}{n}\right), \tag{2.13}$$

$$\text{onde } \widehat{R}_y(h) = \frac{1}{n}\sum_{t=1}^{n-h}(y_t - \bar{y})(y_{t+h} - \bar{y}) \text{ e } \delta(h) = \begin{cases} 1, & \text{quando } h = 0, \\ 0, & \text{caso contrário.} \end{cases}$$

ii. O periodograma é dado por

$$I_z(\lambda) = I_y(\lambda) + \frac{1}{2\pi}\Delta(\omega), \tag{2.14}$$

onde

$$\Delta(\omega) = \frac{\omega^2}{n} - \frac{\omega^2}{n^2}\frac{\sin(n - \frac{1}{2})\lambda}{\sin(\frac{\lambda}{2})} \pm \frac{2\omega}{n}(y_T - \bar{y}) \pm \frac{2\omega}{n}\sum_{h=1}^{n-1}(y_{T-h} + y_{T+h} - \bar{y})\cos(h\lambda) + o_p\left(\frac{1}{n}\right).$$

Prova. *i.* A média amostral do processo $\{z_t\}$ é dada por

$$\bar{z} = \frac{1}{n}\sum_{t=1}^{n} z_t = \frac{1}{n}\sum_{t=1}^{n}\left(y_t + \omega I_t^{(T)}\right) = \frac{1}{n}\sum_{t=1}^{n} y_t + \frac{\omega}{n}\sum_{t=1}^{n} I_t^{(T)} = \bar{y} \pm \frac{\omega}{n}. \tag{2.15}$$

Logo,

$$\begin{aligned}
\widehat{R}_z(h) &= \frac{1}{n}\sum_{t=1}^{n-h}\left\{(y_t - \bar{y}) + \left(\omega I_t^{(T)} \mp \frac{\omega}{n}\right)\right\}\left\{(y_{t+h} - \bar{y}) + \left(\omega I_{t+h}^{(T)} \mp \frac{\omega}{n}\right)\right\} \\
&= \frac{1}{n}\sum_{t=1}^{n-h}(y_t - \bar{y})(y_{t+h} - \bar{y}) + \frac{1}{n}\sum_{t=1}^{n-h}(y_t - \bar{y})\left(\omega I_{t+h}^{(T)} \mp \frac{\omega}{n}\right) \\
&+ \frac{1}{n}\sum_{t=1}^{n-h}\left(\omega I_t^{(T)} \mp \frac{\omega}{n}\right)(y_{t+h} - \bar{y}) + \frac{1}{n}\sum_{t=1}^{n-h}\left(\omega I_t^{(T)} \mp \frac{\omega}{n}\right)\left(\omega I_{t+h}^{(T)} \mp \frac{\omega}{n}\right) \\
&= \widehat{R}_y(h) \pm \frac{\omega}{n}y_{T-h} \mp \frac{\omega}{n^2}\sum_{t=1}^{n-h} y_t \mp \frac{\omega}{n}\bar{y} \pm \frac{\omega}{n}\left(1 - \frac{h}{n}\right)\bar{y} \pm \frac{\omega}{n}y_{T+h} \mp \frac{\omega}{n}\bar{y} \\
&\mp \frac{\omega}{n^2}\sum_{t=1}^{n-h} y_{t+h} \pm \frac{\omega}{n}\bar{y}\left(1 - \frac{h}{n}\right) - \frac{\omega^2}{n^2}\sum_{t=1}^{n-h}\left(I_t^{(T)} - I_{t+h}^{(T)}\right) \pm \frac{\omega^2}{n}\left(1 - \frac{h}{n}\right) \\
&= \widehat{R}_y(h) \pm \frac{\omega}{n}(y_{T-h} - \bar{y}) \pm \frac{\omega}{n}(y_{T+h} - \bar{y}) - \frac{\omega^2}{n^2} + \frac{\omega^2}{n}\delta(h) + o_p\left(\frac{1}{n}\right).
\end{aligned}$$

ii. Pela definição do periodograma tem-se que

$$I_z(\lambda) = \frac{1}{2\pi}\left\{\widehat{R}_z(0) + 2\sum_{h=1}^{n-1}\widehat{R}_z(h)\cos(h\lambda)\right\}, \quad \lambda \in [-\pi, \pi]. \tag{2.16}$$

Logo, pelo item (i),

$$\begin{aligned} I_z(\lambda) &= \frac{1}{2\pi}\left\{\widehat{R}_y(0) \pm \frac{2\omega}{n}(y_T - \bar{y}) - \frac{\omega^2}{n^2} + \frac{\omega^2}{n}\right. \\ &\quad \left. +2\sum_{h=1}^{n-1}\left(\widehat{R}_y(h) \pm \frac{\omega}{n}(y_{T-h} + y_{T+h} - 2\bar{y}) - \frac{\omega^2}{n^2}\right)\cos(h\lambda)\right\} \\ &= \frac{1}{2\pi}\left\{\widehat{R}_y(0) + 2\sum_{h=1}^{n-1}\widehat{R}_y(h)\cos(h\lambda)\right\} + \frac{1}{2\pi}\Delta(\omega), \end{aligned}$$

onde

$$\begin{aligned} \Delta(\omega) &= \frac{\omega^2}{n} - \frac{\omega^2}{n^2} \pm \frac{2\omega}{n}(y_T - \bar{y}) \\ &\quad - \frac{2\omega^2}{n^2}\sum_{h=1}^{n-1}\cos(h\lambda) \pm \frac{2\omega}{n}\sum_{h=1}^{n-1}(y_{T-h} + y_{T+h} - 2\bar{y})\cos(h\lambda) \\ &= \frac{\omega^2}{n} - \frac{\omega^2}{n^2} \pm \frac{2\omega}{n}(y_T - \bar{y}) - \frac{2\omega^2}{n^2}\frac{\sin(n - \frac{1}{2})\lambda}{2\sin(\frac{\lambda}{2})} + \frac{\omega^2}{n^2} \\ &\quad \pm \frac{2\omega}{n}\sum_{h=1}^{n-1}(y_{T-h} + y_{T+h} - 2\bar{y})\cos(h\lambda) + o_p\left(\frac{1}{n}\right). \end{aligned}$$

□

As funções (2.13) e (2.14), apresentadas na Proposição 2, evidenciam a influência do *outlier* sobre as funções amostrais (FACOV e periodograma) de uma série temporal com um dado atípico. Quando $h = 0$, o termo adicional $\dfrac{\omega^2}{n}$ em (2.13) implica diminuição nos valores obtidos pela função de autocorrelação amostral (FAC), definida como $\widehat{\rho}_z(h) = \frac{\widehat{R}_z(h)}{\widehat{R}_z(0)}$.

Os resultados assintóticos na Proposição 3, apresentados também por Chan (1992, 1995), mostram que a FAC é consideravelmente alterada pela presença de outliers. Os resultados evidenciam perda na informação referente à estrutura de autocorrelação do processo, o que ocasiona aumento nos erros de estimação dos parâmetros do modelo.

Proposição 3. *(Chan(1992, 1995)) Seja $z_1, z_2, \ldots, z_n$ um conjunto de observações geradas pelo modelo (2.10) com $\xi_i(B) = 1$.*

i. Para $m = 1$,

$$\lim_{n\to\infty} \lim_{\omega\to\infty} \widehat{\rho}_z(h) = 0.$$

ii. Para $m = 2$ e $T_2 = T_1 + l$, tal que $h < T_1 < T_1 + l < n - h$, temos que

$$\lim_{n\to\infty} \left\{ plim_{\substack{\omega_1\to\infty \\ \omega_2\to\pm\infty}} \widehat{\rho}_z(h) \right\} = \begin{cases} 0, & se\ h \neq l \\ \pm 0.5, & se\ h = l. \end{cases}$$

As Proposições 1 a 3 mostram que a presença de *outliers* em processos lineares estacionários pode afetar seriamente a inferência realizada, alterando sua dinâmica e podendo conduzir a conclusões errôneas sobre sua natureza. Os resultados aqui apresentados corroboram aqueles obtidos por Chang et al. (1988) e por Chen & Liu (1993*b*) para processos ARMA.

O Corolario 1 a seguir apresenta o espectro do processo $\{z_t\}$ quando $\{y_t\}$ segue uma representação ARFIMA(p, d, q).

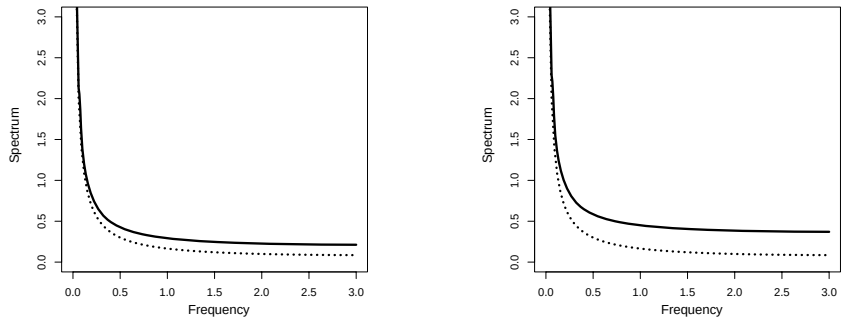
Corolario 1. *Seja $\{y_t\}_{t\in\mathbb{Z}}$ um processo estacionário e invertível ARFIMA(p, d, q). Seja $\{z_t\}_{t\in\mathbb{Z}}$ representado pelo modelo $z_t = y_t + \sum_{i=1}^{m} \omega_i I_t^{(T)}$, onde m é o número total de outliers, o parâmetro desconhecido ω_i representa a magnitude do i-ésimo outlier no tempo T_i e $I_t^{(T_i)}$ é uma variável Bernoulli tal que $Pr\left(I_t^{(T_i)} = -1\right) = Pr\left(I_t^{(T_i)} = 1\right) = \frac{p_i}{2}$ e $Pr\left(I_t^{(T_i)} = 0\right) = 1 - p_i$. O espectro de $\{z_t\}$ é dado por*

$$f_z(\lambda) = \frac{\sigma_\epsilon^2}{2\pi} \frac{|\Theta(e^{-i\lambda})|^2}{|\Phi(e^{-i\lambda})|^2} \left\{ 2\sin\left(\frac{\lambda}{2}\right) \right\}^{-2d} + \frac{1}{2\pi} \sum_{i=1}^{m} \omega_i^2 p_i.$$

O Corolario 1 é conseqüência imediata da Proposição 1.

Como ilustração do Corlolario 1, a Figura 2.1 mostra a translação no espectro de $\{z_t\}$ quando $d = 0.45$ e $m = 1$, considerando magnitudes $\omega = 10, 15$, sendo $y_t = (1-B)^{-d}\epsilon_t$, onde $\{\epsilon_t\}$ é ruído branco com média 0 e variância $\sigma_\epsilon^2 = 1$.

A translação da função de densidade espectral, causada pelos dados atípicos, provoca viés nas estimativas obtidas do parâmetro d, evidenciando a sensibilidade dos estimadores



(a) $\omega = 10$ (b) $\omega = 15$

Figura 2.1: Função de densidade espectral do processo ARFIMA(0,d,0), com $d = 0.45$, contaminado por um outlier do tipo aditivo. A linha pontilhada representa o espectro do processo sem outlier e a linha contínua representa o espectro do processo contaminado.

quando a série temporal contém *outliers*. Em particular, os estudos numéricos apresentados neste trabalho mostram que o estimador sugerido por Geweke & Porter-Hudak (1983) subestima o parâmetro de memória longa na presença de dados atípicos.

CAPÍTULO 3

Estimador GPH Robusto

Na Subseção 2.3.1 foram discutidas as propriedades das funções de autocovariâncias e densidade espectral na presença de *outliers*. Observou-se o viés causado por este tipo de dados na função de autocorrelação amostral e, conseqüentemente, na função periodograma. Os vieses dos estimadores do parâmetro d, em séries temporais de memória longa na presença de dados atípicos, motivam o desenvolvimento de inferência robusta à presença de dados atípicos.

A seguir serão propostos um estimador robusto do parâmetro fracionário d para o processo observado $\{z_t\}$ e uma metodologia para estimação robusta dos parâmetros auto-regressivos e de médias móveis do modelo.

3.1 Estimação robusta

O estimador proposto do parâmetro d (d_{GPHR}) é uma variante do estimador sugerido por Geweke & Porter-Hudak (1983). Para obter o estimador proposto, utilizamos a função de autocorrelação robusta, apresentada por Ma & Genton (2000), com vistas a obter um estimador robustificado do espectro do processo.

3.1.1 Estimador robusto da função de autocovariâncias

Ma & Genton (2000) propuseram um estimador robusto para a FACOV com base na aproximação de escala para a covariância entre duas variáveis aleatórias e no estimador $Q_n(\cdot)$, proposto por Rousseeuw & Croux (1993). O estimador de escala para covariância é dado por

$$\mathrm{COV(X,Y)} = \frac{1}{4ab}[\mathrm{var}(aX + bY) - \mathrm{var}(aX - bY)], \tag{3.1}$$

onde X e Y são variáveis aleatorias, $a = \frac{1}{\sqrt{\mathrm{var}(X)}}$ e $b = \frac{1}{\sqrt{\mathrm{var(Y)}}}$ (Huber 2004). O estimador $Q_n(\cdot)$ é baseado na k-ésima estatística de ordem das $\binom{n}{2}$ distâncias $\{|z_i - z_j|, i < j\}$, sendo dado por

$$Q_n(z) = c \times \{|z_i - z_j|; i < j\}_{(k)}, \tag{3.2}$$

onde $z = (z_1, z_2, \ldots, z_n)'$ é o vetor de dados, c é uma constante usada para garantir consistência ($c = 2.2191$ para distribuição normal) e $k = \left\lfloor \frac{\binom{n}{2}+2}{4} \right\rfloor + 1$. Este procedimento de cálculo implica elevado custo computacional, o qual pode ser diminuído através do uso do algoritmo descrito por Croux & Rousseeuw (1992).

Apartir de (3.1) e (3.2), o estimador robusto para a FACOV é calculado da forma

$$\widetilde{R}(h) = \frac{1}{4}\left[Q^2_{n-h}(u+v) - Q^2_{n-h}(u-v)\right], \tag{3.3}$$

onde u e v são vetores que contêm as primeiras e as últimas $n-h$ observações, respectivamente. O estimador robusto para a função de autocorrelação é dado por

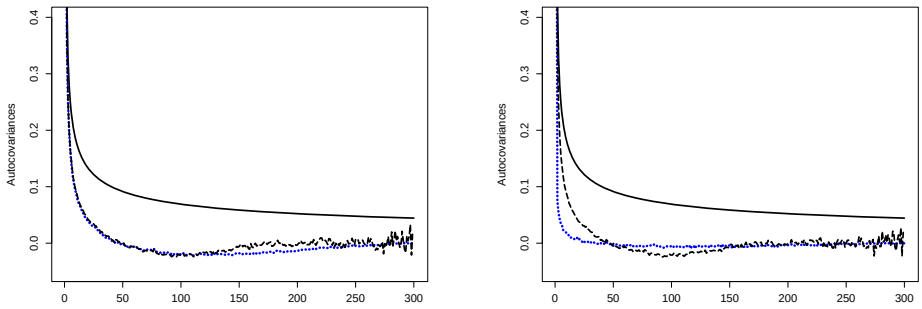
$$\widetilde{\rho}(h) = \frac{Q^2_{n-h}(u+v) - Q^2_{n-h}(u-v)}{Q^2_{n-h}(u+v) + Q^2_{n-h}(u-v)}. \tag{3.4}$$

Ele mantém a propriedade $|\widetilde{\rho}(h)| \leq 1$ e não depende do valor de c.

Ma & Genton (2000) derivam algumas propriedades de robustez do estimador (3.3) e mostram que a variância do mesmo não tem forma fechada.

Como ilustração, a Figura 3.1 mostra o comportamento das FAC teórica e amostrais, clássica e robusta, obtidas para uma série temporal gerada do modelo ARFIMA$(0, d, 0)$, com $d = 0.3$ e tamanho de amostra $n = 300$. A série contaminada é construída substituindo 5% das observações por *outliers* do tipo aditivo com magnitude $\omega = 10$. Note que, na ausência

de dados atípicos (Figura 3.1(a)), as funções amostrais fornecem estimativas relativamente próximas para os primeiros $\lfloor \frac{n}{2} \rfloor$ valores de h, como verificado por Ma & Genton (2000). No caso contaminado (Figura 3.1(b)), a FAC clássica evidencia o viés indicado nas Proposições 1 e 2. Em ambos casos, as estimativas obtidas através da autocorrelação amostral robusta, para valores grandes de h, apresentam comportamento irregular, que é explicado pelo cálculo dos quantis na função $Q_n(\cdot)$.



(a) Dados sem contaminação (b) Dados contaminados

Figura 3.1: Funções de autocorrelação teórica (linha continua) e amostrais, clássica (linha pontilhada) e robusta (linha tracejada), para uma série temporal gerada por um processo ARFIMA$(0, d, 0)$ com $d = 0.3$ e tamanho de amostra $n = 300$.

3.1.2 Estimador robusto da função periodograma

As propriedades de robustez do estimador da FAC, apresentado na Subseção 3.1.1, motivam a utilização do mesmo no cálculo da função periodograma, como descrito a seguir.

Seja $\widetilde{I}(\lambda)$ o estimador robusto da função de densidade espectral dado por

$$\widetilde{I}(\lambda) = \frac{1}{2\pi} \sum_{h=-M}^{M} \kappa\left(\frac{h}{M}\right) \widetilde{R}(h) \cos(h\lambda), \tag{3.5}$$

onde $\kappa(s)$ é função contínua para $s \in (-1, 1)$, com $\kappa(0) = 1$, $\kappa(-s) = \kappa(s)$, e $\widetilde{R}(h)$ é a função de autocovariâncias definida em (3.3). O parâmetro M é função de n, tal que $M \to \infty$ e $\frac{M}{n} \to 0$ quando $n \to \infty$. Note que, para valores grandes de h (h maior ou igual que $\lfloor \frac{3}{4}n \rfloor$), $\widetilde{\rho}(h)$ possui comportamento irregular. Portanto, como descrito na Seção 2.1, torna-se necessária a utilização de uma janela para reduzir a contribuição dos últimos

termos da função de autocovariâncias robusta. A função $\kappa(\cdot)$ é definida como

$$\kappa(s) = \begin{cases} 1, & |s| \leq 1, \\ 0, & |s| > 1, \end{cases}$$

e $M = n^{\beta}$, $0 < \beta < 1$. $\kappa(s)$ corresponde à janela usual do periodograma, como definido em (2.2). O estimador definido em (3.5) será chamado de *pseudo-periodograma robusto truncado* por não apresentar, para tamanhos de amostra finitos, as mesmas propriedades do periodograma usual definido em (2.1).

3.1.3 Estimador robusto do parâmetro de memória longa

Como no caso do estimador GPH, o estimador robusto de d é obtido utilizando a aproximação em (2.8), que sugere a equação de regressão

$$\ln \widetilde{I}(\lambda_j) \approx \beta_0 + \beta_1 \ln \left\{ 2 \sin \left(\frac{\lambda_j}{2} \right) \right\}^2 + e_j^*, \quad j = 1, 2, \ldots, g(n),$$

onde $\beta_0 = \ln f_u(0)$, $\beta_1 = -d$ e os termos de erro $\{e_j^*\}$ são assintoticamente não correlacionados. O estimador GPH robusto pode ser escrito como

$$d_{GPHR} = -\frac{\sum_{i=1}^{g(n)} (x_i - \bar{x}) \ln \widetilde{I}(\lambda_i)}{\sum_{i=1}^{g(n)} (x_i - \bar{x})^2}, \tag{3.6}$$

onde $x_i = \ln \left\{ 2 \sin \left(\frac{\lambda_j}{2} \right) \right\}^2$ e $g(n)$ é como definido anteriormente.

A estimação robusta do parâmetro de memória longa, na série contaminada, não elimina o efeito dos *outliers* nos dados. Por esta razão sugere-se o uso de uma metodologia alternativa para o cálculo das estimativas dos parâmetros auto-regressivos e de médias móveis do modelo ajustado através da função de autocorrelação robusta em (3.4).

3.2 Procedimento para estimação dos parâmetros do modelo ARFIMA

A identificação e a estimação dos parâmetros auto-regressivos e de médias móveis podem ser realizadas através dos seguintes passos:

1. Estimar d no modelo ARFIMA(p, d, q) utilizando o estimador GPH robusto, por exemplo, $\hat{d} = d_{GPHR}$.

2. Calcular $\widehat{U}_t = (1 - B)^{\hat{d}} y_t$.

3. Através da relação funcional $\Phi(B)\widehat{U}_t = \Theta(B)\epsilon_t$, utilizar o procedimento Box-Jenkins (Box et al. 1994), para identificação e estimação dos parâmetros $\phi_1, \phi_2, \ldots, \phi_p$ e $\theta_1, \theta_2, \ldots, \theta_q$ no modelo ARMA(p, q).

   Nota: A estimação obtida no Passo 1 não elimina o efeito do *outlier* na série. Portanto, neste passo sugere-se usar $\widetilde{\rho}(\lambda)$ no sistema de equações de Yule-Walker para obtenção das estimativas dos parâmetros do modelo ARMA(p, q).

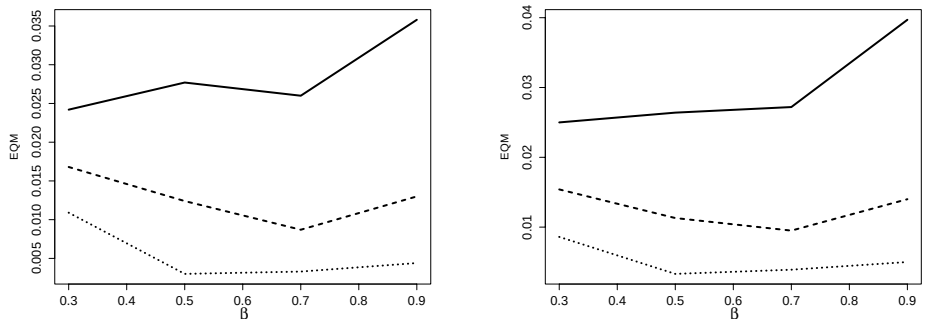
4. Fazer verificação da adequação do modelo (por exemplo, análise de resíduos).

CAPÍTULO 4

Resultados de Simulação

Este capítulo dedica-se à apresentação de resultados de simulação com o objetivo de analisar o comportamento do estimador robusto proposto em amostras finitas. Nos estudos de simulação foram comparadas as estimativas obtidas através dos estimadores d_{GPH} e d_{GPHR}, definidos em (2.9) e (3.6), respectivamente. Os dados gerados são provenientes de um processo ARFIMA(p, d, q) para $d = 0.3$ e 0.45, com tamanhos de amostra $n = 100, 300, 800$ e 3000. Nas séries contaminadas, os dados atípicos foram gerados utilizando probabilidades de ocorrência $p = 0.05$ e 0.1 e magnitudes $\omega = 3, 5$ e 10.

O valor de β, necessário para o cálculo de $\kappa(s)$ utilizado na obtenção do estimador robusto, foi selecionado empiricamente através da minimização dos erros quadráticos médios (EQM) das estimativas do parâmetro de integração fracionária. Como mostra a Figura 4.1, $\beta = 0.7$ é um valor adequado. O valor do *bandwidth* para d_{GPH} e d_{GPHR} foi calculado utilizando $\alpha = 0.5, 0.7$. Os resultados de simulação foram obtidos através da linguagem de programação `Ox` versão `4.02`, realizando 1000 repetições para cada experimento de Monte Carlo e calculando a média, o desvio padrão e o EQM das estimativas.



(a) Dados sem contaminação (b) Dados contaminados

Figura 4.1: Seleção do bandwidth do estimador robusto pelo critério do menor EQM para um ARFIMA$(0, d, 0)$ com $d = 0.3$ e tamanhos de amostra 100 (linha sólida), 300 (linha tracejada) e 800 (linha pontilhada).

4.1 Modelo ARFIMA$(0, d, 0)$

Na Tabela 4.1 são apresentadas as estimativas obtidas utilizando d_{GPH} e d_{GPHR} para $d = 0.3, 0.45$ e $\alpha = \beta = 0.7$. As colunas correspondentes a d_{GPHc} e d_{GPHRc} correspondem às estimativas do parâmetro d obtidas apartir das séries contaminadas.

Os resultados numéricos evidenciam a sensibilidade do estimador GPH à presença de *outliers*; esse estimador subestima o verdadeiro valor do parâmetro d, causando aumento significativo no viés das estimativas obtidas quando a série temporal contém *outliers*. Como mostrado na Seção 2.3.1, os resultados teóricos estão em consonância com a evidência numérica aqui apresentada e sugerem que não é adequada a utilização do estimador d_{GPH} quando a série temporal possui dados atípicos.

As propriedades estatísticas do estimador robusto sugerem que ele é um estimador alternativo atraente para o parâmetro de integração fracionária em séries temporais com *outliers*. Quando o tamanho de amostra aumenta as propriedades de d_{GPHR} melhoram significativamente, como ilustrado na Figura 4.2.

Os resultados contidos na Tabela 4.2 sugerem que o viés do estimador d_{GPHc} é função da magnitude ω. O viés do estimador d_{GPHRc} mantém-se, por outro lado, praticamente inalterado, mesmo para magnitudes relativamente grandes.

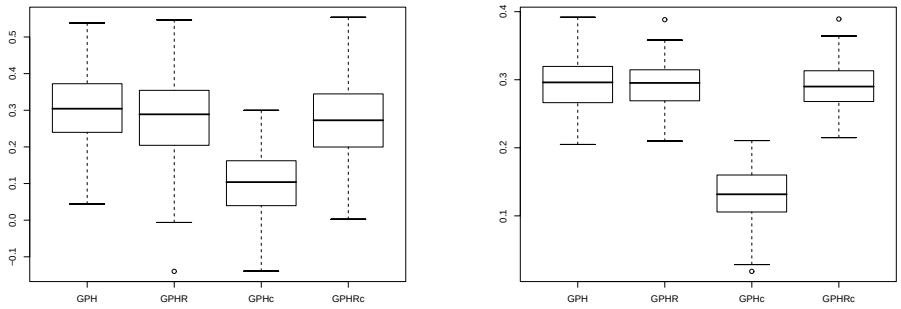
d	n		d_{GPH}	d_{GPH_c}	d_{GPHR}	d_{GPHR_c}
	100	média	0.2988	0.1134	0.2584	0.2449
		desvio padrão	0.1735	0.1619	0.1558	0.1556
		viés	−0.0012	−0.1866	−0.0416	−0.0551
		EQM	0.0301	0.0610	0.0260	0.0272
	300	média	0.3062	0.1007	0.2907	0.2837
0.30		desvio padrão	0.1005	0.0978	0.0926	0.0960
		viés	0.0062	−0.1993	−0.0093	−0.0163
		EQM	0.0101	0.0493	0.0087	0.0095
	800	média	0.3003	0.1184	0.2949	0.2869
		desvio padrão	0.0679	0.0715	0.0573	0.0610
		viés	0.0003	−0.1816	−0.0051	−0.0131
		EQM	0.0046	0.0381	0.0033	0.0039
	100	média	0.4561	0.1923	0.3975	0.3778
		desvio padrão	0.1722	0.1727	0.1506	0.1433
		viés	0.0061	−0.2577	−0.0525	−0.0722
		EQM	0.0297	0.0962	0.0254	0.0258
	300	média	0.4594	0.2015	0.4329	0.4233
0.45		desvio padrão	0.0986	0.0976	0.1041	0.1013
		viés	0.0094	−0.2485	−0.0171	−0.0267
		EQM	0.0098	0.0713	0.0111	0.0110
	800	média	0.4620	0.2306	0.4457	0.4349
		desvio padrão	0.0688	0.0809	0.0562	0.0576
		viés	0.0121	−0.2194	−0.0043	−0.0151
		EQM	0.0049	0.0547	0.0032	0.0035

Tabela 4.1: Estimativas do parâmetro d obtidas para um modelo ARFIMA$(0, d, 0)$ com $\alpha = \beta = 0.7$ e $\omega = 0, 10$.

4.2 Modelo ARFIMA(p, d, q)

A estimação dos parâmetros para o modelo ARFIMA(p, d, q) foi realizada seguindo a metodologia sugerida na Seção 3.2. Na Tabela 4.3 são apresentadas as estimativas do parâmetro d considerando um modelo ARFIMA$(1, d, 0)$ para $d = 0.3$ e $\phi = 0.2, 0.5, 0.7$. Os resultados obtidos mostram uma tendência do estimador d_{GPH} a superestimar o valor do parâmetro d para valores de ϕ próximos de 1. Isto pode ser explicado através do cálculo da densidade espectral do processo ARMA$(1, 0)$, dada por

$$f_y(\lambda) = \frac{\sigma_\epsilon^2}{2\pi(1 - 2\phi\cos\lambda + \phi^2)},$$



(a) Tamanho de amostra $n = 300$ (b) tamanho de amostra $n = 3000$

Figura 4.2: Estimativas do parâmetro d para tamanhos de amostra $n = 300, 3000$ obtidas pelos estimadores para dados contaminados (GPHc e GPHRc, respectivamente) e sem contaminação (GPH e GPHR).

ω	n		d_{GPH_c}	d_{GPHR_c}
3	100	média	0.3747	0.3799
		desvio padrão	0.1953	0.1513
		viés	−0.0753	−0.0701
		EQM	0.0438	0.0278
	800	média	0.4080	0.4309
		desvio padrão	0.0679	0.0576
		viés	−0.0419	−0.0191
		EQM	0.0064	0.0037
5	100	média	0.3108	0.3741
		desvio padrão	0.1934	0.1452
		viés	−0.1392	−0.0759
		EQM	0.0567	0.0268
	800	média	0.3526	0.4270
		desvio padrão	0.0846	0.0568
		viés	−0.0974	−0.0229
		EQM	0.0166	0.0038
10	100	média	0.1923	0.3778
		desvio padrão	0.1727	0.1433
		viés	−0.2577	−0.0722
		EQM	0.0962	0.0258
	800	média	0.2306	0.4349
		desvio padrão	0.0809	0.0576
		viés	−0.2194	−0.0151
		EQM	0.0547	0.0035

Tabela 4.2: Estimativas do parâmetro $d = 0.45$ obtidas no modelo ARFIMA$(0, d, 0)$ com $\omega = 3, 5, 10$ e $\alpha = \beta = 0.7$.

resultado imediato da equação (2.4). Utilizando (2.6), tem-se que o espectro do processo ARFIMA$(1, d, 0)$ é aproximadamente

$$f(\lambda) \approx \frac{\lambda^{-2d}}{(1-\phi)^2}, \quad \text{quando } \lambda \to 0.$$

O parâmetro auto-regressivo positivo contribui para o aumento de memória do processo, isto é, as correlações tornam-se mais significativas do que as do processo ARFIMA$(0, d, 0)$; analogamente, quando $\phi \to 1$, $f(\lambda) \to \infty$.

O espectro do processo ARFIMA$(1, d, 0)$ contaminado pode ser diretamente obtido a partir do Corolário 1, sendo dado por

$$f_z(\lambda) \approx \frac{\lambda^{-2d}}{(1-\phi)^2} + \frac{1}{2\pi}\sum_{i=1}^{m}\omega_i^2 p_i, \quad \text{quando } \lambda \to 0.$$

Note, com base em $f(\lambda)$ e $f_z(\lambda)$ para λ nas proximidades de zero, que os valores de $f_z(\lambda)$ serão maiores que os valores obtidos para $f(\lambda)$, i.e., $f_z(\lambda) \gg f(\lambda)$. (O símbolo "$\gg$" denota "muito maior que".)

Resultados numéricos sobre a estimação do parâmetro d na presença de componentes auto-regressivos e de médias móveis podem ser encontrados em Reisen (1994) e Reisen et al. (2001).

A influência conjunta dos efeitos da inclusão do termo auto-regressivo e da translação na densidade espectral, causada pela presença dos dados atípicos, pode conduzir a estimativas espúrias do parâmetro de integração fracionária. Note, por exemplo, na Tabela 4.3 que quando $\phi = 0.7$, $n = 800$ e $\omega = 10$, a estimativa média fornecida pelo estimador GPH com dados contaminados é 0.3098, significativamente próxima do valor verdadeiro do parâmetro d. Quando $\omega = 20$, todavia, o valor médio das estimativas fornecidas pelo mesmo estimador cai para 0.2366.

As evidências numéricas mostram ainda que, para tamanhos de amostras relativamente grandes, o estimador d_{GPHR} mantém a robustez na estimação do parâmetro d para o modelo ARFIMA$(1, d, 0)$. Seguindo a metodología apresentada na Seção 3.2, a Tabela 4.4 apresenta as estimativas obtidas para o parâmetro auto-regressivo utilizando os diferentes estimadores para a função de autocorrelação (clássico e robusto) nas equações de Yule-Walker. Aqui, $\widehat{\phi}$ e $\widetilde{\phi}$ denotam as estimativas do parâmetro ϕ utilizando $\widehat{\rho}(h)$ e $\widetilde{\rho}(h)$, respec-

tivamente. Analogamente, as estimativas para as séries contaminadas são denotadas por $\widehat{\phi}_c$ e $\widetilde{\phi}_c$, respectivamente.

Os resultados obtidos favorecem a utilização do estimador robusto da FAC no processo de estimação dos parâmetros auto-regressivos em séries temporais na presença de dados atípicos.

ϕ	n		d_{GPH}	d_{GPH_c}	d_{GPHR}	d_{GPHR_c}
0.2	100	média	0.3596	0.1527	0.2735	0.2413
		desvio padrão	0.2829	0.2943	0.2633	0.2479
		viés	0.0596	−0.1473	−0.0265	−0.0587
		EQM	0.0836	0.1083	0.0700	0.0649
	300	média	0.3090	0.1739	0.2532	0.2536
		desvio padrão	0.2150	0.2288	0.1901	0.1874
		viés	0.0090	−0.1261	−0.0468	−0.0464
		EQM	0.0463	0.0683	0.0383	0.0373
	800	média	0.3065	0.1926	0.2687	0.2610
		desvio padrão	0.1451	0.1471	0.1292	0.1323
		viés	0.0065	−0.1074	−0.0313	−0.0390
		EQM	0.0211	0.0332	0.0177	0.0190
0.5	100	média	0.4539	0.2642	0.3533	0.3440
		desvio padrão	0.2927	0.2862	0.2341	0.2528
		viés	0.1539	−0.0358	0.0533	0.0439
		EQM	0.1094	0.0832	0.0577	0.0659
	300	média	0.3409	0.2574	0.2872	0.2839
		desvio padrão	0.2159	0.2070	0.1793	0.1756
		viés	0.0409	−0.0426	−0.0128	−0.0161
		EQM	0.0483	0.0447	0.0323	0.0311
	800	média	0.3179	0.2536	0.2808	0.2701
		desvio padrão	0.1459	0.1562	0.1306	0.1254
		viés	0.0179	−0.0464	−0.0192	−0.0299
		EQM	0.0216	0.0265	0.0174	0.0166
0.7	100	média	0.5848	0.4522	0.4593	0.4436
		desvio padrão	0.3356	0.3599	0.2464	0.2475
		viés	0.2848	0.1522	0.1593	0.1436
		EQM	0.1938	0.1527	0.0861	0.0819
	300	média	0.4299	0.3694	0.3841	0.3731
		desvio padrão	0.2046	0.2021	0.1957	0.1853
		viés	0.1299	0.0694	0.0841	0.0731
		EQM	0.0587	0.0457	0.0454	0.0397
	800	média	0.3410	0.3098	0.3150	0.3062
		desvio padrão	0.1443	0.1423	0.1407	0.1357
		viés	0.0409	0.0098	0.0149	0.0062
		EQM	0.0225	0.0203	0.0200	0.0185

Tabela 4.3: Estimativas do parâmetro d obtidas por $\hat{d}_{GPH}, \hat{d}_{GPHc}$, $\hat{d}_{GPHR}$ e $\hat{d}_{GPHRc}$ para o modelo ARFIMA$(1, 0.3, 0)$ com $\phi = 0.2, 0.5, 0.7$, $\omega = 0, 10$, $\alpha = 0.5$ e $\beta = 0.7$.

ϕ	$\widehat{d}$	n		$\widehat{\phi}$	$\widehat{\phi}_c$	$\widetilde{\phi}$	$\widetilde{\phi}_c$
0.2	0.2413	100	média	0.1653	0.0437	0.1663	0.1520
			desvio padrão	0.1188	0.1116	0.1368	0.1308
			viés	−0.0347	−0.1563	−0.0338	−0.0480
			EQM	0.0153	0.0369	0.0199	0.0194
	0.2536	300	média	0.1948	0.0324	0.1918	0.1820
			desvio padrão	0.0643	0.0509	0.0709	0.0772
			viés	−0.0052	−0.1676	−0.0086	−0.0179
			EQM	0.0042	0.0307	0.0051	0.0063
	0.2610	800	média	0.1959	0.0414	0.1951	0.1873
			desvio padrão	0.0351	0.0371	0.0394	0.0397
			viés	−0.0041	−0.1586	−0.0049	−0.0127
			EQM	0.0012	0.0265	0.0016	0.0017
0.7	0.4510	100	média	0.6406	0.2559	0.6445	0.6158
			desvio padrão	0.0959	0.2085	0.0997	0.1057
			viés	−0.0594	−0.4441	−0.0555	−0.0842
			EQM	0.0127	0.2407	0.0130	0.0183
	0.3534	300	média	0.6838	0.1981	0.6829	0.6579
			desvio padrão	0.0443	0.0889	0.0502	0.0556
			viés	−0.0162	−0.5019	−0.0171	−0.0421
			EQM	0.0022	0.2598	0.0028	0.0049
	0.3049	800	média	0.6955	0.2092	0.6943	0.6719
			desvio padrão	0.0245	0.0613	0.0265	0.0271
			viés	−0.0045	−0.4908	−0.0057	−0.0281
			EQM	0.0006	0.2446	0.0007	0.0015

Tabela 4.4: Estimativas do parâmetro ϕ para o modelo ARFIMA$(1, d, 0)$ com $\omega = 0,10$, $\alpha = 0.5$ e $\beta = 0.7$ para as séries diferenciadas com as estimativas obtidas pelos estimadores d_{GPH} e d_{GPHR}.

CAPÍTULO 5

Aplicações

Neste capítulo apresentamos duas aplicações da metodologia descrita na Seção 3.2. A primeira aplicação utiliza a série temporal do nível mínimo anual da beira do rio Nilo no periodo que se estende de 622 a 1284 D.C. Na literatura, tem sido sugerido que os modelos ARFIMA$(0, d, 0)$ são adequados para analisar esse conjunto de dados (ver Figura 5.1). Para efeitos de comparação com a estimativa obtida através do estimador apresentado no Capítulo 3, a Tabela 5.1 mostra as estimativas do parâmetro de diferênça fracionária através dos estimadores propostos por Agostinelli & Bisaglia (2003), Beran (1994) e Robinson (1994).

	$\widehat{d}$
Robinson (1994)	0.4338
Beran (1994)	0.4000
Agostinelli & Bisaglia (2003)	0.4160
GPH Robusto (d_{GPHR})	0.4161

Tabela 5.1: Valores das estimativas do parâmetro d para os dados do rio Nilo.

Observa-se que o valor obtido por d_{GPHR} está consideravelmente próximo do valor

fornecido pelo estimador de Agostinelli & Bisaglia (2003), indicando, assim, que não é necessário utilizar toda a informação do espectro, como no caso dos estimadores paramétricos, para obter boas estimativas do parâmetro d.

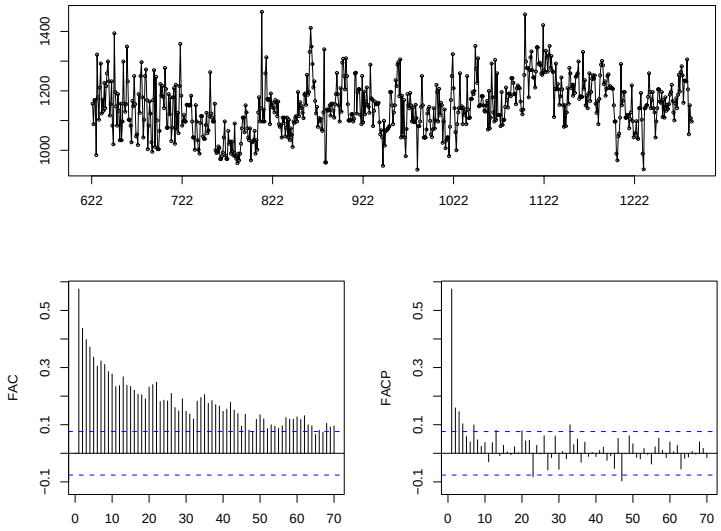


Figura 5.1: Dados do nível do rio Nilo e funções amostrais de autocorrelação.

A segunda aplicação utiliza a taxa mensal de variação do Índice Nacional de Preços ao Consumidor Amplo (IPCA) do Brasil. O IPCA é a medida de inflação utilizada pelo Banco Central do Brasil (BCB) para orientar sua política monetária, ou seja, é a medida de inflação que o BCB observa quando estabelece a taxa Selic. O índice é calculado a partir de uma cesta de consumo média de famílias que recebem entre 1 e 40 salários mínimos, a responsabilidade pela sua construção sendo do Instituto Brasileiro de Geografia e Estatística (IBGE). O período aqui considerado se estende de janeiro de 1995 a outubro de 2006. Como pode ser observado na Figura 5.2, o valor para o mês de novembro do ano 2002 parece ser um dado atípico de tipo aditivo. A função de autocorrelação amostral sugere que o processo gerador dos dados é ARIMA fracionário.

Seguindo a metodologia apresentada na Seção 3.2, as estimativas do parâmetro de diferença fracionária obtidas através dos estimadores d_{GPH} e d_{GPHR} são 0.181 e 0.315, respectivamente. Utilizando os filtros de diferença fracionária e as estimativas obtidas para o parâmetro d, as séries diferenciadas são dadas por: $\widehat{U}_{GPH,t} = (1-B)^{d_{GPH}} IPCA_t$ e

$\widehat{U}_{GPHR,t} = (1-B)^{d_{GPHR}} IPCA_t$. Posteriormente à filtragem, foram identificadas as ordens das componentes auto-regressivas (p) e de médias móveis (q) usando o critério de informação de Akaike corrigido (AICC); ver, e.g., Brockwell & Davis (2006). Os valores do critério AICC são apresentados na Tabela 5.2. Observa-se que o menor AICC obtido para as duas séries diferenciadas foi associado ao modelo ARMA$(1,0)$.

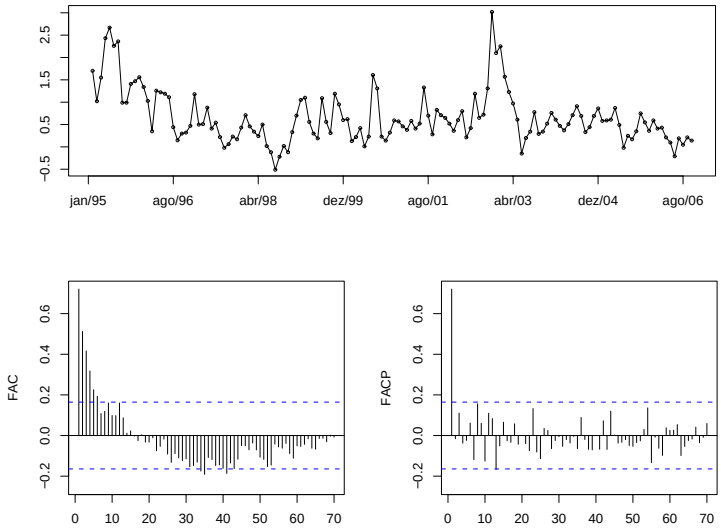


Figura 5.2: Dados IPCA e funções amostrais de autocorrelação.

Utilizando as equações de Yule-Walker, os modelos ajustados são:

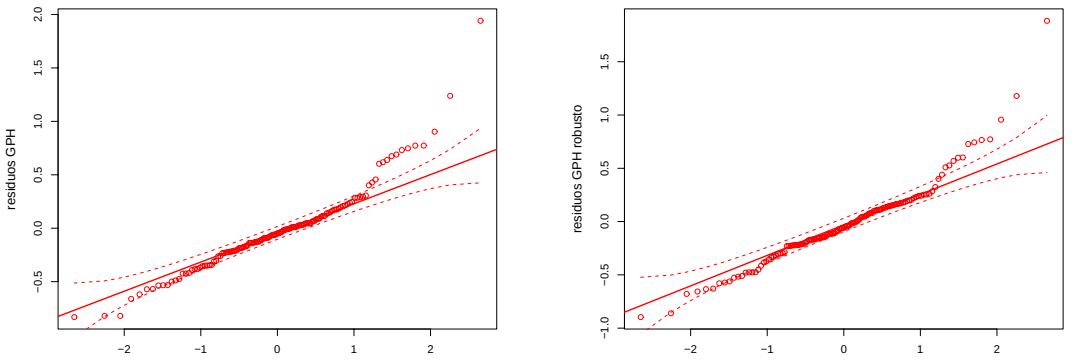
$$\widehat{U}_{GPH,t} = 0.4718\widehat{U}_{GPH,t-1},$$
$$\widehat{U}_{GPHR,t} = 0.2999\widehat{U}_{GPHR,t-1},$$

onde os erros-padrão dos parâmetros estimados são 0.0789 e 0.0728, respectivamente.

Foram realizados testes de adequação do ajuste para verificar o comportamento de ruído branco dos resíduos nos modelos ajustados. Os testes Ljung-Box (GPH: p-valor= 0.1634, GPHR: p-valor= 0.1369) e McLeod - Li (GPH: p-valor= 0.9985 e GPHR: p-valor= 0.9980) confirmam a ausência de autocorrelação serial nos dados. O teste de Jarque-Bera rejeita, em ambos os modelos, a hipótese de normalidade dos resíduos aos níveis usuais de significância, o que pode ser confirmado através dos gráficos quantil-quantil apresentados na Figura 5.3.

GPH		*GPHR*	
Modelo	**AICC**	**Modelo**	**AICC**
ARMA$(1,0)$	129.99	ARMA$(1,0)$	128.46
ARMA$(1,1)$	130.72	ARMA$(2,0)$	128.67
ARMA$(2,0)$	130.87	ARMA$(1,1)$	129.14
ARMA$(3,0)$	131.07	ARMA$(3,0)$	129.64
ARMA$(2,1)$	132.88	ARMA$(2,1)$	131.33
ARMA$(3,1)$	134.29	ARMA$(3,1)$	132.93
ARMA$(1,3)$	137.01	ARMA$(1,2)$	133.09
ARMA$(1,2)$	137.41	ARMA$(1,3)$	135.01
ARMA$(2,2)$	138.89	ARMA$(2,3)$	137.68
ARMA$(2,3)$	139.93	ARMA$(2,2)$	140.77
ARMA$(3,2)$	146.76	ARMA$(3,2)$	144.16
ARMA$(3,3)$	152.17	ARMA$(3,3)$	152.03

Tabela 5.2: Valores do critério AICC obtidas na determinação das ordens p e q a partir das séries $\widehat{U}_{GPH,t}$ e $\widehat{U}_{GPHR,t}$.



(a) Resíduos modelo ARMA$(1,0)$ para $\widehat{U}_{GPH,t}$ (b) Resíduos modelo ARMA$(1,0)$ para $\widehat{U}_{GPHR,t}$

Figura 5.3: Gráficos quantil-quantil dos resíduos dos modelos ARMA$(1,0)$ com bandas de confiança de 95%.

Observa-se que os resíduos obtidos para o modelo ARMA$(1,0)$ através do estimador d_{GPHR} apresentam comportamento menos irregular do que os resíduos obtidos através do estimador d_{GPH}, fazendo com que uma maior quantidade de resíduos esteja entre as bandas de confiança indicativas de normalidade. Cabe ressaltar que a análise de resíduos foi realizada para cada um dos modelos apresentados na Tabela 5.2, mas nenhum deles

conseguiu satisfazer a suposição de normalidade.

Nós comparamos os desempenhos preditivos dos modelos ARFIMA$(1, d, 0)$ com $d = d_{GPH} = 0.181$ e $d = d_{GPHR} = 0.315$. As medidas estatísticas consideradas para a avaliação das previsões são: erro quadrático médio (EQM) e erro total (ET). O erro total é a diferença entre a soma das observações e a soma das previsões. Cada medida é calculada para previsões $1, 6$ e 12 passos à frente e os resultados encontram-se apresentados na Tabela 5.3.

Passos	**Critério**	**GPH**	**GPHR**
1	**EQM**	0.0067	0.0007
	ET	−0.0816	0.0258
6	**EQM**	0.1340	0.0797
	ET	−1.8741	−1.1877
12	**EQM**	0.0863	0.0529
	ET	−2.8353	−1.8421

Tabela 5.3: Medidas de precisão das previsões para 1, 6 e 12 passos à frente.

Observa-se, com o critério do EQM, que as previsões obtidas com o estimador $GPHR$ apresentam maior precisão tanto para horizontes de previsão curtos quanto para horizontes longos. A mesma conclusão é obtida quando o critério da avaliação é o erro total.

Dos resultados anteriores, pode-se concluir que o estimador d_{GPHR} constitui uma boa opção para estimar o parâmetro d na presença de *outliers* de tipo aditivo; adicionalmente, a metodologia sugerida na Seção 3.2 é útil no que tange à estimação dos parâmetros do modelo ARFIMA em presença de dados atípicos.

CAPÍTULO 6

Conclusões

A presente dissertação apresenta uma metodologia robusta para a estimação dos parâmetros do modelo ARFIMA(p, d, q) em séries temporais na presença de *outliers* do tipo aditivo. A metodologia é baseada num procedimento semi-paramétrico realizado em dois passos. No primeiro passo, estima-se o parâmetro de integração fracionária através da regressão do logaritmo do *pseudo-periodograma truncado* representado pela equação (3.5). Posteriormente, estimam-se os parâmetros auto-regressivos e de médias móveis a partir do estimador robusto da função de autocovariância, sugerido por Ma & Genton (2000).

Os resultados teóricos e as evidências empíricas mostram a sensibilidade do estimador GPH à presença de dados atípicos, indicando que este não é um estimador apropriado para a estimação do parâmetro d na presença desse tipo de dados. Nosso estimador robusto apresenta propriedades estatísticas desejáveis no que concerne à estimação do parâmetro d em séries temporais com propriedade de memória longa contaminadas por *outliers* do tipo aditivo. As evidências numéricas sugerem que o método robusto proposto é um procedimento alternativo atraente para a estimação dos parâmetros de um modelo ARFIMA(p, d, q) com propriedade de memória longa.

Referências Bibliográficas

Agostinelli, C. & Bisaglia, L. (2003), Robust estimation of ARFIMA processes, Technical report, Università Cà Foscari di Venezia.

Baillie, R. (1996), 'Long memory process and fractional integration in econometrics', *Journal of Econometrics* **73**, 5–59.

Baillie, R. T. & Chung, S. (2002), 'Modeling and forecasting from trend-stationary long memory models with applications to climatology', *International Journal of Forecasting* **18**, 215–226.

Barkoulas, J. T. & Baum, C. F. (1998), 'Fractional dynamics in Japanese financial time series', *Pacific-Basin Finance Journal* **6**, 115–124.

Beran, J. (1994), 'On a class of M-estimators for Gaussian long-memory models', *Biometrika* **81**, 755–766.

Beran, J. (1995), 'Maximum likelihood estimation of the differencing parameter for invertible short and long memory autoregressive integrated moving average models', *Journal of the Royal Statistical Society* **57**(B), 659–672.

Bhardwaj, G. & Swanson, N. (2006), 'An empirical investigation of the usefulness of

ARFIMA models for predicting macroeconomic and financial time series', *Journal of Econometrics* **131**, 539–578.

Bisaglia, L. & Guégan, D. (1998), 'A comparison of techniques of estimation in long-memory process', *Computational Statistics & Data Analysis* **27**, 61–81.

Box, G., Jenkins, G. & Reinsel, G. (1994), *Time Series Analysis: Forecasting and Control*, third edn, Prentice Hall.

Box, G. & Tiao, G. C. (1975), 'Intervention analysis with applications to economic and enviromental problems', *Journal of the American Statistical Association* **70**, 70–79.

Brockwell, P. & Davis, R. (2002), *Introduction to Time Series and Forecasting*, second edn, Springer Verlag.

Brockwell, P. & Davis, R. (2006), *Time Series: Theory and Methods*, second edn, Springer Verlag.

Chan, W. (1992), 'A note on time series model specification in the presence outliers', *Journal of Applied Statistics* **19**, 117–124.

Chan, W. (1995), 'Outliers and financial time series modelling: a cautionary note', *Mathematics and Computers in Simulation* **39**, 425–430.

Chang, I., Tiao, G. C. & Chen, C. (1988), 'Estimation of time series parameters in presence of outliers', *Technometrics* **30**, 1936–204.

Chen, C. & Liu, L. (1993*a*), 'Forecasting time series with outliers', *Journal of Forecasting* **12**, 13–35.

Chen, C. & Liu, L. (1993*b*), 'Joint estimation of model parameters and outlier effects in time series', *Journal of the American Statistical Association* **88**, 284–297.

Croux, C. & Rousseeuw, P. J. (1992), 'Time-efficient algorithms for two highly robust estimators of scale', *Computational Statistics* **1**, 1–18.

Cunado, J., Gil-Alana, L. A. & Péres de Gracia, F. (2004), 'Real convergence in Taiwan: a fractionally integrated approach', *International Review of Financial Analysis* **13**, 265–276.

Dahlhaus, R. (1989), 'Efficient parameter estimation for self-similar processes', *The Annals of Statistics* **17**, 1749–1766.

Deutsch, S. J., Richards, J. E. & Swain, J. (1990), 'Effects of a single outlier on ARMA identification', *Communications in Statistics: Theory and Methods* **19**, 2207–2227.

Doukhan, P., Oppenheim, G. & Taqqu, M. (2003), *Theory and Applications of Long-Range Dependence*, Birkhäuser.

Fox, A. J. (1972), 'Outliers in time series', *Journal of the Royal Statistical Society* **34**(B), 350–363.

Fox, R. & Taqqu, M. S. (1986), 'Large-sample properties of parameters estimates for strongly dependent stationary gaussian time series', *The Annals of Statistics* **14**, 517–532.

Franses, P. H. & Ooms, M. (1997), 'A periodic long-memory model for quarterly UK inflation', *International Journal of Forecasting* **13**, 117–126.

Geweke, J. & Porter-Hudak, S. (1983), 'The estimation and application of long memory time series model', *Journal of Time Series Analysis* **4**, 221–238.

Gil-Alana, L. A. (2004), 'Long memory in the U.S. interest rate', *International Review of Financial Analysis* **13**, 265–276.

Granger, C. W. J. & Joyeux, R. (1980), 'An introduction to long-memory time series models and fractional differencing', *Journal of Time Series Analysis* **1**, 15–30.

Haldrup, N. & Nielsen, M. O. (2007), 'Estimation of fractional integration in the presence of data noise', *Computational Statistics & Data Analysis* **51**, 3100–3114.

Hauser, M. (1999), 'Maximum likelihood estimators for ARMA and ARFIMA models: a Monte Carlo study', *Journal of Statistical Planning and Inference* **80**, 229–255.

Hosking, J. R. (1981), 'Fractional differencing', *Biometrika* **68**, 165–176.

Huber, P. J. (2004), *Robust Statistics*, third edn, John Wiley & Sons.

Hurvich, C. M., Deo, R. & Brodsky, J. (1998), 'The mean square error of Geweke and Porter-Hudak's estimator of the memory parameter of a long-memory time series', *Journal of Time Series Analysis* **19**, 19–46.

Ledolter, J. (1989), 'The effect of additive outliers on the forecast from ARMA models', *International Journal of Forecasting* **5**, 231–240.

Lobato, I. & Robinson, P. M. (1996), 'Averaged periodogram estimation of long memory', *Journal of Econometrics* **73**, 303–324.

Ma, Y. & Genton, M. (2000), 'Highly robust estimation of the autocovariance function', *Journal of Time Series Analysis* **21**, 663–684.

Priestley, M. B. (1983), *Spectral Analysis and Time Series*, Academic Press.

Reisen, V. (1994), 'Estimation of the fractional difference parameter in the ARIMA(p, d, q) model using the smoothed periodogram', *Journal of Time Series Analysis* **15**, 335–350.

Reisen, V. A., Rodrigues, A. & Palma, W. (2006), 'Estimation os seasonal fractionally integrated processes', *Computational Statistics & Data Analysis* **50**, 568–582.

Reisen, V., Abraham, B. & Lopes, S. (2001), 'Estimation of parameters in ARFIMA processes: A simulation study', *Communications in Statistics: Simulation and Computation* **30**, 787–803.

Reisen, V., Abraham, B. & Toscano, E. (2000), 'Parametric and Semiparametric Estimations of Stationary Univariate ARFIMA Models', *Brazilian Journal of Probability and Statistics* **14**, 185–206.

Reisen, V., Abraham, B. & Toscano, E. (2002), 'Effect of parameter estimation on estimating the forecast error variace in an ARFIMA processes: a simulation study and an example', *Statistical Methods* **4**, 21–37.

Reisen, V., Cribari-Neto, F. & Jensen, M. (2003), 'Long memory inflationary dynamics: The case of brazil', *Studies in Nonlinear Dynamics & Econometrics* **7**, 1–16.

Robinson, P. M. (1994), 'Semiparametric analysis of long-memory time series', *The Annals of Statistics* **22**, 515–539.

Robinson, P. M. (1995*a*), 'Gaussian semiparametric estimation of long range dependence', *Annals of Statistics* **23**, 1630–1661.

Robinson, P. M. (1995*b*), 'Log-periodogram regression of time series with long range dependence', *Annals of Statistics* **23**, 1048–1072.

Rousseeuw, P. J. & Croux, C. (1993), 'Alternatives to the median absolute deviation', *Journal of the American Statistical Association* **88**, 1273–1283.

Smith, J., Taylor, N. & Yadav, S. (1997), 'Comparing the bias and misspecification in ARFIMA models', *Journal of Time Series Analysis* **18**, 507–527.

Sowell, F. B. (1992), 'Maximum likelihood estimation of stationary univariate fractionally integrated time series models', *Journal of Econometrics* **53**, 165–188.

Sun, Y. & Phillips, P. (2003), 'Nonlinear log-periodogram regression for perturbed fractional processes', *Journal of Econometrics* **115**, 355–389.

Tsay, R. S. (1986), 'Time series model specification in the presence of outliers', *Journal of the American Statistical Association* **81**, 132–141.

Velasco, C. (2000), 'Non-Gaussian log-periodogram regression', *Econometric Theory* **16**, 44–79.

Wei, W. (2005), *Time Series Analysis: Univariate and Multivariate Methods*, Addison Wesley.