Intervalos de confiança baseados em Deviance para os hiperparâmetros em modelos estruturais

Thiago Barbosa Ceccotti

Universidade Federeal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística

Intervalos de confiança baseados em Deviance para os hiperparâmetros em modelos estruturais

Thiago Barbosa Ceccotti

Dissertação apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de Minas Gerais para a obtenção do título de Mestre em Estatística.

Resumo

Este trabalho se propõe a comparar diferentes procedimentos para a obtenção de intervalos de confiança para os hiperparâmetros em modelos estruturais. As metodologias geralmente empregadas incluem métodos baseados na distribuição assintótica dos estimadores de máxima verossimilhança, assim como intervalos utilizando a técnica bootstrap. Contudo, o primeiro método apresenta problemas de fronteira para parâmetros de variância, além de não ser eficaz com dados não gaussianos e o segundo tem um alto custo computacional. Este trabalho apresenta três métodos para a construção de intervalos de confiança baseados em verossimilhança. O primeiro é uma aproximação de uma região de confiança a partir do teste da razão de verossimilhança, o segundo é um intervalo de confiança marginal e o terceiro é baseado na função *signed root deviance profile*. Estes métodos visam contornar problemas do método assintótico no caso de pequenas amostras e problemas de fronteira do intervalo, além de serem alternativas computacionalmente menos custosa que o método bootstrap. É feita uma comparação, via simulação Monte Carlo, buscando estabelecer as vantagens e desvantagens de cada método. De maneira geral, pode-se concluir que o método assintótico não é recomendado para casos não-gaussianos e que o intervalo *signed root deviance profile* é o método com melhores coberturas e apresenta um tempo computacional expressivamente menor que o intervalo bootstrap, também utilizado na literatura para a construção de intervalos de confiança. Esta dissertação introduz na literatura um novo tipo de intervalo de confiança para os hiperparâmetros dos modelos estruturais, onde além de se evitar os problemas de fronteira do método assintótico ganha-se em tempo computacional frente ao método alternativo bootstrap, sem perder a assimetria presente neste. Estes novos intervalos apresentam também melhores coberturas para os hiperparâmentros e funcionam muito bem para séries temporais pequenas.

Abstract

The objective of this work is to compare different confidence interval procedures for hyperparameters of structural models. The usual procedures include the asymptotic method, based on the asymptotic distribuition of the maximum likelihood estimator, as well as a bootstrap based confidence interval. This work presents three methods based on the likelihood test. The first is a marginal aproximation of confidence regions, the second is based on the profile deviance function and the third method is the *Signed root Deviance Profile*. Those methods avoid the problems associated with the asymptotic method for small samples, as intervals generated outside the parametric space. They are also an alternative for bootstrap methods, being computationaly more efficient. A comparison is performed, via Monte Carlo simulation, in order to establish advantages and disadvantages for each method. The results show that these methods possess a better coverage rate than the asymptotic and bootstrap procedures.

Agradecimentos

Expresso aqui minha gratidão a todos os professores que me ajudaram e motivaram neste caminho, principalmente à Prof[a]. Glaura Franco e ao Prof. Thiago Rezende meus orientadores de mestrado.
Agradeço aos colegas de mestrado e aos funcionários da secretaria, sempre dispostos a ajudar e compartilhar experiências.
Às intituições de fomento à pesquisa que fazem possível uma melhora científica no Brasil.
Agradeço também a paciência, compreensão e ajuda de amigos e familiares.

"What! you have solved it already?"
"Well, that would be too much to say.
I have discovered a suggestive fact, that is all.
It is, however, very suggestive.
The details are still to be added."
Sherlock Holmes, Sign of Four

Lista de Figuras

Lista de Tabelas

Abreviaturas

EMV Estimador de Máxima Verossimilhança

ME Modelos Estruturais

MNL Modelo de Nível Local

MTL Modelo de Tendência Linear Local

MEB Modelo Estrutural Básico

FEE Forma de Espaço de Estados

FK Filtro de Kalman

BFGS Algoritmo Broyden-Fletcher-Goldfarb-Shanno

ICA Intervalo de Confiança Assintótico

ICB Intervalo de Confiança Bootstrap Percentílico

IC Intervalo de Confiança

RC Região de Confiança

ICM Intervalo de Confiança Marginal

ICD Intervalo de Confiança baseado na estatística Deviance

ICS Intervalo de Confiança *Signed Root Deviance Profile*

MC Monte Carlo

Sumário

Capítulo 1

Introdução

Uma série temporal é um conjunto de dados coletados e ordenados no tempo. Séries temporais aparecem constantemente no cotidiano, por exemplo em dados meteorológicos, taxas de desemprego e preços de ações, assim como nas áreas de processamentos de sinais, finanças e para monitoramento epidemiológico, entre outras.

Existem várias metodologias para a modelagem deste tipo de dados. Entre as mais conhecidas, destacam-se o alisamento exponencial, desenvolvido por Holt(1957) e Winters(1960), os modelos de Box & Jenkins (1976) e os modelos estruturais (Harvey,1989). Por ser um modelo que decompõe a série em componentes não-observáveis como sazonalidade, nível e tendência, os modelos estruturais têm uma interpretação mais intuitiva. Uma comparação da previsão de valores futuros usando os modelos estruturais e os de Box & Jenkins mostra bons argumentos a favor do primeiro, segundo trabalho de Harvey & Todd (1983), um fato importante, visto que a modelagem de séries temporais pretende, em geral, prever valores futuros. O mesmo trabalho mostra que a metodologia de Box & Jenkins pode apresentar problemas em séries pequenas gerando um modelo inapropriado.

Assim, neste trabalho serão utilizados os modelos estruturais (Harvey, 1989). Estes modelos utilizam a forma de espaço de estados e o filtro de Kalman (Kalman, 1960) para a construção da função de verossimilhança. Sob a abordagem Bayesiana estes modelos são conhecidos como modelos dinâmicos (West & Harrison, 1997). As variâncias das distribuições destas componentes não-observáveis, de nível, tendência, sazonalidade e erro, são ditas hiperparâmetros. Este estudo foca na construção de intervalos de confiança para estes hiperparâmetros.

A forma usual para calcular intervalos de confiança para os hiperparâmetros é baseada na distribuição assintótica do estimador de máxima verossimilhança (EMV). Contudo, para amostras pequenas, o EMV pode não satisfazer as propriedades assintóticas (Pfanzagl, 1994). Este problema já foi discutido para o caso dos modelos estruturais em Quenneville & Singh (2000) e em Pfeffermann & Tiller (2005). Sendo assim, é necessário buscar outras maneiras para calcular intervalos além do método assintótico.

Uma outra opção para construção de intervalos de confiança é utilizar o método de reamostragem bootstrap (Efron, 1979). Uma adaptação deste aos modelos estruturais, e outros

modelos que possam ser escritos na forma de espaço de estados, foi feita por Stoffer & Wall (1991). O trabalho destes autores possibilitou a construção de intervalos de confiança bootstrap para os parâmetros de estado (Pfeffermann & Tiller, 2005; Rodriguez & Ruiz, 2012), para as observações futuras (Rodriguez & Ruiz, 2009) e para os hiperparâmetros (Franco et. al, 2008).

Uma outra possibilidade para obtenção de intervalos de confiança é baseada na distribuição assintótica do teste da razão de verossimilhança (Wilks, 1938), que é mais robusto para amostras pequenas. Este método é conhecido como intervalo de confiança baseado na função deviance. Espera-se que estes intervalos sejam computacionalmente mais eficientes que os construídos pelo método bootstrap, que necessita um processo de maximização da verossimilhança para cada reamostra. Esta eficiência é de importância para análises de séries econômicas, por exemplo, onde decisões devem ser tomadas em tempo hábil. Além disso, este método é atrativo, pois permite que os intervalos de confiança sejam assimétricos e não tenham problemas de fronteira como os intervalos de confiança assintóticos.

Este método baseado na verossimilhança já foi utilizado na literatura em análise de dados genéticos (Neale & Miller, 1997), para resolver o problema de separação em regressões logísticas (Heinze & Schemper, 2002), para a construção de intervalos na presença de parâmetros de perturbação (Rolke et al., 2005) e comparado com o método bootstrap para estimadores de população do tipo captura-recaptura (Evans et al, 1996; Gimenez et al., 2005), entre outros. Contudo, estes intervalos ainda não foram utilizados em modelos estruturais.

Além do intervalo baseado na função deviance, serão avaliados também os intervalos construídos via *signed root deviance profile* (Chen & Jennrich, 1996). Uma vantagem deste método sobre o de verossimilhança usual é que os intervalos retornados por ele são de fácil interpretação e não necessitam de um processo de maximização em dois passos. Este também é um método ainda não aplicado aos modelos estruturais.

Desta forma, este trabalho pretende aplicar e comparar os diversos métodos citados acima quanto à cobertura e amplitude dos intervalos de confiança para os hiperparâmetros em modelos estruturais, via simulações Monte Carlo. Pretende-se também estimar o tempo computacional gasto por cada um deles, facilitando a escolha destes métodos em casos reais onde tempo de processamento é uma variável importante.

Esta dissertação tratará no Capítulo 2 sobre os Modelos Estruturais: a forma de espaço de estados e o filtro de Kalman, construção de intervalos de confiança assintóticos e os intervalos bootstrap. No Capítulo 3 é feita uma descrição sobre os intervalos baseados em verossimilhança e o *signed root deviance profile*. Seguindo com uma explicação dos métodos computacionais no Capítulo 4 e simulações no Capítulo 5, aplicações a séries de dados reais no Capítulo 6 e concluindo no capítulo final com uma comparação entre os métodos.

Capítulo 2

Modelos Estruturais

A decomposição de uma série temporal y_t via Modelos Estruturais (ME), em termos de suas componentes não observáveis é dada por:

$$y_t = \mu_t + \beta_t + \gamma_t + \epsilon_t; \quad \epsilon_t \sim N(0, \sigma_\epsilon^2),$$

para $t = 1, 2, ..., n$, onde μ_t representa o nível, β_t sua tendência e γ_t a componente sazonal de y_t.

Os modelos mais simples são o Modelo de Nível Local (MNL), o Modelo de Tendência Linear local (MTL) e o Modelo Estrutural Básico (MEB), já revisados em detalhe por Harvey (1990).

Uma série $y_t, t = 1, 2, ..., n$ que, aparentemente, oscile em torno de um nível constante pode ser simplesmente modelada por um MNL (ver Figura 2.1)

$$\begin{aligned} y_t &= \mu_t + \epsilon_t; & \epsilon_t &\sim N(0, \sigma_\epsilon^2) \\ \mu_t &= \mu_{t-1} + \eta_t; & \eta_t &\sim N(0, \sigma_\eta^2) \end{aligned}$$

onde ϵ_t e η_t são não-correlacionados para $t = 1, 2, ..., n$.

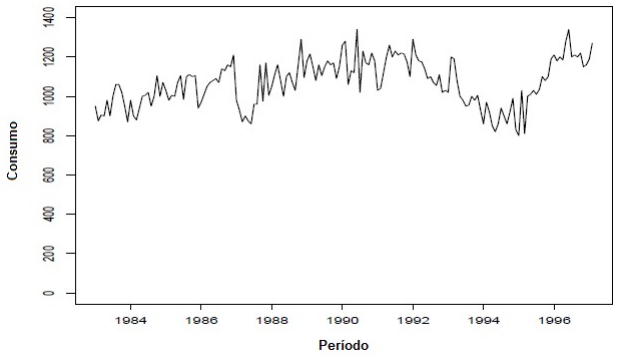

Figura 2.1: Exemplo MNL. Série do consumo de eletricidade na região nordeste do Brasil.

Caso este nível varie linearmente com o tempo, uma modelagem via o MTL é indicada (ver Figura 2.2)

$$\begin{aligned} y_t &= \mu_t + \epsilon_t; & \epsilon_t &\sim N(0, \sigma_\epsilon^2) \\ \mu_t &= \mu_{t-1} + \beta_{t-1} + \eta_t; & \eta_t &\sim N(0, \sigma_\eta^2) \\ \beta_t &= \beta_{t-1} + \xi_t; & \xi_t &\sim N(0, \sigma_\xi^2) \end{aligned}$$

onde ξ_t ,ϵ_t e η_t são mutuamente não-correlacionados para $t = 1, 2, ..., n$.

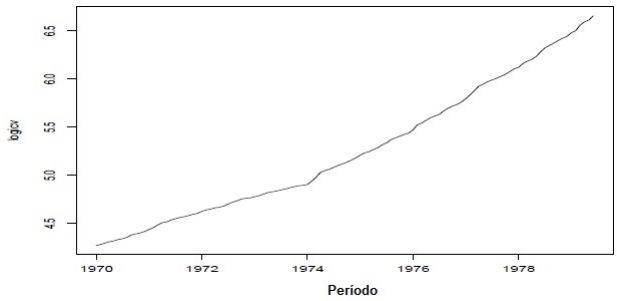

Figura 2.2: Exemplo MTL. Série do logaritmo natural do índice do custo de vida na cidade de São Paulo.

O MEB é o MTL acrescido de uma componente sazonal (ver Figura 2.3)

$$\begin{aligned} y_t &= \mu_t + \gamma_t + \epsilon_t; & \epsilon_t &\sim N(0, \sigma_\epsilon^2) \\ \mu_t &= \mu_{t-1} + \beta_{t-1} + \eta_t; & \eta_t &\sim N(0, \sigma_\eta^2) \\ \beta_t &= \beta_{t-1} + \xi_t; & \xi_t &\sim N(0, \sigma_\xi^2) \\ \gamma_t &= -\gamma_{t-1} - ... - \gamma_{t-(s-1)} + \omega_t; & \omega_t &\sim N(0, \sigma_\omega^2) \end{aligned}$$

onde s indica o número de períodos sazonais, e ω_t, ξ_t ,ϵ_t e η_t são mutuamente não-correlacionados para $t = 1, 2, ..., n$.

Generalizando, pode-se escrever estes modelos básicos na forma de espaço de estados (FEE), o que reduz o número de equações e abrange uma classe maior de modelos, como a regressão linear, modelos ARMA e até mesmo modelagem de séries multivariadas. As equações da FEE são:

$$\begin{aligned} y_t &= z_t^T \alpha_t + \epsilon_t; & \epsilon_t &\sim N(0, h_t) \\ \alpha_t &= T_t \alpha_{t-1} + R_t \eta_t; & \eta_t &\sim N(0, Q_t), \end{aligned} \tag{1}$$

onde $t = 1, 2, ..., n$, η_t é um vetor de ruídos não correlacionados, com matriz de covariância Q_t diagonal e independente de ϵ_t; T_t e R_t são matrizes que indentificam o modelo e z_t um vetor que identifica o modelo.

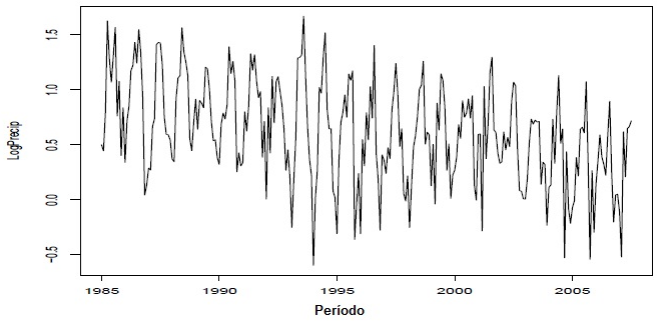

Figura 2.3: Exemplo MEB. Série do logaritmo natural da precipitação de SO_4 em Nova Iorque.

Note que α_t é o vetor que indica os componentes não observáveis do modelo, que podem ser estocásticos ou determinísticos. A equação para y_t é conhecida como equação das observações e, para α_t, equação de estado. Além disso o vetor de estados inicial α_0 é tal que $E(\alpha_0) = a_0$, $var(\alpha_0) = P_0$ e $E(\eta_t \alpha_t) = 0, \forall t$. Neste trabalho é suposta homocedasticidade, $h_t = h$ e $Q_t = Q$. Estudos sobre a modelagem e identificação de modelos heteroscedasticos podem ser encontrados em Broto & Ruiz (2009).

Para ilustrar a FEE, serão apresentados os MNL, MTL e MEB nesta forma. A FEE aplicada ao MNL é escrita com as seguintes variáveis:

$$\begin{array}{llll} z_t^T = 1, & \epsilon_t = \epsilon_t, & h_t = \sigma_\epsilon^2, & R_t = 1, \\ T_t = 1, & \eta_t = \eta_t, & Q_t = \sigma_\eta^2, & \alpha_t = \mu_t. \end{array}$$

Já para o MTL, fica claro que T_t e Q_t são matrizes e α_t e z_t vetores:

$$\begin{array}{llll} z_t^T = [1 \quad 0], & \epsilon_t = \epsilon_t, & h_t = \sigma_\epsilon^2, & R_t = 1, \\ T_t = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, & \eta_t = \begin{bmatrix} \eta_t \\ \xi_t \end{bmatrix}, & Q_t = \begin{bmatrix} \sigma_\eta^2 & 0 \\ 0 & \sigma_\xi^2 \end{bmatrix}, & \alpha_t = \begin{bmatrix} \mu_t \\ \beta_t \end{bmatrix}. \end{array}$$

Finalmente, o MEB com s períodos sazonais apresenta as seguintes matrizes:

$$Q_t = \begin{bmatrix} \sigma_\eta^2 & & & & \overrightarrow{0} & \\ & \sigma_\xi^2 & & & & \\ & & \sigma_\omega^2 & & & \\ & & & 0 & & \\ & \overrightarrow{0} & & & \ddots & \\ & & & & & 0 \end{bmatrix}_{(s+1)\text{x}(s+1)}, \qquad \alpha_t = \begin{bmatrix} \mu_t \\ \beta_t \\ \gamma_t \\ \gamma_{t-1} \\ \vdots \\ \gamma_{t-s+2} \end{bmatrix}_{(s+1)\text{x}1}, \quad h_t = \sigma_\epsilon^2,$$

$$T_t = \begin{bmatrix} 1 & 1 & & & \overrightarrow{0} & & \\ 0 & 1 & & & & & \\ & & -1 & -1 & ... & -1 & -1 \\ & & 1 & 0 & ... & 0 & 0 \\ \overrightarrow{0} & & 0 & 1 & ... & 0 & 0 \\ & & \vdots & \vdots & \ddots & \vdots & \vdots \\ & & 0 & 0 & ... & 1 & 0 \end{bmatrix}_{(s+1)\text{x}(s+1)}, \qquad \eta_t = \begin{bmatrix} \eta_t \\ \xi_t \\ \omega_t \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{(s+1)\text{x}1}, \qquad z_t^T = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{(s+1)\text{x}1}^T,$$

$$R_t = 1, \qquad \epsilon_t = \epsilon_t,$$

onde s é o número de períodos sazonais e $\overrightarrow{0}$ uma matriz de zeros.

O Filtro de Kalman (FK) (Kalman,1960) decompõe a série através de equações recursivas que atualizam sequencialmente o vetor de estado α_t, não observado, baseado na informação $Y_{t-1} = \{y_{t-1}, y_{t-2}, ..., y_0\}$ disponível até o tempo $t-1$.

A esperança e a variância condicionais de α_t e y_t são dadas, respectivamente, por:

$$\begin{aligned} a_{t|t-1} = E(\alpha_t|Y_{t-1}) &= T_t E(\alpha_t|Y_t) = T_t a_{t-1} \\ P_{t|t-1} = var(\alpha_t|Y_{t-1}) = T_t var(\alpha_{t-1}|Y_{t-1}) T_t^T + R_t Q_t R_t^T &= T_t P_{t-1} T_t^T + R_t Q_t R_t^T \\ E(y_t|Y_{t-1}) &= z_t^T a_{t|t-1} \\ F_t = var(y_t|Y_{t-1}) &= z_t^T P_{t|t-1} z_t + h_t. \end{aligned}$$

Para se obter as equações de atualização usa-se propriedades da distribuição normal multivariada. Sejam X e Z variáveis aleatórias tais que $(X, Z) \sim N(\mu, \Sigma)$, onde $\mu = (\mu_1, \mu_2)^T$ e $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$. Então, $(X|Z = c) \sim N(\mu_{1|2}, \Sigma_{1|2})$, onde $\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(c - \mu_2)$ e $\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. Assim, tomando $X = \alpha_t$ e $Z = y_t$ obtém-se as equações de atualização dos estimadores para α_t condicionados na informação disponível, Y_t:

$$\begin{aligned} a_t = E(\alpha_t|Y_t) &= a_{t|t-1} + T_{t+1}^{-1} K_t \upsilon_t \\ P_t = var(\alpha_t|Y_t) &= P_{t|t-1} - P_{t|t-1} z_t F_t^{-1} z_t^T P_{t|t-1}, \end{aligned}$$

onde $\upsilon_t = y_t - z_t^T a_{t|t-1}$ é o erro de previsão a um passo à frente e $K_t = T_{t+1} P_{t|t-1} z_t F_t^{-1}$ é a matriz de ganho de Kalman.

Assim, segundo Harvey (1989) as equações simplificadas para o FK são:

$$\begin{aligned} F_t &= z_t^T P_{t|t-1} z_t + h_t \\ K_t &= T_{t+1} P_{t|t-1} z_t F_t^{-1} \\ \upsilon_t &= y_t - z_t^T a_{t|t-1} \\ a_{t+1|t} &= T_{t+1} a_{t|t-1} + K_t \upsilon_t \\ P_{t+1|t} &= T_{t+1} P_{t|t-1} T_{t+1}^T - K_t F_t K_t^T + R_{t+1} Q_{t+1} R_{t+1}^T. \end{aligned}$$

Com o FK é possível construir a função de verossimilhança, de onde pode-se estimar os hiperparâmetros $\psi = (h, Q_{11}, ..., Q_{(p-1)(p-1)})^T \epsilon \mathbb{R}_+^p$, onde p é o número de parâmetros do modelo. No caso do MEB, por exemplo: $\psi = (h, Q_{11}, Q_{22}, Q_{33}) = (\sigma_\eta^2, \sigma_\epsilon^2, \sigma_\xi^2, \sigma_\omega^2)$.

A função de log-verossimilhança obtida pelo FK é:

$$logL(\psi; Y_t) = -\frac{n}{2} log(2\pi) - \frac{1}{2}\Sigma_{t=1}^n log|F_t| - \frac{1}{2}\Sigma_{t=1}^n \frac{\upsilon_t^2}{F_t},$$

que por ser não-linear em ψ deve ser maximizada via métodos numéricos. Neste trabalho é usado o algoritmo Broyden-Fletcher-Goldfarb-Shanno (BFGS) (Shanno,1970).

2.1 Intervalos de Confiança Assintóticos

Obtidas as estimativas de máxima verossimilhança para ψ, denotadas por $\hat{\psi}$, pode-se construir o intervalo de confiança assintótico (ICA).

Usando o fato de que para grandes amostras $\hat{\psi} \sim N(\psi, I^{-1}(\psi))$, onde $I(\psi)$ é a matriz de informação de Fisher e I_{kk} seus elementos da diagonal principal, o ICA para $\psi_k; k = 1, 2, ..., p$, com nível $100(1-\rho)\%$ é dado por:

$$\Big[\hat{\psi}_k - z_{\rho/2}\sqrt{I_{kk}^{-1}(\hat{\psi})}; \quad \hat{\psi}_k + z_{\rho/2}\sqrt{I_{kk}^{-1}(\hat{\psi}))}\Big],$$

onde $z_{\rho/2}$ é o quantil $\rho/2$ da Normal padrão.

Contudo, o cálculo da matriz de informação de Fisher não é uma tarefa trivial e para modelos complexos, como os discutidos neste trabalho, utiliza-se uma aproximação desta matriz para o cálculo do ICA.

Harvey (1989) mostrou que a matriz de informação de Fisher esperada para os modelos estruturais é dada por:

$$I_{ij}(\psi) = \frac{1}{2}\sum_t \Big[tr\Big(F_t^{-1}\frac{\partial F_t}{\partial \psi_i} F_t^{-1}\frac{\partial F_t}{\partial \psi_j}\Big)\Big] + E\Big[\sum_t \frac{\partial \upsilon_t}{\partial \psi_i}^T F_t^{-1}\frac{\partial \upsilon_t}{\partial \psi_j}\Big];$$

onde $i, j = 1, 2, ..., p$, $t = 1, 2, ..., n$ e $tr(A)$ é o traço da matriz A.

Para o cálculo desta matriz, Harvey(1989) propôs uma forma numérica para o cálculo das derivadas de υ_t e F_t, facilitando assim o cálculo da matriz de informação. Este método é

apresentado em detalhes no Capítulo 4. Uma outra aproximação baseada em um algoritmo recursivo, para o cálculo da matriz de informação de Fisher, pode ser vista em Cavanaugh & Shumway (1996).

Além da dificuldade com o cálculo da matriz hessiana, é esperado que, para amostras pequenas, as propriedades assintóticas do EMV não sejam satisfeitas. Assim o ICA pode apresentar problemas diversos, como intervalos com limites fora do espaço paramétrico o que é conhecido como problema de fronteira. Além disto, por ser um intervalo simétrico, pode não captar possiveis assimetrias da distribuição do EMV.

Para sanar o problema de fronteira costuma-se utilizar transformações nos parâmetros, construindo-se intervalos para este novo parâmetro transformado, por exemplo via método delta, e depois reescalando este intervalo para voltar à escala original. Uma transformação comum para parâmetros de variância, σ^2, é a utilização do logaritmo natural. Uma vantagem desta transformação é reduzir a assimetria da função de verossimilhança na direção de σ^2. Além disso, aplicando a função exponencial para retornar o intervalo para a escala de σ^2obtém-se um intervalo de confiança (IC) dentro do espaço paramétrico.

Assim o ICA para $log(\sqrt{\psi_k}); k = 1, 2, ..., p$, com nível $100(1-\rho)\%$ é dado por:

$$\Big[log(\sqrt{\hat{\psi}_k}) - z_{\rho/2}\sqrt{I^{-1}_{kk}(\hat{\psi})}/(2\hat{\psi}_k); \quad log(\sqrt{\hat{\psi}_k}) + z_{\rho/2}\sqrt{I^{-1}_{kk}(\hat{\psi})}/(2\hat{\psi}_k)\Big],$$

onde $z_{\rho/2}$ é o quantil $\rho/2$ da Normal padrão.

O ICA transformado, para ψ, tem como extremos $(e^{LI})^2$ e $(e^{LS})^2$, onde LI é o limite inferior do ICA para $log(\sqrt{\psi_k})$ e LS é o limite superior deste mesmo intervalo.

Esta abordagem apresenta, contudo, um problema quando o valor da estimativa $\hat{\psi}_k$ é muito pequeno. Neste caso, ela pode gerar intervalos para ψ_k com valores extremos, por exemplo tendendo a infinito.

Frente a estas dificuldades do ICA, tais como problemas de cálculo numérico, de fronteira e amostras pequenas, outros métodos para construção dos IC são estudados.

2.2 Intervalos de Confiança Bootstrap

É esperado que, para amostras pequenas, o EMV não atenda às propriedades assintóticas, logo o ICA pode apresentar vários problemas. Uma alternativa nestes casos é utilizar o método bootstrap para construir a distribuição empírica do EMV.

O bootstrap é um método de reamostragem proposto por Efron (1979). Para tal reamostragem existem duas opções: reamostrar da distribuição geradora dos dados, caso conhecida, ou reamostrar dentro da amostra. Esta última forma é conhecida como bootstrap não-paramétrico, que será o utilizado neste trabalho por ser uma técnica mais geral. Este método baseia-se no fato de que a distribuição boostrap F^* converge, em probabilidade quando o número de replicações bootstrap tende ao infinito, para a distribuição empírica estimada dos dados $\hat{F}$. Esta por sua vez converge, em probabilidade quando o tamanho

amostral tende ao infinito, para a distribuição dos dados F (Efron & Tibshirani, 1993).

As reamostras são obtidas supondo que cada observação tenha igual massa de probabilidade, dita distribuição empírica $\hat{F}$. Seja $X = (X_1, X_2, ..., X_n)$ uma amostra aleatória de uma distribuição qualquer. A amostra bootstrap é calculada com reposição sobre a amostra original e é representada como $X^* = (X_1^*, X_2^*, ..., X_n^*)$. A proposta original do bootstrap supõe independência entre as observações. Como dados de uma série temporal carregam interdependência é necessário fazer uma adaptação do método original. No caso de modelos estruturais, Stoffer & Wall (1991) propuseram construir as séries bootstrap através de reamostragem dos resíduos do modelo ajustado. O método é apresentado a seguir.

Do FK, encontram-se as inovações $\upsilon_t \sim N(0, F_t)$. Note que o FK retorna estes valores para valores fixos dos hiperparâmetros, $\upsilon_t = \upsilon_t(\psi)$. Padronizando as inovações de forma a garantir média 0 e variância 1, tem-se:

$$e_t(\hat{\psi}) = \frac{\upsilon_t(\hat{\psi}) - \bar{\upsilon_t}(\hat{\psi})}{\sqrt{F_t(\hat{\psi})}},$$

onde $t = 1, ..., n$ e $\hat{\psi}$ é o EMV de ψ.

Reamostrando de $e_t(\hat{\psi})$ tem-se os erros bootstrap $e_t^*(\hat{\psi})$, que serão utilizados nas equações do FK para gerar y_t^*.

Seja $S_t = [a_t, y_{t-1}]^T$, onde a_t é uma estimativa para o vetor de estados α_t. Tome

$$S_{t+1} = \begin{bmatrix} T_t & 0 \\ z_t & 0 \end{bmatrix} S_t + \begin{bmatrix} T_t \upsilon_t z_t^T F_t^{-1} \sqrt{F_t} \\ \sqrt{F_t} \end{bmatrix} e_t. \tag{2}$$

Substituindo e_t por e_t^* em (2) calcula-se y_t^*. Usa-se, então, o FK em y_t^* para obter a função de verossimilhança e, a seguir, as estimativas de máxima verossimilhança dos hiperparâmetros na série bootstrap.

Intervalos de confiança bootstrap podem ser construídos pelo método percentílico (ICB) (Efron & Tibshirani, 1993). Nesse método geram-se B séries y_t^*, com suas B estimativas bootstrap de $\hat{\psi}, \hat{\psi}^*$. Um ICB para ψ_k; $\quad k = 1, ..., p$, com $100(1-\rho)\%$ de confiança, é simplesmente

$$[\hat{\psi}^*_{k(\frac{\rho}{2})}; \quad \hat{\psi}^*_{k(1-\frac{\rho}{2})}],$$

ou seja, os percentis $\frac{\rho}{2}$ e $(1 - \frac{\rho}{2})$ da distribuição bootstrap de $\hat{\psi}_k^*$. O ICB não apresenta problemas de fronteira e pode ser assimétrico.

Capítulo 3

Intervalos de confiança baseados na função deviance

O método assintótico para construir intervalos de confiança é baseado na distribuição assintótica do EMV. Contudo, a distribuição do EMV em amostras pequenas pode diferir da distribuição assintótica. Além disso, por ser um intervalo simétrico, o ICA pode conter valores fora do espaço paramétrico, o que é conhecido como problema de fronteira. Sabe-se também que o bootstrap pode ser computacionalmente muito custoso caso sejam necessárias muitas reamostragens. Uma alternativa mais robusta para amostras pequenas pode ser derivada da razão de verossimilhança.

Seja $\boldsymbol{\theta}$ um vetor de parâmetros de dimensão p e $\hat{\boldsymbol{\theta}}$ o seu EMV. Considere o seguinte teste de hipótese:

$$H_0 : \boldsymbol{\theta} = \theta_{\mathbf{0}}$$
$$H_1 : \boldsymbol{\theta} \neq \theta_{\mathbf{0}}.$$

Tome $\hat{\boldsymbol{\theta}}_0$ como sendo o EMV para $\boldsymbol{\theta}$ sob H_0. A razão

$$\Lambda = \frac{L(\hat{\boldsymbol{\theta}}_0)}{L(\hat{\boldsymbol{\theta}})}$$

é chamada de razão de verossimilhança (RV). Desta forma, uma região de confiança (RC) para $\boldsymbol{\theta}$ com $100(1-\rho)\%$ de confiança pode ser obtida pelos valores de $\boldsymbol{\theta}$ que satisfazem

$$-2log(\Lambda) \leq \chi^2_\rho(p)$$

onde $\chi^2_\rho(p)$ é o quantil ρ da distribuição Qui-quadrado com p graus de liberdade (Wilks, 1938).

A Figura 3.1 mostra um exemplo para o limite da região de confiança de 95% obtida para um MNL ($\theta = (\sigma^2_\eta, \sigma^2_\epsilon) = (0,5;1)$), isto é, as raízes da função $-2log(\Lambda) - \chi^2_{0,05}(2)$.

Contudo, regiões de confiança não têm uma fácil interpretação e seu custo computacional pode ser alto. Logo uma possível alternativa é construir IC marginais para cada parâmetro fixando os parâmetros desconhecidos nas suas respectivas EMV. Desta forma, a

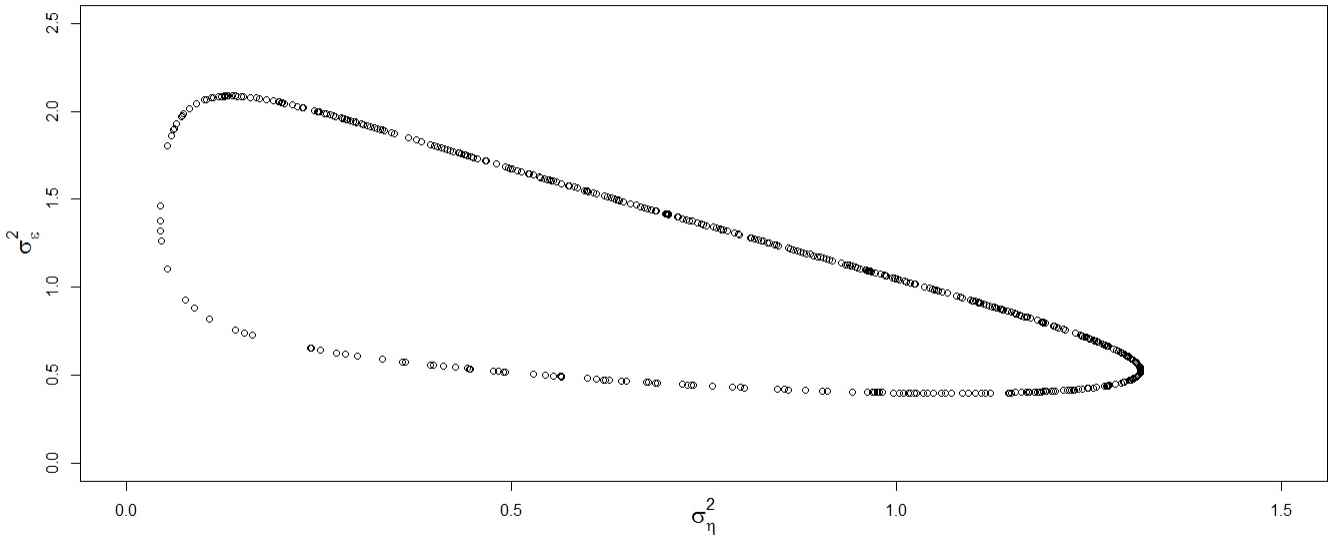

Figura 3.1: Exemplo de construção da RC para $(\sigma^2_\eta, \sigma^2_\epsilon)$ usando a RV.

função $-2log(\Lambda)$ fica em função de um único parâmetro e, consequentemente, é utilizada a distribuição Qui-quadrado com 1 grau de liberdade ($p = 1$). Estes serão os intervalos utilizados na simulação, realizada no Capítulo 5 denotados como intervalos de confiança marginais (ICM). Por exemplo no caso do MNL, um ICM para σ^2_η com $100(1-\rho)\%$ de confiança é composto pelos valores de σ^2_η que satisfazem

$$-2log\left(\frac{L(\sigma^2_\eta, \hat{\sigma}^2_\epsilon)}{L(\hat{\sigma}^2_\eta, \hat{\sigma}^2_\epsilon)}\right) \leq \chi^2_\rho(1).$$

A cobertura do ICM pode também ser aferida na região gerada pela interseção de todos os intervalos marginais para cada parâmetro, de forma que todos terão a mesma cobertura. Esta interseção dos ICM é uma aproximação da região de confiança que seria gerada com a distribuição $\chi^2_{(p)}$.

A Figura 3.2 mostra um exemplo para os intervalos de confiança de 95% aproximados obtidos para os parâmetros σ^2_η e σ^2_ϵ de um MNL ($\theta = (\sigma^2_\eta, \sigma^2_\epsilon) = (0,5;1)$) juntamente com a RC já apresentada na Figura 3.1. A altura do retângulo representa o ICM para σ^2_ϵ e sua largura representa o ICM para σ^2_η. O ICM, considerando $\chi^2_{0,05}(1)$, para σ^2_η foi [0,07; 0,74] e para σ^2_ϵ foi [0,72; 1,77]. Pode-se observar que o ICM gera uma subestimação da RC uma vez que foi usado 1 grau de liberdade.

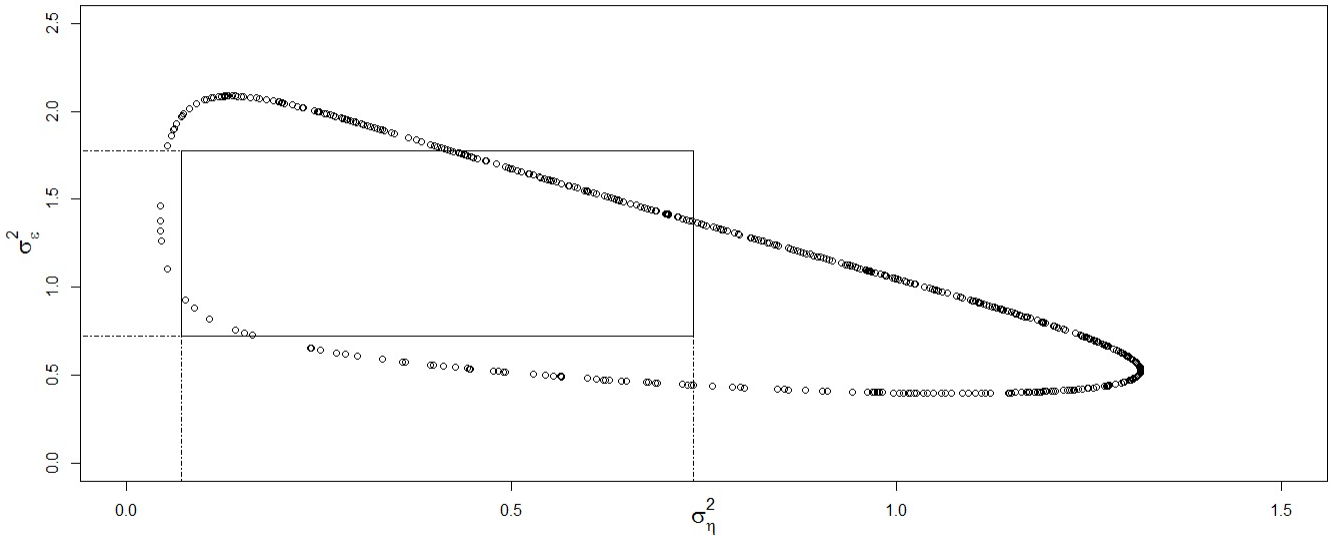

Figura 3.2: Exemplo de construção da RC aproximada para $(\sigma^2_\eta, \sigma^2_\epsilon)$ usando o ICM, representado pelo retângulo.

3.1 Intervalo de Confiança baseado na estatística Deviance

Quando um modelo estatístico é construído para explicar as relações entre as variáveis do estudo pode haver o interesse em inferir sobre apenas parte do vetor paramétrico para a compreensão do fenômeno em estudo. Neste caso, pode-se definir um vetor de parâmetros de interesse (λ) e um vetor de parâmetros de perturbação (δ), tal que $\boldsymbol{\theta} = (\lambda, \delta)$ é o vetor de parâmetros do modelo.

Por exemplo, seja o vetor paramétrico dado por $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_p)$. Neste caso, pode-se tomar $\lambda = \theta_k$ como parâmetro de interesse e $\delta = \theta_{(-k)}$, ou seja o vetor $\boldsymbol{\theta}$ sem a k-ésima componente, como parâmetro de perturbação.

A função Deviance, dada por $D(\theta_k) = -2[logL(\theta_k, \tilde{\theta}_{(-k)}) - logL(\hat{\boldsymbol{\theta}})]$ tem distribuição $\chi^2(1)$. O valor $\tilde{\theta}(-k)$ é o EMV da função de verossimilhança restrita a θ_k fixado no valor que deseja-se calcular a Deviance.

Desta maneira, como cada avaliação de $D(\theta_k)$ necessita de uma maximização para encontrar $\tilde{\theta}_{(-k)}$ é computacionalmente mais interessante considerar $\tilde{\theta}_{(-k)}$ fixo no EMV irrestrito $\hat{\theta}_{(-k)}$. Esta aproximação desconsidera possíveis não-ortogonalidades entre os parâmetros.

Para os modelos estruturais $\theta = \psi = (h, Q_{11}, ..., Q_{(p-1)(p-1)})^T$, logo um intervalo de confiança para ψ_k, com $100(1-\rho)\%$ de confiança, é composto pelos valores de ψ_k que satisfazem

$$D(\psi_k) = -2[logL(\psi_k, \hat{\psi}_{(-k)}) - logL(\hat{\psi})] \leq \chi^2_\rho(1),$$

em que $\hat{\psi}_{(-k)}$ é o EMV irrestrito de $\psi_{(-k)}$. Os extremos do intervalo são os valores de ψ_k que satisfazem a igualdade $D(\psi_k) = \chi^2_\rho(1)$, logo é necessário um algoritmo para encontrar o zero da função. Estes serão os intervalos utilizados na simulação realizada na Seção 5, denotados como intervalos de confiança baseados na estatística Deviance (ICD). A Figura 3.3 mostra um exemplo para o quantil $\chi^2_{0,05}(1) = 3,84$, encontrando as raízes da função $D(\sigma^2_\eta)$ para o MNL ($\psi = (0,5;1)$). Os limites inferior e superior do IC são os valores de σ^2_η que satisfazem a igualdade $D(\sigma^2_\eta) = \chi^2_{0,05}(1)$.

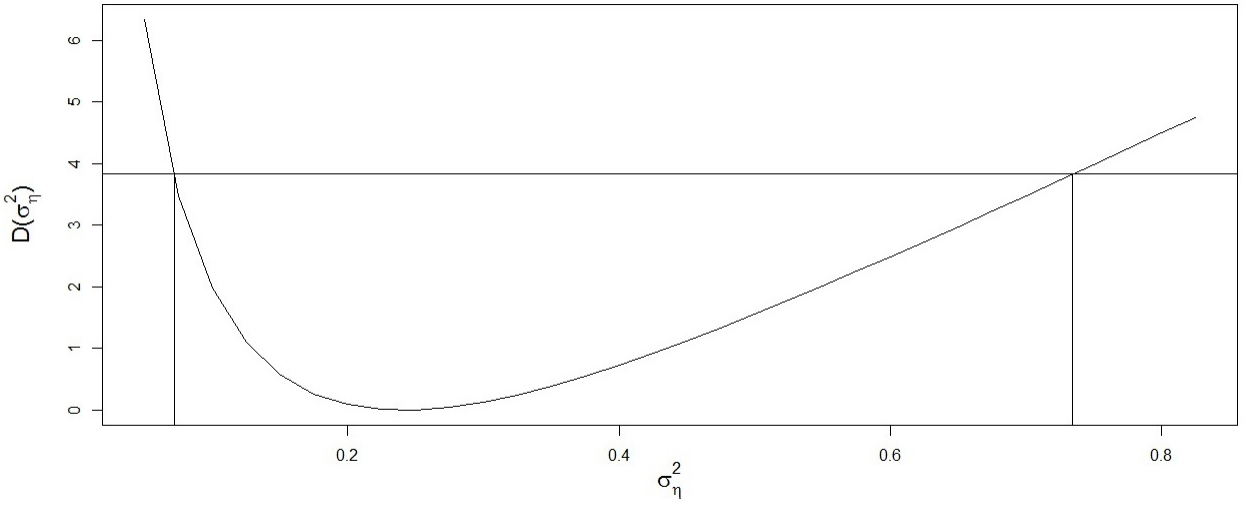

Figura 3.3: Exemplo de construção do ICD para o parâmetro σ^2_η.

Um algoritmo robusto, simples e competitivo em relação a métodos mais complexos é o Método Pégaso para obtenção de raízes (Dowell & Jarratt, 1972), que será descrito em detalhes no Capítulo 4. Espera-se também que este método seja mais rápido que o método bootstrap, uma vez que a função de verossimilhança já foi construida usando o FK e não será necessário fazer várias reamostragens bootstrap seguidas de maximização das novas funções de verossimilhança.

Para simplificar os cálculos, acabou-se usando aproximações no ICM e no ICD. A diferença entre eles, basicamente, é a cobertura e a região. No ICM é construída uma RC e a cobertura aferida nela, enquanto no ICD são construídos IC para cada hiperparâmetro e a cobertura é aferida em cada um deles separadamente.

3.2 Intervalo Signed Root Deviance Profile

Uma alternativa ao ICD é o IC *Signed Root Deviance Profile* (Chen & Jennrich, 1996), baseado na estatística sinal da razão de verossimilhança, $sinal(\theta - \hat{\theta})\frac{L(\hat{\theta}_0)}{L(\hat{\theta})}$ (Barndorff-Nielsen, 1986).

Seja $\hat{\theta}$ o EMV de θ. A função *Signed Root Deviance Profile*, para θ_k é definida como:

$$z^*(\theta_k) = sinal(\theta_k - \hat{\theta}_k)\sqrt{-2(logL(\theta_k, \tilde{\theta}_{(-k)}) - logL(\hat{\theta}))},$$

onde $k = 1, 2, ..., p$ e $\tilde{\theta}_{(-k)}$ é o EMV da função de verossimilhança restrita a θ_k fixado no valor que deseja-se calcular a função.

Novamente, substitui-se $\tilde{\theta}_{(-k)}$ por $\hat{\theta}_{(-k)}$ visando um menor tempo computacional. No texto de Chen & Jennrich (1996) é proposta uma maneira de encontrar o IC sem utilizar processos de maximização. Esta abordagem contudo necessita da matriz hessiana, que como dito no Capítulo 2 não é uma tarefa trivial.

Os valores de θ_k que satisfazem $-z_0 \leq z^*(\theta_k) \leq z_0$, onde z_0 é o quantil $\rho/2$ da Normal Padrão, pertencem ao IC Signed Root Deviance Profile(ICS) para θ_k.

Pode-se então escrever o ICS de cobertura $100(1 - \rho)\%$ para os modelos estruturais fazendo $\theta_k = \psi_k; \quad k = 1, 2, ..., p$ em função de z^*:

$$z^{*-1}(-z_0) \leq \psi_k \leq z^{*-1}(z_0), \tag{3}$$

onde z_0 é o quantil $\rho/2$ da Normal Padrão.

A Figura 3.4 ilustra o procedimento acima no caso do MNL ($\psi = (0,5; 1)$) para σ^2_η e assumindo um nível $\rho = 0.05$. Os limites do IC são os valores de σ^2_η que satisfazem $z^*(\sigma^2_\eta) = -1,96$ e $z^*(\sigma^2_\eta) = 1,96$.

Este tipo de IC tem uma construção e interpretação direta, ao contrário do ICM, que no caso multiparamétrico pode gerar dificuldades de interpretação, e é uma melhoria sobre o ICD. O ICS também tem a vantagem de poder ser assimétrico e não apresenta problemas de fronteira.

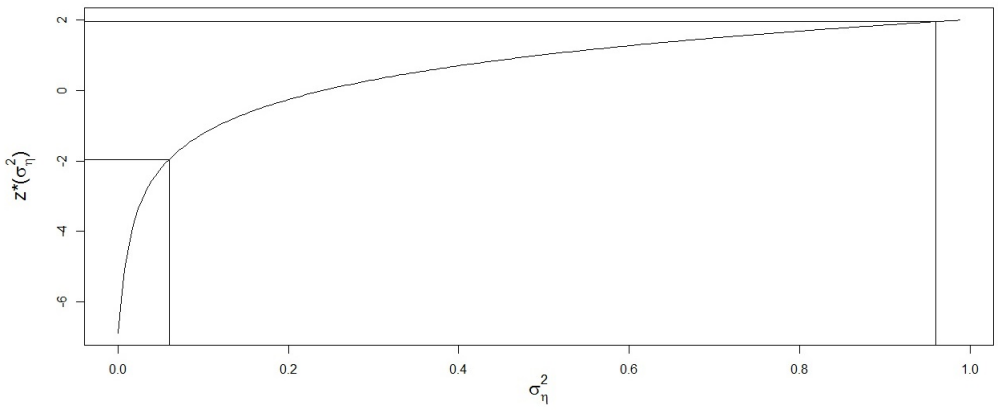

Figura 3.4: Exemplo de construção do ICS para o parâmetro σ^2_η.

Capítulo 4

Métodos computacionais

A Seção 3 cita os vários métodos computacionais auxiliares para o cálculo dos intervalos propostos. Esta seção explicará brevemente a ideia por trás destes algoritmos.

4.1 Método numérico para cálculo da matriz de informação de Fisher

Para a construção do ICA é necessário calcular a matriz de informação de Fisher:

$$I_{ij}(\psi) = \frac{1}{2}\sum_t \Big[tr\Big(F_t^{-1}\frac{\partial F_t}{\partial \psi_i}F_t^{-1}\frac{\partial F_t}{\partial \psi_j}\Big)\Big] + E\Big[\sum_t \frac{\partial \upsilon_t}{\partial \psi_i}^T F_t^{-1}\frac{\partial \upsilon_t}{\partial \psi_j}\Big];$$

onde $i, j = 1, 2, ..., p$, $t = 1, 2, ..., n$ e $tr(A)$ é o traço da matriz A.

O método descrito por Harvey (1989) para construção das derivadas de F_t e υ_t consiste em adicionar a um parâmetro ψ_i um valor de perturbação δ_i, fixados os outros parâmetros. A escolha de δ_i, neste caso, foi documentada em Franco et.al. (2008). A seguir, obtém-se novos valores para υ_t e F_t: $\upsilon_t^{(i)}$ e $F_t^{(i)}$, respectivamente e utiliza-se a aproximação $\frac{\partial \upsilon_t}{\partial \psi_i} = \frac{\upsilon_t^{(i)} - \upsilon_t}{\delta_i}$ e $\frac{\partial F_t}{\partial \psi_i} = \frac{F_t^{(i)} - F_t}{\delta_i}$ no cálculo da matriz de informação de Fischer. O Algoritmo 1 ilustra o procedimento descrito acima.

4.2 Método Pégaso para obtenção de raízes

Os limites do ICM são calculados com o algoritmo Pégaso (Dowell & Jarratt, 1972) para encontrar raízes de uma função qualquer. Este algoritmo é uma melhora ao método de aproximação linear da Falsa Posição (Campos, 2007). São consideradas raízes para $D(\psi_k) = \chi^2_\rho(1)$ valores de ψ_k pertencentes ao intervalo $[\chi^2_\rho(1) - \tau; \chi^2_\rho(1) + \tau]$, onde $k = 1, 2, ..., p$ e ρ é o nível de confiança do intervalo. O erro que tolera-se cometer em relação à verdadeira raiz é representado por τ.

Algoritmo 1 Algoritmo para construção da diagonal principal da matriz de informação de Fisher

Entrada: $y_t, \psi, \delta_i, z_t, T_t, R_t$

Com o Filtro de Kalman obtém-se $F_t, K_t, a_t, P_t, \upsilon_t$

Para i de 1 a p

Some a ψ_i uma quantidade δ_i

$(F_t^{(i)}, K_t^{(i)}, a_t^{(i)}, P_t^{(i)}, \upsilon_t^{(i)})$ é o resultado do Filtro de Kalman com o novo valor de ψ

Tome $\frac{\partial \upsilon_t}{\partial \psi_i}$ como a aproximação $\frac{\upsilon_t^{(i)} - \upsilon_t}{\delta_i}$

Tome $\frac{\partial F_t}{\partial \psi_i}$ como a aproximação $\frac{F_t^{(i)} - F_t}{\delta_i}$

Modifique a entrada ii da matriz de informação de Fisher para

$$\frac{1}{2}\sum_t \left[tr\left(F_t^{-1}\frac{\partial F_t}{\partial \psi_i}F_t^{-1}\frac{\partial F_t}{\partial \psi_i}\right)\right] + E\left[\sum_t \frac{\partial \upsilon_t}{\partial \psi_i}^T F_t^{-1}\frac{\partial \upsilon_t}{\partial \psi_i}\right]$$

Fim para

Retorna: Matriz de informação de Fisher

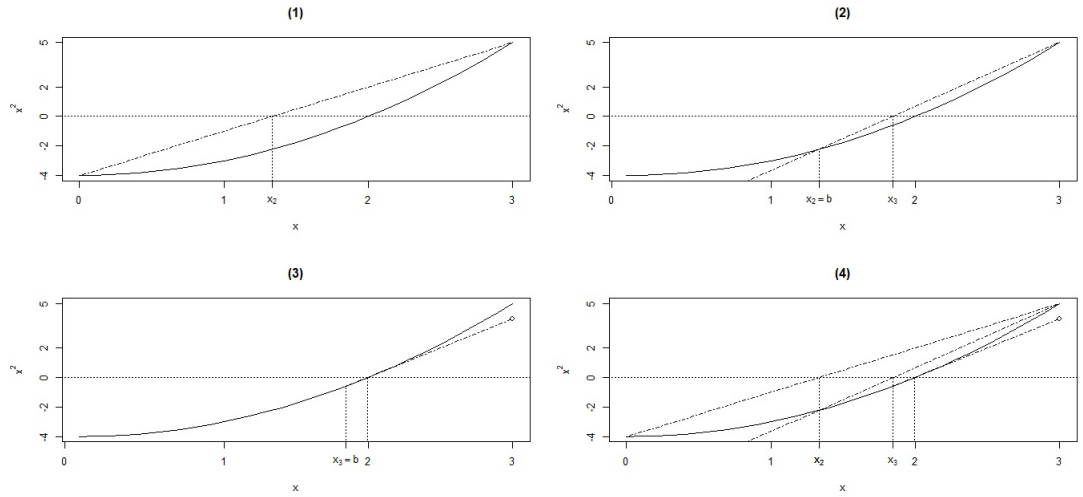

Figura 4.1: Exemplo de construção do ICS para o parâmetro σ_η^2.

Este algoritmo é baseado em aproximações lineares da função de interesse, obtendo uma sequência de valores que convergem para a verdadeira raiz, isolada em um intervalo conhecido $[a, b]$. Esta sequência é obtida pela fórmula:

$$x_{j+1} = x_j - \frac{f(x_j)}{f(x_j) - f(x_{j-1})}(x_j - x_{j-1})$$

onde $x_0 = a$, $x_1 = b$, $j = 0, 1, 2, ...$ e $f(x)$ uma função genérica.

Para garantir o isolamento da raiz é necessário que $f(x_{j-1})f(x_j) \leq 0$. Caso o valor de $f(x_{j+1})$ tenha mesmo sinal que o valor de $f(x)$, i.e: $f(x_{j+1})f(x) \geq 0$, para evitar problemas de retenção de ponto, e consequente não convergência, o valor de $f(x_{j-1})$ é multiplicado por

Algoritmo 2 Algoritmo Pégaso

entrada: $f, a, b, \tau, \text{MaxIter}$
Tome F_a como $f(a)$
Tome F_b como $f(b)$
Tome x como b
Tome F_x como $f(b)$
Inicialize o número de iterações Iter em 0
Enquanto flag $= 0$
  ΔX recebe $-F_x/(F_b - F_a) * (b - a)$
  x recebe $x + \Delta X$
  F_x recebe $f(x)$
  Se $(|\Delta X| \leq \tau)$ **e** $(|F_x| \leq \tau)$ **ou** Iter $\geq$ MaxIter
   flag recebe 1
  Fim se
  Se $F_x * F_b \leq 0$
    a recebe b
    F_a recebe F_b
  Senão
    F_a recebe $F_a * F_b/(F_b + F_x)$
  Fim se
  b recebe x
  F_b recebe F_x
  Iter $\leftarrow$ Iter $+ 1$

Fim Enquanto

Retorne: x, F_x, Iter

$f(x_j)/(f(x_j)f(x_{j+1}))$, tornando possível traçar a aproximação linear por um ponto não pertencente à curva de $f(x)$. Esta atualização também acaba por agilizar o método (ver Figura 4.1).

A Figura 4.1 mostra quatro iterações do algoritmo: no primeiro passo, tem-se o intervalo em que a raíz está isolada $[x_0, x_1]$, traça-se a aproximação linear da função e encontra-se x_2. A seguir atualizam-se os extremos do intervalo, que passa a ser $[x_2, x_1]$, traça-se a aproximação linear e encontra-se x_3. Como $f(x_3)$ e $f(x_2)$ têm o mesmo sinal, atualiza-se o valor de $f(x_1)$ da maneira descrita acima, gerando um ponto fora do gráfico da função. O valor x_4 é encontrado utilizando este ponto. A imagem final une os passos anteriores em um único gráfico.

O Algoritmo 2 mostra como programar este método.

O método Pégaso tem uma implementação fácil e, por ser uma aproximação linear, não necessita de cálculo de derivadas, como o método de Newton ou outros métodos que utilizam

tangente. Além disso é um método robusto e competitivo quando comparado com métodos mais sofisticados (Campos, 2007).

Contudo, algoritmos para encontrar raizes necessitam de um intervalo onde a raiz está isolada e que a função assuma sinais contrários nos extremos deste intervalo. Uma maneira de encontrar mais de uma raiz, e isolá-las, dado um intervalo maior, é particionar este em intervalos menores e executar o Pégaso em cada um deles. Esta metodologia é utilizada na função uniroot.all do pacote rootSolve do R (Soetaert, 2009).

4.3 Busca Binária

Para o ICS o algoritmo Pégaso não apresentou bons resultados, principalmente devido à natureza assintótica da função *Signed Root Deviance Profile* z^*, que pode apresentar grandes variações perto de 0. Contudo, para o cálculo deste intervalo, basta aferir quais valores de ψ_i satisfazem à Equação (3), com $i = 1, 2, ..., p$.

A partir de um intervalo $(0; b]$ onde acredita-se estar o parâmetro, faz-se uma discretização em N pontos e utiliza-se uma busca para encontrar os extremos aproximados do intervalo. Ou seja, valores de ψ_i que sejam os mais próximos a $z^*(\psi_i) = -z_0$ e $z^*(\psi_i) = z_0$. Note que a escolha de N influencia no erro cometido ao escolher os extremos do intervalo. Por exemplo, se $b = 3$ e $N = 600$, o erro máximo, em ψ seria de $b/N = 0,005$.

Uma maneira de encontrar estes extremos do IC seria buscar de forma crescente e linear ao longo da discretização feita. Contudo, como a função z^* é crescente, pode-se utilizar um procedimento de busca binária (Cormen et al., 1990), que consiste em percorrer a discretização de pontos procurando recursivamente na metade do espaço que contem o valor desejado.

Discretizado o intervalo, a busca por um valor z começa pelo ponto médio $m = b/2$ e segue para o subintervalo apropriado: $(0; m)$ se $z^*(m) \geq z$ ou $(m; b]$ se $z^*(m) \leq z$, atualiza-se o ponto médio e os valores extremos do intervalo de busca e prosegue-se de maneira recursiva a busca. A busca é finalizada quando o valor procurado está entre $z^*(m - \frac{1}{N})$ e $z^*(m + \frac{1}{N})$ e retorna o índice $M \in \{m - \frac{1}{N}; m; m+; \frac{1}{N}\}$ de imagem mais próxima a z. O Algoritmo 3 implementa esta busca.

Esta abordagem reduz muito o tempo computacional, se comparado com a busca linear. Enquanto o tempo da busca linear cresce linearmente com o valor de N escolhido, o tempo da busca binária cresce em função do $log_2(N)$.

Algoritmo 3 Busca Binária

Entrada: f, b, z, N

m recebe 0
M recebe N
Enquanto $M > m$
 i recebe o menor inteiro mais próximo de $\frac{M+m}{2}$
 j recebe $(i-1) * b/N$
 F recebe $f(j)$
 Se $f(j - \frac{1}{N}) < z < f(j + \frac{1}{N})$
  M recebe m
 Senão
  Se $F < z$
   m recebe $i + 1$
  Senão
   M recebe $i - 1$
  Fim Se/Senão
 Fim Se/Senão
Fim enquanto
Retorna: $x \in \{j - \frac{1}{N}; j; j + \frac{1}{N}\}$ de imagem mais próxima a z

Capítulo 5

Estudo Monte Carlo

Estudos Monte Carlo (MC) foram feitos para comparar os intervalos de confiança descritos nas Seções 2 e 3: Assintótico (ICA), Bootstrap (ICB), Marginal (ICM), Deviance (ICD) e Signed Root Deviance Profile (ICS). Em um primeiro estudo, os erros das observações seguem uma distribuição gaussiana ou Normal. Em seguida, é feito um estudo considerando os erros com uma distribuição não-gaussiana para as observações, visando verificar a robustez dos intervalos para este caso.

A cobertura nominal foi fixada em 95% e os intervalos são comparados quanto à taxa de cobertura e amplitude. As simulações feitas avaliam os IC para o Modelo de Nível Local (MNL), Modelo de Tendência Local (MTL) e Modelo Estrutural Básico (MEB), com sazonalidade $s = 12$.

Um estudo preliminar, com 200 simulações Monte Carlo, foi realizado com a intenção de verificar a cobertura dos IC descritos para amostras grandes. Para isso foram utilizadas séries de tamanho $n = 1000$. As outras simulações para tamanhos de amostra menores: 60, 200 e 500, foram realizadas considerando 1000 replicações Monte Carlo.

É apresentada também uma tabela de comparação de tempo computacional entre os métodos a fim de melhor classificá-los. A verossimilhança é maximizada com o algoritmo quasi-newton BFGS (Shanno,1970). Este faz um máximo de 50 iterações para achar o valor de máximo. Para facilitar a reprodução e verificação dos resultados são explicitadas as constantes utilizadas nos métodos computacionais.

O pacote SsfPack (Koopman et al., 1999) do software Ox (Doornik, 2006) contém algoritmos para o cálculo do Filtro de Kalman e da função de verossimilhança, sendo o utilizado na implementação. Para gerar os dados é considerado um *burn-in* igual a 100 de forma a evitar influência de valores iniciais.

O número de iterações bootstrap é $B = 500$, este valor é baseado em estudos apresentados em Franco & Santos (2010). O valor de δ_i para o método numérico de cálculo da matriz de informação de Fisher é fixado em $0,0001$, também segundo estudos em Franco et al.(2008). Valores menores não influenciam muito na amplitude dos intervalos e valores

maiores aumentam a mesma.

O intervalo $[0,000001; 5]$ é utilizado para construir os ICD e este é particionado em $P = 200$ subintervalos menores de mesmo tamanho. São consideradas raízes para $D(\psi) = \chi^2_{0,05}(1)$ valores de função pertencentes ao intervalo $[\chi^2_{0,05}(1) - 0,0001; \chi^2_{0,05}(1) + 0,0001]$, isto é, tomando $\tau = 0,0001$. O algoritmo do método Pégaso para encontrar raízes faz um máximo de 50 iterações em cada subintervalo.

Para o ICS o intervalo $[0,000001; 3]$ é discretizado em $N = 600$ pontos. A escolha de N influencia no erro cometido ao escolher os extremos do intervalo, neste caso de $0,005$. Os extremos destes intervalos iniciais podem ser escolhidos considerando uma estimativa de ψ, seja por EMV ou bootstrap.

Visando mensurar a eficiência computacional dos métodos foi calculado o tempo utilizando a função *today()* do Ox. As simulações foram feitas em um Intel Core 2Quad com 4GB de memória ram em ambiente Windows 7 Pro.

5.1 Resultados para os erros com distribuição gaussiana

A Tabela 5.1 apresenta as coberturas médias de cada método para amostras de tamanho grande ($n = 1000$) para o Modelo de Nível Local (MNL), Modelo de Tendência Local (MTL) e Modelo Estrutural Básico (MEB), considerando 200 repetições MC.

Tabela 5.1: Coberturas dos intervalos propostos para uma amostra de tamanho $n = 1000$.

Modelo	ψ	ICA	ICB	ICM	ICD	ICS
MNL	$\sigma^2_\eta = 0,5$	**93,0**	**93,5**	85,5	92,0	**93,5**
	$\sigma^2_\epsilon = 1$	**95,0**	**95,0**	85,5	91,5	**96,0**
MTL	$\sigma^2_\eta = 0,5$	**96,0**	**95,0**	87,5	**95,5**	**96,5**
	$\sigma^2_\xi = 0,1$	**94,5**	**94,0**	87,5	91,5	**95,0**
	$\sigma^2_\epsilon = 1$	91,0	**93,5**	87,5	100	**94,5**
MEB	$\sigma^2_\eta = 0,5$	**95,0**	**95,0**	85,0	92,0	**97,0**
	$\sigma^2_\xi = 0,1$	**96,0**	**94,0**	85,0	92,0	**93,5**
	$\sigma^2_\omega = 0,03$	**93,0**	91,5	85,0	100	**95,0**
	$\sigma^2_\epsilon = 1$	**96,0**	**94,0**	85,0	100	**96,0**

Obs.: Em negrito: coberturas a uma distância de 2 pontos percentuais do nível nominal

Com os dados da Tabela 5.1 fica evidente a pouca efetividade do ICM e do ICD. No ICD é feita uma aproximação que fixa os parâmetros de ruído em seus EMV irrestritos. Esta mesma abordagem é utilizada na construção do ICM, contudo este processo não considera a aleatoriedade destes parâmetros de ruído, de tal forma que as coberturas destes IC ficam distantes do nível fixado de 95%. Nota-se também a subestimação da RC feita pelo ICM, como comentado na Figura 3.1.

Desta forma, serão excluidos estes métodos para as comparações a seguir. Os outros métodos tiveram suas coberturas bem próximas do nível fixado de 95%.

A Tabela 5.2 mostra os resultados para o MNL, a Tabela 5.3 para o MTL e a 5.4 para o MEB, para amostras de tamanho 60, 200 e 500, considerando 1000 repetições MC.

Tabela 5.2: Resultados da simulação MC para o MNL com erros Normais para as observações.

		ICA		ICB		ICS	
n	ψ	Cobertura	Amplitude	Cobertura	Amplitude	Cobertura	Amplitude
60	$\sigma_\eta^2 = 0,5$	87,1	0,94	90,0	1,05	**94,9**	1,05
	$\sigma_\epsilon^2 = 1$	91,1	1,12	89,6	1,07	92,7	1,20
200	$\sigma_\eta^2 = 0,5$	92,6	0,51	**93,4**	0,53	**95,2**	0,57
	$\sigma_\epsilon^2 = 1$	**94,2**	0,61	**93,3**	0,62	**95,6**	0,64
500	$\sigma_\eta^2 = 0,5$	92,7	0,33	92,3	0,33	92,2	0,33
	$\sigma_\epsilon^2 = 1$	**93,7**	0,40	**93,0**	0,40	**94,5**	0,40

Obs.: Em negrito estão as coberturas que estão a uma distância de 2 pontos percentuais do nível nominal

Tabela 5.3: Resultados da simulação MC para o MTL com erros Normais para as observações.

		ICA		ICB		ICS	
n	ψ	Cobertura	Amplitude	Cobertura	Amplitude	Cobertura	Amplitude
60	$\sigma_\eta^2 = 0,5$	98,1	2,40	99,7	1,99	**94,7**	1,64
	$\sigma_\xi^2 = 0,1$	80,3	0,27	88,6	0,25	**95,2**	0,34
	$\sigma_\epsilon^2 = 1$	**93,5**	1,48	**96,6**	1,45	**95,6**	1,48
200	$\sigma_\eta^2 = 0,5$	**96,1**	1,35	**96,6**	1,24	**94,2**	1,24
	$\sigma_\xi^2 = 0,1$	**93,8**	0,15	92,4	0,15	88,1	0,16
	$\sigma_\epsilon^2 = 1$	**94,9**	0,83	**96,1**	0,84	**95,5**	0,85
500	$\sigma_\eta^2 = 0,5$	**93,7**	0,86	**94,0**	0,84	**94,5**	0,84
	$\sigma_\xi^2 = 0,1$	92,1	0,09	**93,6**	0,11	**94,8**	0,10
	$\sigma_\epsilon^2 = 1$	**94,8**	0,53	**94,8**	0,54	**95,0**	0,54

Obs.: Em negrito estão as coberturas que estão a uma distância de 2 pontos percentuais do nível nominal

Nota-se nas Tabelas 5.2 a 5.4 que o ICS é, na maioria das vezes, o intervalo com cobertura mais próxima do nível fixado de 95%. Em geral, as coberturas ficam mais próximas do nível de 95% e as amplitudes diminuem com o crescimento do tamanho amostral. Este fato é esperado, uma vez que os EMV dos parâmetros ficam com uma variânca menor com o aumento do tamanho amostral.

Um estudo de tempo de simulação (em minutos) para os métodos ICA, ICB e ICS aplicados para o MNL, MTL e MEB, é apresentado na Tabela 5.5.

O método mais rápido, como esperado, foi o ICA. Apesar das desvantagens citadas no Capítulo 2, este método pode ser utilizado para aplicações que necessitam de um tempo computacional mínimo, se o tamanho da amostra for grande. Para amostras de tamanho

Tabela 5.4: Resultados da simulação MC para o MEB com erros Normais para as observações.

		ICA		ICB		ICS	
n	ψ	Cobertura	Amplitude	Cobertura	Amplitude	Cobertura	Amplitude
60	$\sigma^2_\eta = 0,5$	88,0	2,35	**94,0**	2,06	92,1	1,57
	$\sigma^2_\xi = 0,1$	78,5	0,26	90,0	0,37	92,2	0,33
	$\sigma^2_\omega = 0,03$	99,7	0,50	99,2	0,19	97,5	0,50
	$\sigma^2_\epsilon = 1$	91,5	1,82	**94,5**	2,26	**95,6**	1,82
200	$\sigma^2_\eta = 0,5$	91,7	1,36	**93,7**	1,57	**95,2**	1,28
	$\sigma^2_\xi = 0,1$	88,7	0,15	**93,5**	0,15	**94,9**	0,16
	$\sigma^2_\omega = 0,03$	86,1	0,08	88,1	0,07	**94,1**	0,09
	$\sigma^2_\epsilon = 1$	**93,1**	0,91	**93,9**	1,26	**95,5**	0,99
500	$\sigma^2_\eta = 0,5$	92,2	0,88	**93,0**	1,03	**95,6**	0,88
	$\sigma^2_\xi = 0,1$	92,8	0,10	**94,2**	0,10	**94,4**	0,10
	$\sigma^2_\omega = 0,03$	91,5	0,04	87,8	0,04	**94,0**	0,04
	$\sigma^2_\epsilon = 1$	**93,8**	0,59	91,8	0,74	**95,0**	0,81

Obs.: Em negrito estão as coberturas que estão a uma distância de 2 pontos percentuais do nível nominal

Tabela 5.5: Tempo, em minutos, necessário para 1000 simulações dos IC nos modelos MNL, MTL e MEB.

	MNL			MTL			MEB		
n	ICA	ICB	ICS	ICA	ICB	ICS	ICA	ICB	ICS
60	0,03	27,25	2,33	0,18	178,55	27,50	5,26	4467,35	1142,52
200	0,10	64,95	4,45	0,42	379,75	43,25	19,43	14217,55	3083,08
500	0,22	150,38	10,88	0,98	887,78	96,96	17,33	15204,15	3127,26

pequeno, as simulações mostram que este é o método com o pior desempenho quanto à cobertura dos intervalos.

Já o ICB, apesar de ser um método sem as desvantagens do ICA e com coberturas próximas das deste método, apresenta um tempo computacional elevado quando comparado com o ICS, cerca de 10 vezes mais lento, além de apresentar coberturas um pouco mais distantes do nível fixado.

O ICS, que foi o método com coberturas mais próximas do nível fixado de 95%, apresenta um tempo computacional mais eficiente que o do ICB, método alternativo ao ICA disponível na literatura.

5.2 Resultados para erros com distribuição não-gaussiana

Um grande atrativo tanto do ICB quanto do ICS é a possibilidade de construir os IC para os hiperparâmetros do modelo sem as limitações do ICA. Neste contexto, a qualidade desses IC é averiguada para dados não-gaussianos.

Para tanto, serão gerados dados com distribuição Gama na equação das observações da FEE. O ajuste, contudo é feito com o FK Gaussiano, apresentado no Capítulo 2.

Os parâmetros de forma a e escala b, são escolhidos de maneira que a variância desta distribuição Gama seja igual a 1. A distribuição é centrada em 0 via subtração da média teórica da mesma, $\frac{a}{b}$. Desta maneira, tem-se os erros das observações com média 0 e variância h_t, seguindo a FEE em (1). Assim, o parâmetro de forma foi escolhido como $a = \frac{16}{9}$ e parâmetro de escala $b = \frac{4}{3}$ de forma que $h_t = \sigma_\epsilon^2 = \frac{a}{b^2} = 1$, como no caso das simulações anteriores. O vetor de estados continua tendo distribuição normal.

O desenho do estudo MC é o mesmo da seção anterior. A Tabela 5.6 mostra os resultados para o MNL, a Tabela 5.7 para o MTL e a 5.8 para o MEB com dados gerados da maneira descrita acima.

Tabela 5.6: Resultados da simulação MC para o MNL com erros Gama para as observações.

		ICA		ICB		ICS	
n	ψ	Cobertura	Amplitude	Cobertura	Amplitude	Cobertura	Amplitude
60	$\sigma_\eta^2 = 0,5$	85,7	0,94	90,2	1,07	**94,1**	1,07
	$\sigma_\epsilon^2 = 1$	85,0	1,13	87,2	1,24	87,0	1,20
200	$\sigma_\eta^2 = 0,5$	90,5	0,51	**93,1**	0,53	**93,2**	0,53
	$\sigma_\epsilon^2 = 1$	85,9	0,62	87,6	0,67	86,9	0,64
500	$\sigma_\eta^2 = 0,5$	**94,0**	0,33	**94,6**	0,34	**94,4**	0,34
	$\sigma_\epsilon^2 = 1$	87,3	0,40	89,8	0,43	88,6	0,40

Obs.: Em negrito estão as coberturas que estão a uma distância de 2 pontos percentuais do nível nominal

Tabela 5.7: Resultados da simulação MC para o MTL com erros Gama para as observações.

		ICA		ICB		ICS	
n	ψ	Cobertura	Amplitude	Cobertura	Amplitude	Cobertura	Amplitude
60	$\sigma_\eta^2 = 0,5$	**96,8**	2,46	99,7	2,03	92,6	1,07
	$\sigma_\xi^2 = 0,1$	78,7	0,27	88,6	0,25	**94,2**	0,33
	$\sigma_\epsilon^2 = 1$	89,3	1,51	89,9	1,52	90,4	1,48
200	$\sigma_\eta^2 = 0,5$	97,1	1,35	97,1	1,24	**95,0**	0,85
	$\sigma_\xi^2 = 0,1$	89,3	0,15	**93,2**	0,87	**94,1**	0,16
	$\sigma_\epsilon^2 = 1$	89,9	0,84	92,1	0,87	91,2	0,85
500	$\sigma_\eta^2 = 0,5$	**94,1**	0,86	**93,6**	0,85	**94,6**	0,84
	$\sigma_\xi^2 = 0,1$	**93,3**	0,09	**94,1**	0,09	**95,3**	0,10
	$\sigma_\epsilon^2 = 1$	90,9	0,52	92,4	0,55	91,6	0,54

Obs.: Em negrito estão as coberturas que estão a uma distância de 2 pontos percentuais do nível nominal

Em geral, o ICS tem a cobertura mais próxima do nível fixado de 95%. Os métodos apresentados nas tabelas diminuem as amplitudes com o crescimento do tamanho amostral, além de ficarem com coberturas mais próximas do nível de 95%. Como esperado, o ICA tem coberturas muito distantes do nível de 95% principalmente para as amostras de tamanho 60 e 200, já que este método é construído baseado na suposição de normalidade dos dados. É também interessante notar que o ICB não apresenta um bom desempenho, apesar de não

Tabela 5.8: Resultados da simulação MC para o MEB com erros Gama para as observações.

		ICA		ICB		ICS	
n	ψ	Cobertura	Amplitude	Cobertura	Amplitude	Cobertura	Amplitude
60	$\sigma^2_\eta = 0,5$	88,4	2,34	**94,9**	2,06	**93,7**	1,61
	$\sigma^2_\xi = 0,1$	77,9	0,27	91,2	0,37	**93,2**	0,35
	$\sigma^2_\omega = 0,03$	100	0,49	99,8	0,19	98,2	0,48
	$\sigma^2_\epsilon = 1$	89,0	1,80	91,7	2,23	**93,0**	1,77
200	$\sigma^2_\eta = 0,5$	92,4	1,30	92,6	1,53	97,6	1,27
	$\sigma^2_\xi = 0,1$	89,8	0,15	**93,8**	0,16	**95,4**	0,17
	$\sigma^2_\omega = 0,03$	86,8	0,08	87,8	0,07	**94,8**	0,09
	$\sigma^2_\epsilon = 1$	92,6	0,92	91,8	1,27	**93,4**	1,00
500	$\sigma^2_\eta = 0,5$	**94,0**	0,89	92,4	1,04	**94,6**	0,89
	$\sigma^2_\xi = 0,1$	91,6	0,10	**93,8**	0,10	**94,6**	0,10
	$\sigma^2_\omega = 0,03$	86,4	0,04	87,6	0,04	92,0	0,05
	$\sigma^2_\epsilon = 1$	92,0	0,59	91,2	0,75	**93,0**	0,62

Obs.: Em negrito estão as coberturas que estão a uma distância de 2 pontos percentuais do nível nominal

fazer nenhuma suposição sobre a distribuição dos dados.

Um estudo de tempo de simulação (em minutos) para os métodos ICA, ICB e ICS aplicados para o MNL, MTL e MEB, é apresentado na Tabela 5.9, utilizando 1000 simulações Monte Carlo.

Tabela 5.9: Tempo, em minutos, necessário para 1000 simulações dos IC nos modelos MNL, MTL e MEB com erro Gama nas observações.

	MNL			MTL			MEB		
n	ICA	ICB	ICS	ICA	ICB	ICS	ICA	ICB	ICS
60	0,03	32,70	3,33	0,18	182,46	40,02	3,63	4480,33	1149,50
200	0,10	74,58	5,26	0,43	388,40	67,20	7,48	6828,06	1509,88
500	1,33	165,98	11,95	1,05	911,70	130,50	17,91	15411,70	3181,85

As tabelas de tempo seguem o mesmo padrão das simulações com dados gaussianos. Ainda que, em alguns poucos casos, o ICB tenha tido melhores coberturas, o tempo computacional do ICS aliado ao seu melhor desempenho quanto à taxa de cobertura são as grandes vantagens deste método.

Capítulo 6

Aplicações

Para ilustrar a utilização dos métodos de construção de IC para os hiperparâmetros dos ME descritos, serão apresentadas duas aplicações a dados reais. A primeira série é a de receitas com previdência arrecadada pelas Entidades Abertas de Previdência Complementar (EAPC), a segunda é a da log-incidência de casos de dengue em Belo Horizonte.

6.1 Receita arrecadada pelas EAPC

A série temporal em estudo é a da receita arrecadada, em bilhões de reais, pelas EAPC. Esta arrecadação de previdência complementar é supervisionada pelo governo e institui planos privados previdenciários mediante contribição facultativa dos participantes. Os dados foram retirados de boletins estatísticos da Superintendência de Seguros Privados (SUSEP), disponíveis para consulta através do SES (sistema gerador de estatísticas dos mercados supervisionados).

Os dados são mensais no período de junho de 2002 a dezembro de 2010, totalizando 103 observações. É tomado o logaritmo natural dos dados de maneira com que estes satisfaçam a hipotese de normalidade. A série da log-receita apresenta um p-valor de 0.07 para o teste de normalidade de Shapiro-Wilk.

A Figura 6.1 mostra a série em questão versus o tempo, com o ajuste de um MEB, já que a mesma apresenta uma clara sazonalidade. O ajuste foi feito desconsiderando as 13 primeiras observações, que foram utilizadas para calibrar o FK.

A Tabela 6.1 mostra os valores das estimativas de máxima verossimilhança para os hiperparâmetros, bem como a mediana, média e erro padrão das estimativas para as séries bootstrap, obtido utilizando-se 500 reamostragens. Pode-se notar uma baixa magnitude para o hiperparâmetro σ_ξ^2 (valor da ordem de magnitude 10^{-5}). Isto pode indicar que a inclinação da tendência seja não-estocástica. A mediana é tomada como referência para o IC bootstrap uma vez que os hiperparâmetros apresentam uma distribuição empírica assimétrica (ver Figura 6.2).

A Figura 6.3 mostra os resíduos do ajuste com o MEB, bem como estes resíduos contra os valores ajustados, um qq-plot, um histograma e os gráficos da FAC e FACP destes resíduos. Destes gráficos pode-se ver que os resíduos satisfazem as hipóteses de independência, homos-

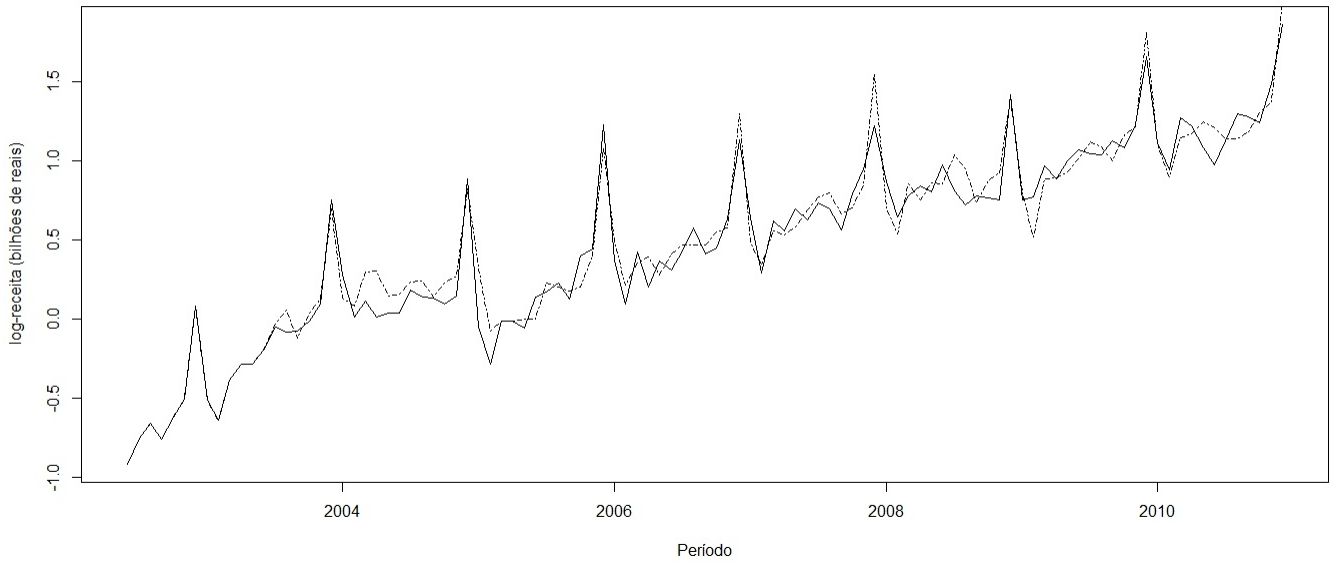

Figura 6.1: Gráfico do ajuste (linha tracejada) do MEB à série da log-receita arrecadada pelas EAPC (linha contínua).

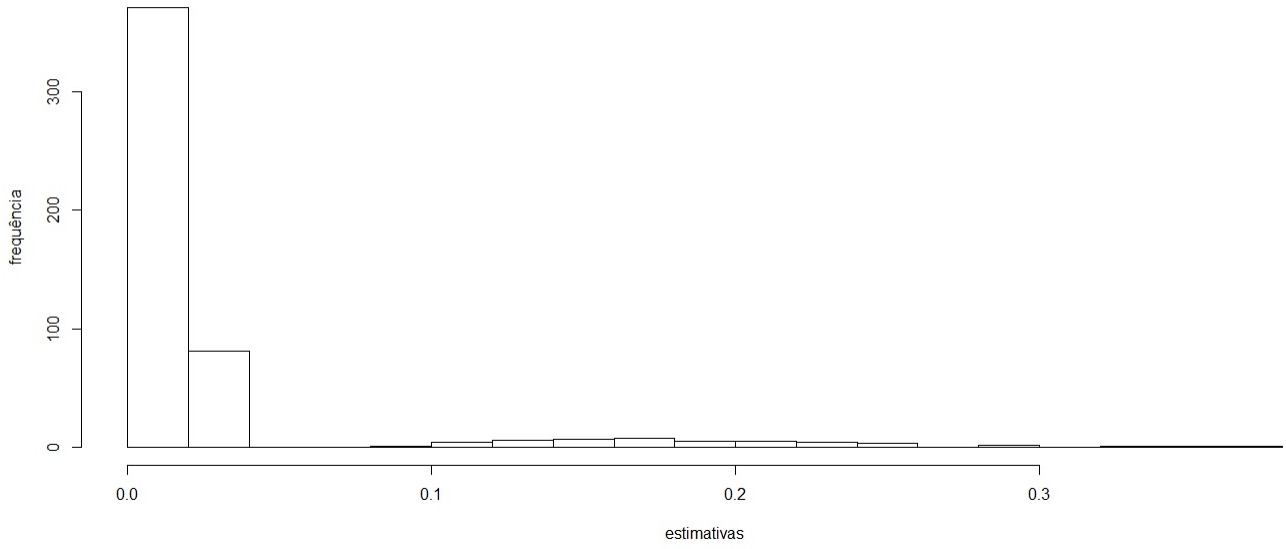

Figura 6.2: Histograma para as estimativas de σ_η^2 obtidas nas séries bootstrap.

cedasticidade e normalidade.

Tabela 6.1: Estimativa de máxima verossimilhança dos hiperparâmetros, mediana, média e erro padrão das estimativas dos hiperparâmetros das séries bootstrap

	σ^2_η	σ^2_ξ	σ^2_ω	σ^2_ϵ
EMV	$0,017$	$1\text{x}10^{-8}$	$0,018$	$0,001$
mediana	$0,016$	$1\text{x}10^{-9}$	$0,015$	$0,001$
média	$0,031$	$1\text{x}10^{-5}$	$0,014$	$0,008$
erro padrão	$0,055$	$1\text{x}10^{-4}$	$0,007$	$0,018$

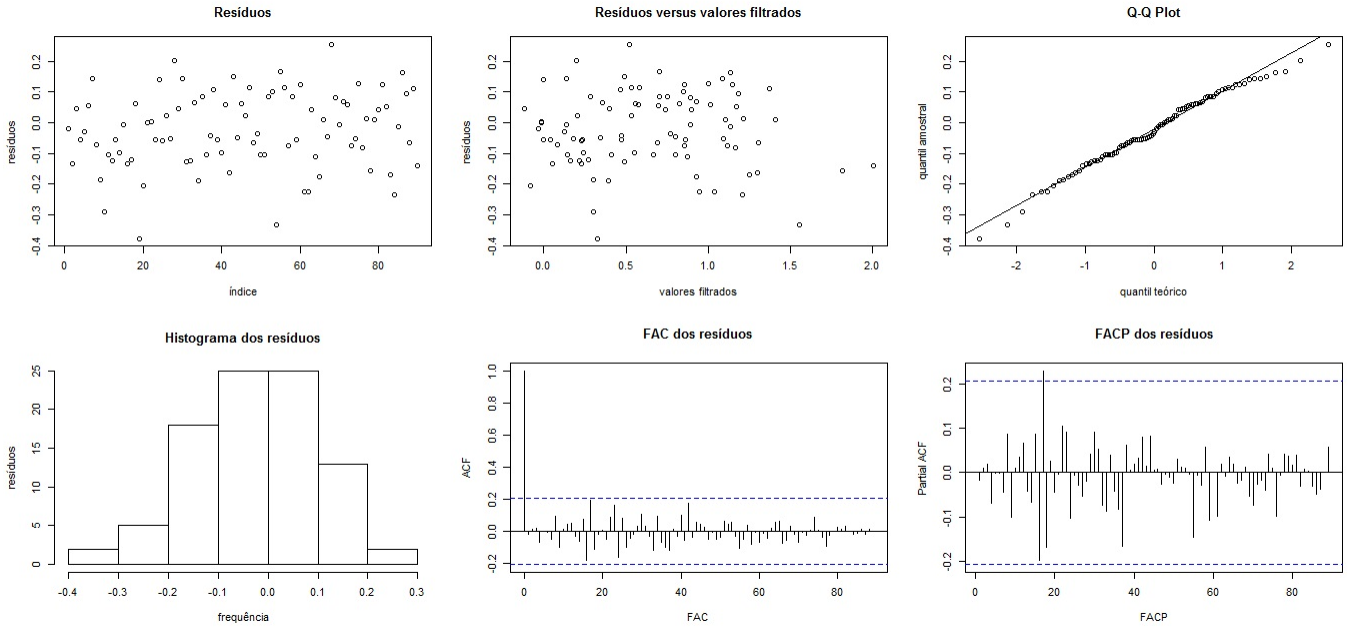

Figura 6.3: Análise de resíduos do modelo ajustado à série EAPC.

A seguir, são construídos os diferentes IC apresentados na Seção 3, que estão apresentados na Tabela 6.2.

Tabela 6.2: Intervalos de confiança para a série da log-receita arrecadada pelas EAPC.

θ	ICA		ICB		ICS	
σ^2_η	$-1\text{x}10^{-4}$	0,036	$1\text{x}10^{-6}$	0,218	0,009	0,018
σ^2_ξ	$-1\text{x}10^{-4}$	$1\text{x}10^{-4}$	$1\text{x}10^{-17}$	$1\text{x}10^{-4}$	$1\text{x}10^{-12}$	$1\text{x}10^{-4}$
σ^2_ω	$0,003$	$0,032$	$1\text{x}10^{-22}$	$0,025$	0,004	0,031
σ^2_ϵ	-0,021	0,021	$1\text{x}10^{-9}$	0,071	$1\text{x}10^{-6}$	0,019

Este exemplo deixa claro como valores muito pequenos para as estimativas dos hiperparâmetros são comuns, o que inviabiliza a modificação proposta no Capítulo 2 para o ICA. Desta forma, fica evidente a maneira com que o ICA gera intervalos fora do espaço paramétrico. Com relação ao ICB, este apresenta consistentemente limites inferiores muito

baixos. Pode-se observar também que os intervalos para o hiperparâmetro σ_ξ^2 apresentam limites inferior e superior muito próximos de zero, confirmando a suspeita de que esta componente de tendência pode ser não-estocástica.

6.2 Log-incidência de casos de dengue em Belo Horizonte

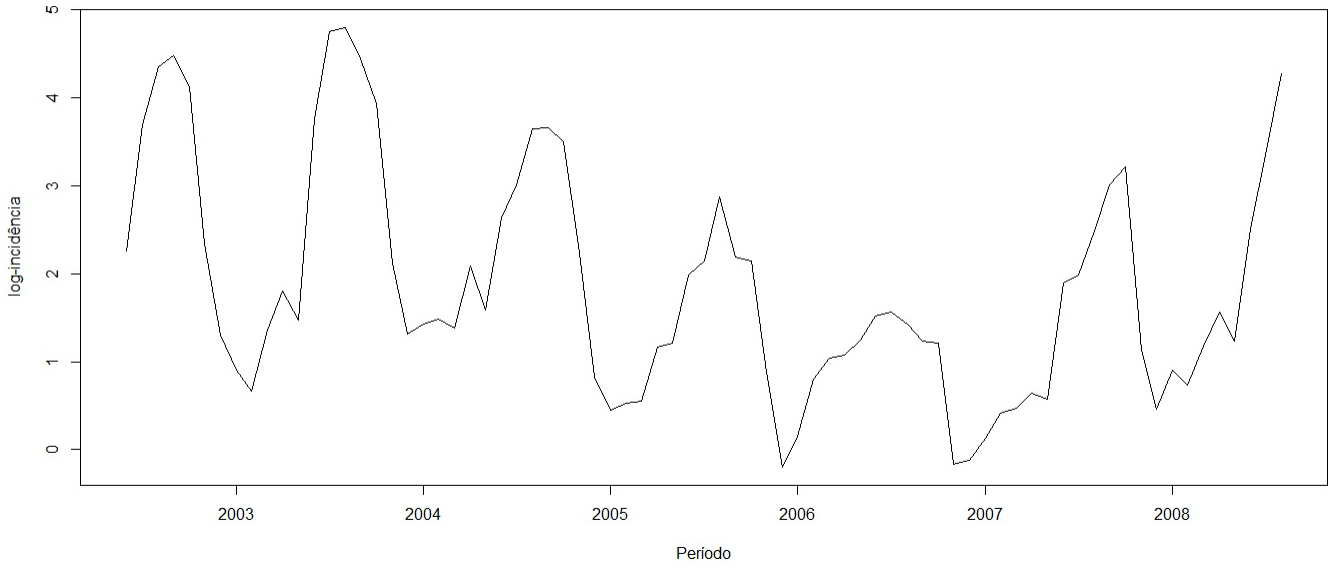

Figura 6.4: Série da log-incidência de casos de dengue em Belo Horizonte.

A série temporal em estudo é a do logaritmo natural da incidência de casos de dengue na cidade de Belo Horizonte. A incidência em um determinado instante de tempo é dada pelo número de casos notificados dividido pela população sob risco neste mesmo instante de tempo multiplicado por 100.000. Como a incidência é um número positivo, é tomado o logaritmo natural destes valores, para que a distribuição dos dados se aproxime mais de uma distribuição Normal.

Os dados são mensais no período de janeiro de 2002 a março de 2008, totalizando 75 observações. A Figura 6.4 mostra a série em questão versus o tempo. Nota-se uma clara sazonalidade na série, sugerindo um MEB, o que se deve ao fato de que a procriação do mosquito vetor da doença ocorre geralmente durante o verão. Essas informações foram cedidas pela Secretaria Municipal de Saúde de Belo Horizonte.

O histograma dos dados, apresentado na Figura 6.5 indica uma assimetria na distribuição. Contudo, será ajustado um modelo com erros Gaussianos, já que baseado nos resultados das simulações do Capítulo 5, os intervalos ICS apresentam boa cobertura também em situações

não-Gaussianas.

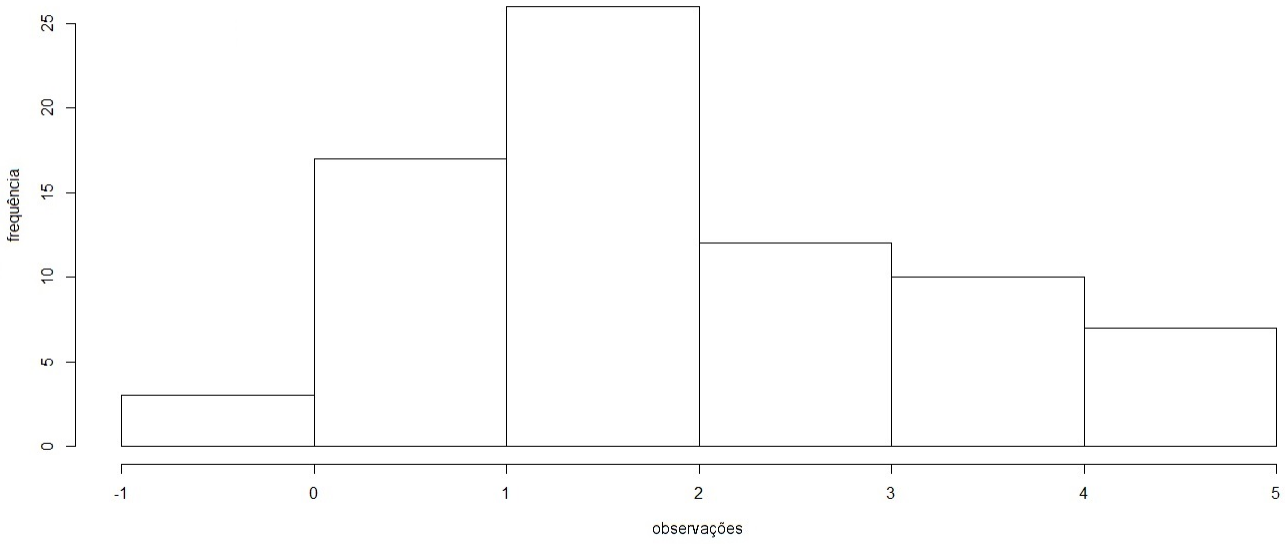

Figura 6.5: Histograma dos dados da Dengue.

A Figura 6.6 mostra o ajuste do MEB à série desconsiderando as 14 primeiras observações, que foram utilizadas para calibrar o FK. A Tabela 6.3 mostra o valor da estimativa de máxima verossimilhança para os hiperparâmetros, bem como a mediana, média e erro padrão das estimativas para as séries bootstrap. Novamente nota-se uma baixa magnitude dos hiperparâmetros de tendência e sazonalidade.

Tabela 6.3: Estimativa de máxima verossimilhança dos hiperparâmetros, mediana, média e erro padrão das estimativas dos hiperparâmetros das séries bootstrap

	σ^2_η	σ^2_ξ	σ^2_ω	σ^2_ϵ
EMV	$0,1643$	$1\text{x}10^{-6}$	$1\text{x}10^{-6}$	$1\text{x}10^{-5}$
mediana	$0,1508$	$1\text{x}10^{-8}$	$1\text{x}10^{-8}$	$1\text{x}10^{-5}$
média	$0,1479$	$0,0003$	$0,0003$	$0,0074$
erro padrão	$0,0374$	$0,0017$	$0,0011$	$0,0302$

A Figura 6.7 mostra os resíduos do ajuste com o MEB, bem como estes resíduos contra os valores ajustados, um qq-plot, um histograma e os gráficos da FAC e FACP destes resíduos. Destes gráficos pode-se verificar que os resíduos satisfazem as hipótese de independência, contudo não satisfazem a hipótese de normalidade, como esperado, e também não são homoscedasticos, o que pode ter ocorrido devido à distribuição dos dados ser não-gaussiana.

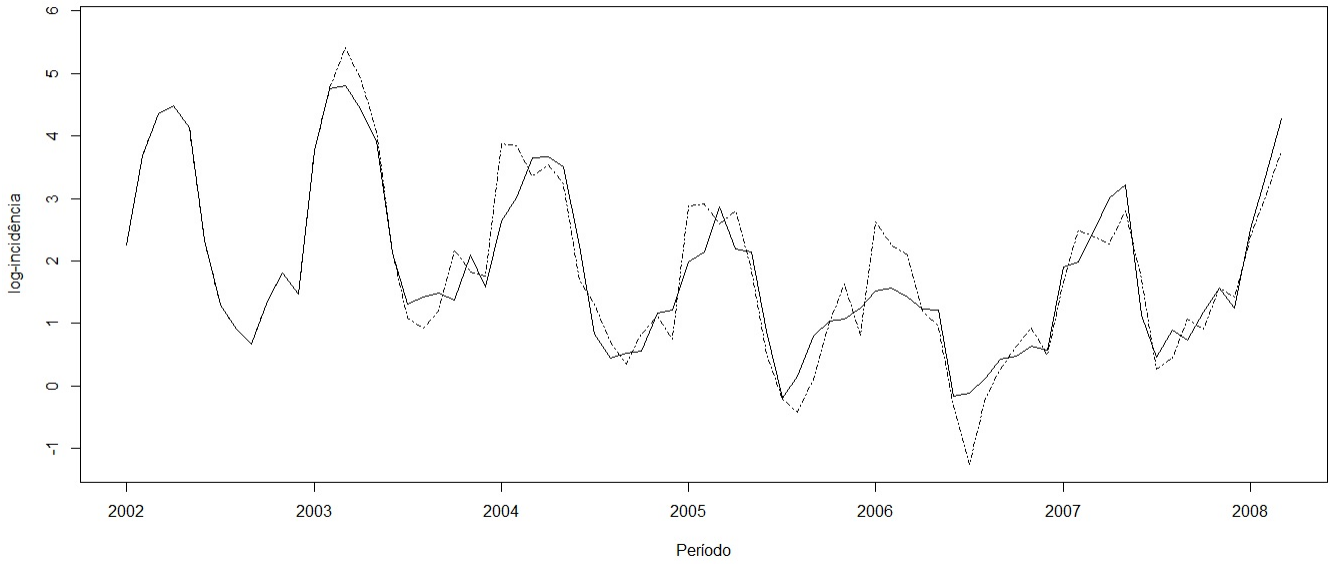

Figura 6.6: Gráfico do ajuste (linha tracejada) do MEB à série da Dengue (linha contínua).

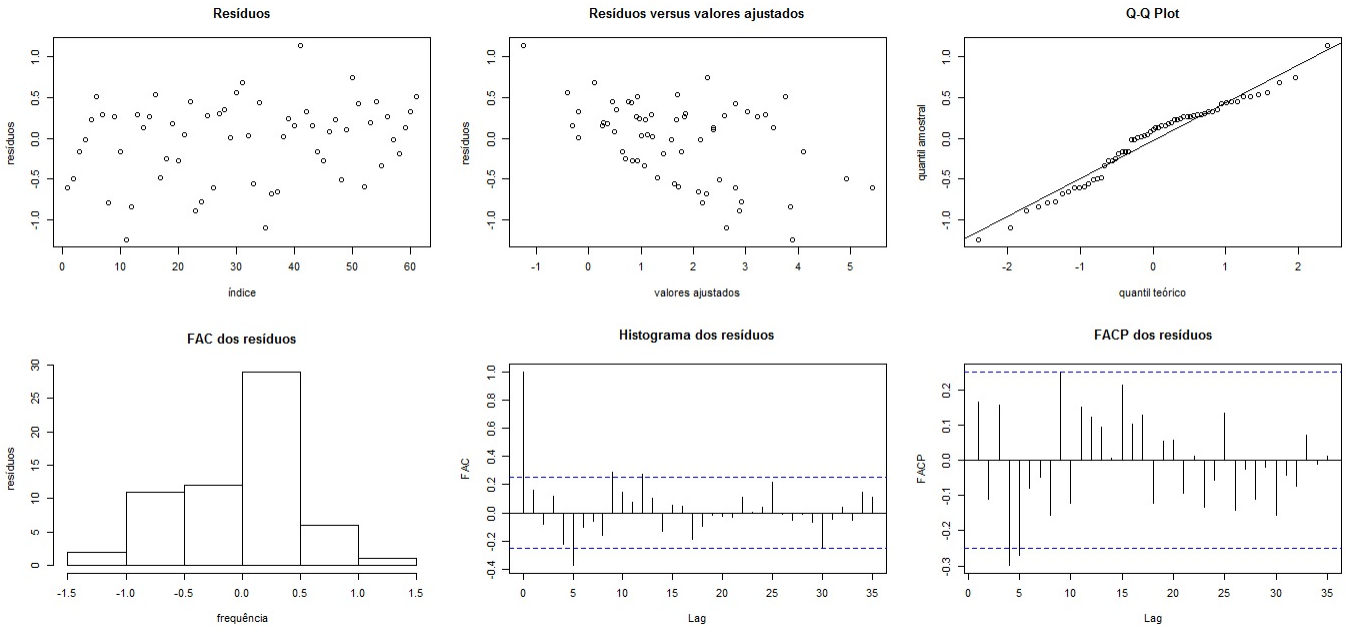

Figura 6.7: Análise de resíduos da série ajustada.

A seguir, são construídos os diferentes IC apresentados na Seção 3, que estão descritos na Tabela 6.4.

Nota-se a tendência observada na Tabela 5.8 do ICS de gerar intervalos com baixa ampli-

Tabela 6.4: Intervalos de confiança para a série da log-incidência dos casos de dengue em Belo Horizonte.

θ	ICA		ICB		ICS	
σ^2_η	0,0491	0,2794	0,0668	0,2159	0,1180	0,2390
σ^2_ξ	$-0,0016$	0,0016	$1\text{x}10^{-15}$	0,0003	$1\text{x}10^{-8}$	$0,0037$
σ^2_ω	$-0,0056$	0,0056	$1\text{x}10^{-21}$	0,0039	$1\text{x}10^{-8}$	0,0071
σ^2_ϵ	$-0,0532$	0,0534	$1\text{x}10^{-10}$	0,0453	$1\text{x}10^{-8}$	0,0311

tude. Pode-se observar também que os intervalos para os hiperparâmetros σ^2_ξ e σ^2_ω apresentam limites inferior e superior muito próximos de zero. Contudo, dadas suas magnitudes não há evidências para afirmar que as componentes de nível e tendência sejam determinísticas.

Capítulo 7

Conclusões

Este trabalho se propôs a apresentar um novo método para construção de intervalos de confiança para os hiperparâmetros dos modelos estruturais baseado na estatística deviance. Esperava-se desta nova metodologia um melhor tempo computacional, em relação ao método boostrap, além de não possuir problemas de fronteira, como o método assintótico. O trabalho apresenta também uma seção com detalhamento dos métodos computacionais empregados.

Afim de aferir a eficiência dos métodos quanto à cobertura dos intervalos e ao tempo computacional foram feitas simulações Monte Carlo. No caso de dados gaussianos, vê-se a grande eficência do ICS, tendo coberturas próximas do nível nominal fixado e um tempo computacional cerca de 10 vezes menor que o ICB. Ainda que com grandes amostras o ICA se comporte muito bem, com boas coberturas e um tempo mínimo, o ICS acaba sendo o método indicado para todos os modelos.

No caso não-gaussiano foram realizadas análises para as observações com distribuição Gama. O tempo computacional do ICS juntamente com coberturas próximas do nível nominal tornam este o método mais indicado para este tipo de dados. Não recomenda-se o uso do ICA para este tipo de dados.

Os métodos foram também utilizados em aplicações a dados reais. Neste caso notou-se os problemas do ICA e observou-se o bom comportamento do ICS, como previsto nas simulações. Notou-se também que o ICB é influenciado pela distribuição assimétrica dos hipeparâmetros.

Os resultados comparativos, apresentadas na Seção 5, de tempo computacional em conjunto com os de cobertura e amplitude reforçam a melhora na eficiência computacional que o método ICS apresenta frente ao método ICB, um método alternativo que contornava os problemas apresentados pelo ICA.

Referências Bibliográficas

[1] Barndorff-Nielsen, O. Inference on Full or Partial Parameters Based on the Standardized Signed Log Likelihood Ratio. *Biometrika*, vol.73, pp.307-322, 1986.

[2] Box, G. & Jenkins, G. *Time Series Analysis: Forecasting and Control.* New York: John Wiley & Sons, 1976.

[3] Broto, C. & Ruiz, E. Testing for Conditional Heteroscedasticity in the Components of Inflation. *Studies in Nonlinear Dynamics & Econometrics*, vol.13, 2009.

[4] Campos, F. *Algoritmos numéricos.* Belo Horizonte: LTC Editora, 2007.

[5] Cavanaugh, J. & Shumway, R. On computng the expected Fisher information matrix for state-space model parameters. *Statistics & Probability Letters*, vol.26, pp.347-355, 1996.

[6] Chen, J. & Jennrich, R.I. The Signed Root Deviance Profile and Confidence Intervals in Maximum Likelihood Analysis. *Journal of the American Statistical Association*, vol.91, pp.993-998, 1996.

[7] Cormen, T.H., Leiserson, C.E. & Rivest, R.L. *Introduction to Algorithms.* Cambridge: MIT Press and McGraw-Hill, 1990.

[8] Doornik, J.A. *Object-Oriented Matrix Programming using Ox 5th edition.* London: Timberlake Consultants Press, 2006.

[9] Dowell,M. & Jarrat,P. The Pegasus Method for computing the root of an equation. *BIT Numerical Mathematics*, vol.12, pp.503-508, 1972.

[10] Efron, B. Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, vol.7, pp. 1-26, 1979.

[11] Efron, B. & Tibshirani, R. *An Introduction to the Bootstrap.* London: Chapman & Hall, 1993.

[12] Evans, M.A., Kim, H.M. & OBrien, T.E. An Application of Profile-Likelihood Based Confidence Interval to Capture-Recapture Estimators. *Journal of Agricultural, Biological and Environmental Statistics*, vol. 1, pp. 131-140, 1996.

[13] Franco, G.C., Santos, T.R., Ribeiro, J.A. & Cruz, F.R.B Confidence Intervals for the Hyperparameters in Structural Models. *Communications in Statistics*, vol. 37, pp. 486-497, 2008.

[14] Franco, G.C. & Santos, T.R. Inference for the Hyperparameters of Structural Models Under Classical and Bayesian Perspectives: A Comparison Study. *Communications in Statistics - Simulation and Computation*, vol. 39, pp. 1671-1689, 2010.

[15] Gimenez, O.; Choquet, R.; Lamor, L.; Scofield, P.; Fletcher, D.; Lebreton, J. & Pradel, R. Efficient profile likelihood confidence intervals for capture-recapture model. *Journal of Agricultural, Biological and Enviromental Statistic*, vol. 10, pp. 184-196, 2005.

[16] Harvey, A.C. *Forecasting, Structural Time Series Models and the Kalman Filter.* Cambridge: University Press, 1989.

[17] Harvey, A.C. *The Econometric Analysis of Time Series.* New Jersey: Prentice-Hall, 1990.

[18] Harvey, A.C. & Todd, P.H.J. Forecasting economic time series with structural and Box-Jenkins models. *Journal of Business and Economic Statistics*, vol.1, pp.299-315, 1983.

[19] Heinze,G. & Chemper,M. A solution to the problem of separation in logistic regression. *Statistics in Medicine*, vol.21, pp.2409-2419, 2002.

[20] Holt, C.C. Forecasting seasonals and trends by exponentially weighted moving averages. *ONR Research Memorandum*, vol.1, pp.52, 1957.

[21] Kalman, R.E. A new approach to linear filltering and prediction problems. *Trans. ASME J. Basic Eng.*, vol.82, pp.35-45, 1960.

[22] Koopman, S.; Shephard, N. & Doornik, J. Statistical algorithms for models in state space form using SsfPack 2.2. *Econometrics Journal*, vol.2, pp.107-160, 1999.

[23] Neale, M.C. & Miller, M.B. The Use of Likelihood Based Confidence Intervals in Genetic Models. *Behavior Genetics*, vol.27, pp. 113-120, 1997.

[24] Pfanzagl, J. *Parametric statistical theory with the assistance of R.* Berlin: Walter de Gruyter, 1994.

[25] Pfefferman, D. & Tiller, R. Bootstrap approximation to predictions MSE for State-Space models with estimated parameters. *Journal of Time Series Analysis*, vol.26, pp.893-916, 2005.

[26] Quenneville, B. & Singh, A. Bayesian Prediction Mean Squared Error for State Space Models with Estimated Parameters. *Journal of Time Series Analysis*, vol.21, pp.219-236, 2000.

[27] Rodriguez, A. & Ruiz, E. Bootstrap Prediction Intervals in State Space Models. *Journal of the Time Series Analysis*, vol. 30, pp. 167-178, 2009.

[28] Rodriguez, A. & Ruiz, E. Bootstrap predictions Mean Squared Erros of unobserved states based on the Kalman Filter with estimated parameters. *Computational Statistics and Data Analysis*, vol. 56, pp. 62-74, 2012.

[29] Rolke, W.A; López, A.M. & Conrad, J. Limits and confidence intervals in the presence of nuisance parameters. *Nucl.Instrum. Methods Phys. Res.A.*, pp. 493-551, 2005.

[30] Shanno, D.F. Conditioning of quasi-Newton methods for function minimization *Mathematics of Computation*, vol.24, pp.647-656, 1970

[31] Soetaert, K. rootSolve: Nonlinear root finding, equilibrium and steady-state analysis of ordinary differential equations *R package version 1.4.*, 2009

[32] Stoffer, D.S. & Wall, K.D. Bootstrapping state-space models: Gaussian maximum likelihood estimation and the Kalman Filter *Journal of the American Statistical Association*, vol.86, pp.1024-1033, 1991

[33] Wilks, S.S. The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *The Annals of Mathematical Statistics*, vol.9, pp.60, 1938.

[34] West, M. & Harrison, J. *Bayesian Forecasting and Dynamic Models.* New York: Springer, 1997.

[35] Winters, P.R. Forecasting sales by exponentially weighted moving averages. *Management Science*, vol. 6, pp. 324-342, 1960.