Universidade Federal do Rio Grande do Sul
Instituto de Matemática e Estatística
Departamento de Estatística

Anais

IX SEMANÍSTICA

IX Semana Acadêmica do Departamento de Estatística

da UFRGS

http://www.ufrgs.br/semanistica

Porto Alegre - 15, 16 e 17 de outubro de 2018

Organização:

Promoção:

Conteúdo

1 Cartaz da IX SEMANÍSTICA

2 Cronograma da IX SEMANÍSTICA

3 Introdução

A IX Semana Acadêmica da Estatística (SEMANÍSTICA) será realizada nos dias 15, 16 e 17 de outubro de 2018, no Instituto de Matemática e Estatística - IME, Campus do Vale da UFRGS, Porto Alegre, RS. O evento engloba os mais variados temas dentro da área acadêmica e profissional.

O objetivo principal da SEMANÍSTICA é promover o desenvolvimento, aprimoramento e a divulgação da Estatística, entre diferentes perspectivas, acadêmica e/ou prática no campo de aplicação. A proposta da IX SEMANÍSTICA é incentivar a integração entre estudantes, professores e profissionais de diversas áreas que utilizam a Estatística como suporte de decisão em suas respectivas áreas de conhecimento.

Como objetivos específicos da SEMANÍSTICA, podem-se citar: divulgar as contribuições recentes dos pesquisadores participantes promovendo-se o intercâmbio entre cientistas, alunos e profissionais aplicados; promover um maior contato entre pesquisadores do Departamento de Estatística da UFRGS e pesquisadores de outros departamentos, propiciando futuros trabalhos de pesquisa conjuntos; intensificar o contato e o intercâmbio científico entre profissionais da Região Sul e a iniciativa privada dentro das realidades do Estado do Rio Grande do Sul e do MERCOSUL; divulgar os diferentes métodos e aplicações de Estatística para discentes da graduação em Estatística, bem como discentes de pós-graduação e graduação das mais diversas áreas correlatas, tais como: Economia, Administração, Engenharia e Biomédicas.

Para maiores informações sobre a IX SEMANÍSTICA (Semana Acadêmica da Estatística 2018) podem ser encontradas no site www.ufrgs.br/semanistica.

4 Agradecimentos

A IX SEMANÍSTICA - Semana Acadêmica do Departamento de Estatística da UFRGS não teria sido possível sem o apoio das seguintes agências financiadoras e instituições:

- DEST-UFRGS - Departamento de Estatística da UFRGS
- IME-UFRGS - Instituto de Matemática e Estatística da UFRGS
- PROPESQ-UFRGS - Pró-Reitoria de Pesquisa da UFRGS
- UFRGS - Universidade Federal do Rio Grande do Sul

A Comissão Organizadora da IX SEMANÍSTICA agradece a colaboração de todos que se dedicaram anonimamente e sem interesses pessoais, em promover a integração entre alunos, professores e profissionais em estatística.

Comissão Organizadora

5 Comissão Organizadora Docente

- Danilo Marcondes Filho (Departamento de Estatística-UFRGS)
- Márcio Valk (Departamento de Estatística-UFRGS)
- Guilherme Pumi (Departamento de Estatística-UFRGS)
- Liane Werner (Departamento de Estatística-UFRGS)
- Gabriela Cybis (Departamento de Estatística-UFRGS)
- Márcia Elisa Soares Echeveste (Departamento de Estatística-UFRGS)
- Cleber Bisognin (Departamento de Estatística-UFSM)

6 Comissão Científica

- Márcio Valk (Departamento de Estatística-UFRGS)
- Gabriela Cybis (Departamento de Estatística- UFRGS)
- Danilo Marcondes Filho (Departamento de Estatística-UFRGS)
- Guilherme Pumi (Departamento de Estatística-UFRGS)
- Liane Werner (Departamento de Estatística-UFRGS)

7 Comissão Organizadora Discente

- Gabriel da Cunha (Bacharel em Estatística - UFRGS)
- Júlia Bürgel Borsato (Curso de Estatística - UFRGS)
- Juliana Souza (Curso de Estatística - UFRGS)
- Maicon Fridrich Gottselig (Curso de Estatística - UFRGS)
- Martha Reichel (Curso de Estatística - UFRGS)
- Pietá Ribeiro (Curso de Estatística - UFRGS)
- Roger Moreira (Curso de Estatística - UFRGS)
- Gabriel Fagundes (Curso de Estatística - UFRGS)

8 Apresentação

A programação da IX SEMANÍSTICA - Semana Acadêmica do Departamento de Estatística da Universidade Federal do Rio Grande do Sul englobou as seguintes atividades:

- Duas conferências envolvendo uma professora pesquisadora do DEST(Departamento de Estatística) e uma professora aposentada.
- 3 Minicursos envolvendo manipulação e visualização de dados e edição de textos, sendo dois deles ministrados por professores do curso de Bacharelado em Estatística da Universidade Federal do Rio grande do Sul e um deles ministrado por um mestrando em Ciência da Computação do PPGC - UFRGS.
- Comunicações orais apresentadas pelos participantes do evento;

9 Programação

Conferências:

(M1) Minicurso 1 - Prof. Dr. Rodrigo Citton e Prof. Dr. Markus Stein - Professores do Departamento de Estatística - UFRGS

Título: Pintando de Bordando no R: ggplot2 e Rmarkdown

(M2) Minicurso 2 - Taiane Prass - Professora do Departamento de Estatística - UFRGS

Título: Introdução ao LaTeX

(M3) Minicurso 3 - Kazuki Monteiro Yokoyama - Mestrando no Programa de Pós-Graduação em Ciência da Computação - PPGC/ UFRGS

(C1) Conferência 1 - Profª. Dinara Fernandez - Professora do Departamento de Estatística - UFRGS

Título: AAA: Três dimensões do Bacharelado em Estatística da UFRGS

(C2) Conferência 2 – Profª. Drª. Gabriela Cybis - Professora do Departamento de Estatística - UFRGS

Título: Integrando diferentes tipos de dados para caracterizar a diversidade do vírus da gripe

10 Minicursos

Pintando e bordando no R: ggplot2 e Rmarkdown

Prof. Dr. Rodrigo Citton e Prof. Dr. Markus Stein
Professores do Departamento de Estatística - UFRGS

Resumo

É fundamental para todo profissional ligado à análise de dados a boa comunicação dos resultados. A máxima "uma imagem vale mais que mil palavras"se aplica mais uma vez neste contexto. A geração de gráficos deve ser tarefa rotineira de estatísticos e analistas de dados e a alta qualidade deve ser perseguida. O pacote ggplot2 do R atinge este objetivo sem um alto custo de programação. Outra ferramenta que vem ganhando importância para a boa comunicação estatística é o R Markdown. Este pacote do R integra funcionalidades de edição de texto e análise de dados para a geração de relatórios dinâmicos nos mais diversos formatos: HTML, LaTeX, PDF, WORD, SLIDES, entre outros. Neste breve tutorial apresentaremos de forma simples as principais funções destes dois pacotes que irão lhe possibilitar o compartilhamento de suas análises com um público mais amplo.

Introdução ao LaTeX

Taiane Prass
Professora do Departamento de Estatística - UFRGS

Resumo

LaTeX é uma implementação da linguagem TeX, criada em 1978, amplamente utilizada na edição de textos científicos. Diferentemente de editores como Word, o LaTeX não apresenta uma interface amigável, o que deixa muitas pessoas pensando: "Por que eu deveria abandonar algo simples e adotar algo mais complexo?" Neste minicurso discutiremos as vantagens e desvantagens do LaTeX na elaboração de documentos e apresentações. Apresentaremos noções básicas de configuração de estilos (artigo, relatório, livro), uso de pacotes, formatação de páginas, tabelas, figuras e equações matemáticas.

Introdução ao Python

Kazuki Monteiro Yokoyama
Mestrando no Programa de Pós-Graduação em Ciência da Computação - PPGC/UFRGS

Resumo

A linguagem de programação Python tem destacado-se nas comunidades da estatística e machine learning por sua versatilidade, produtividade e amplo ecossistema de bibliotecas. Esse curso apresentará os conceitos básicos da linguagem e como ela pode ser utilizada para resolver problemas com dados.

11 Conferências

Conferência 1

AAA: Três dimensões do Bacharelado em Estatística da UFRGS

Profª. Dinara Fernandez
Professora do Departamento de Estatística - UFRGS

Resumo

A gênese do curso de Bacharelado em Estatística da UFRGS, sua trajetória até hoje e um olhar relativamente as ameaças e oportunidades futuras.

Conferência 2

Integrando diferentes tipos de dados para caracterizar a diversidade do vírus da gripe

Profª. Drª. Gabriela Cybis
Professora do Departamento de Estatística - UFRGS

Resumo

O vírus da gripe infecta anualmente de 10 a 20% da população mundial e traz custos econômicos e de saúde pública significativos. A vacinação é uma das nossas principais ferramentas de controle para o vírus. Entretanto, devido à rápida evolução do vírus, a vacina deve ser atualizada todo ano para proteger contra as novas variantes do vírus que estarão circulando na próxima temporada de gripe. Assim, ampliar nosso conhecimento dos processos de evolução genética e imunogênica do vírus é fundamental para entendimento do comportamento futuro da gripe, o que pode levar a melhorias no design da vacina. Nesse contexto, considerarei o potencial da integração desses dados para realizar previsões epidemiológicas.

12 Comunicações Orais

Comunicação Oral 1:

Comparação de Modelos de Regressão Para Dados de Contagem Inflacionados de Zeros por Meio de Simulações

Maicon Michael Fridrich Gottselig, Juliana Sena de Souza, Silvana Schneider

Resumo: Modelos inflacionados de zeros são ferramentas importantes no desarme de dados não identicamente distribuídos provenientes da mistura de duas populações com processos distintos. Esta classe de modelos é evidenciada por Diane Lambert (1992) que postula uma família de modelos de mistura que permite a modelagem de dados com excesso de zeros, lidando com a sobredispersão decorrente desta característica. Posto isso, este trabalho tem como foco executar por meio de simulações computacionais uma comparação de modelos de contagem sob ótica de excesso de zeros. Os seguintes modelos: ZIP (Zero-Inflated Poisson), ZIG (Zero-Inflated Geometric), ZIB (Zero-Inflated Binomial), ZINB (Zero- Inflated Negative Binomial), ZIPIG (Zero-Inflated Poisson Inverse Gaussian), ZIBB (Zero-Inflated Beta Binomial), ZIBNB (Zero-Inflated Beta Negative Binomial), ZICMP (Zero-Inflated Conway-Maxwell Poisson) e ZIDelaporte (Zero-Inflared Delaporte); São utilizados como base para simulações e ajustes cruzados afim de avaliar e testar adaptilidade de cada modelo a diferentes cenários de sobredispersão e inflação de zeros. Notou-se que modelos os modelos relativamente novos ZID e ZICMP performam muito bem e se posicionam paralelamente aos modelos ZIPIG e ZINB. Negativamente destacam-se os modelos ZIBNB, ZIB, ZIBB e ZIG que não obtiveram estimativas satisfatórias

Comunicação Oral 2:

Inferência Estatística para Classificação de Sinais Cardíacos

Mikaela Baldasso, Marcio Valk

Resumo: Doenças cardiovasculares são responsáveis por milhões de mortes anualmente, segundo a Organização Mundial da Saúde e, dado isso, várias são as iniciativas, em todo o mundo, que visam estimular o desenvolvimento de novas técnicas que permitam diagnosticar e prevenir essas enfermidades. Diferentes técnicas de diagnósticos são utilizadas para detectar e prevenir esses desfechos, em que busca-se, principalmente, utilizar métodos não invasivos, baratos e que resultem em respostas rápidas e confiáveis, como por exemplo, aqueles baseados em Eletrocardiogramas e Fonocardiogramas. A partir disso, nosso objetivo nesse trabalho é utilizar a estatística para fazer inferência sobre classificação, ou seja, mensurar a confiabilidade de uma técnica de diagnóstico, em particular testar o método baseado em U-estatística para classificação e agrupamento de dados.

Comunicação Oral 3:

Um novo modelo probabilístico para dados restritos ao intervalo unitário.

Tatiane Fontana Ribeiro, Renata Rojas Guerra, Fernando Arturo Peña-Ramírez, Pierre Louis Termidor

Resumo: São inúmeras as situações nas quais o objeto de estudo consiste em variáveis com suporte no intervalo unitário. Dentre as quais citam-se: taxas, proporções e índices. Embora possa ser utilizado, nesses casos, o modelo clássico: distribuição beta e outros já existentes na literatura, é importante dispor de outros modelos probabilísticos alternativos. Neste contexto, objetiva-se propor uma nova distribuição de probabilidade unitária, bem como estudar algumas de suas características estatísticas e matemáticas e estimar seus parâmetros via máxima verossimilhança. Para tanto, propõe-se uma transformação em uma dada variável aleatória que limita a imagem da nova variável obtida ao intervalo (0; 1). Foi avaliado o desempenho dos estimadores de máxima verossimilhança em amostras de tamanho finito através de simulações de Monte Carlo. Obtiveram-se resultados razoáveis em termos de acurácia e precisão das estimativas, mesmo para amostras de tamanho 20.

Comunicação Oral 4:

Estudo simulado envolvendo Cartas de Controle Multivariadas.

Eduardo de Oliveira Correa, Danilo Marcondes Filho

Resumo: Processos industriais geram dados acerca de inúmeras variáveis de interesse correlacionadas. Buscando um monitoramento mais robusto de tais processos,cartas de controle baseados em técnicas estatísticas multivariadas foram desenvolvidos. Destacam-se as cartas de controle Qui-Quadrado (χ^2) e da Variância Generalizada (W). Estas estatísticas permitem um monitoramento simultâneo do vetor de médias e da matriz de covariâncias das variáveis, respectivamente, a cada nova amostra do processo. Este trabalho apresenta um estudo por simulação para investigar o poder de detecção das cartas χ^2 e W. A partir de um processo simulado com 4 variáveis e uma estrutura de covariância, descontroles são impostos tanto no vetor de médias quanto na matriz de covariâncias do processo sob controle. Os resultados mostram que a sensibilidade da carta W aumenta para a detecção de modificações maiores na estrutura de covariância original das variáveis. Já em relação à carta χ^2, podemos notar que alterações no vetor de médias nas direções comuns de variância das variáveis (isto é, na direção das suas covariâncias) são detectadas com menos sensibilidade em relação às alterações que não estão nas suas direções de covariância.

Comunicação Oral 5:

Estudo de Simulações na Estimação de Parâmetros dos Processos k-Factor GARMA$(p;u;\alpha;q)_{S\alpha S}$

Cleber Bisognin, Sílvia R.C. Lopes, Leticia Menegotto

Resumo: Neste trabalho estamos interessados em estudar séries temporais com as características de longa dependência, sazonalidade e alta variabilidade. Os processos k-Factor GARMA $(p;u;\alpha;q)$ com inovações α-estáveis simétricas, denotados por k-Factor GARMA $(p;u;\alpha;q)_{S\alpha S}$, nos permitem trabalhar com tais séries temporais. Séries de agregados monetários e rendimentos financeiros são exemplos para aplicações destes processos. O principal objetivo é verificar as condições de estacionariedade, invertibilidade e propor estimadores para os parâmetros destes processos. Para tanto, estendemos o estimador para os processos SARFIMA$(p;d;q)$ $\times$(P;D;Q)$_{sS\alpha S}$, proposto por Ndongo et al. [2010], para os processos k-Factor GARMA $(p;u;\alpha;q)_{S\alpha S}$. Neste estimador utilizamos as funções periodograma normalizado suavizado e periodograma suavizado de correlações como estimadores da função poder de transferência [Stein, 2012]. Foram realizadas simulações de Monte Carlo para verificar a acurácia das estimativas dos parâmetros e para tal foram analisados o vício, o erro quadrático médio (EQM) e a variância (Var) das estimativas. Constatamos que ambos os estimadores propostos, apresentaram boas estimativas, no sentido de baixos vício, erro quadrático médio e variância para todos os parâmetros na maioria dos casos analisados. Verificou-se também que quanto menor o valor do $0 < \alpha < 2$ (parâmetro relacionado a variabilidade dos dados, quanto menor α maior a variabilidade da série temporal) menor é a acurácia das estimativas para o parâmetro λ do processo.

Comunicação Oral 6:

Estudo da Sensibilidade do Bayes Factor para seleção de modelos

Lauren Alves Vieira; Gabriela Bettella Cybis

Resumo: Métodos bayesianos filogenéticos são uma ferramenta central na biologia evolutiva. Dentre estes o Modelo de Variável Latente estima correlações entre características fenotípicas (contínuas e categóricas ordinais ou nominais), controlando para história evolutiva entre os indivíduos amostrados. Nas aplicações deste modelo é comum a escolha de prioris pouco informativas, geralmente adotando a distribuição conjugada Wishart Inversa para matriz de covariâncias do modelo. Nossos resultados prévios evidenciaram uma possível sensibilidade do método de seleção de modelos quanto a escolha da priori, de modo que modelos com maior número de graus de liberdade (gl), pareciam ser favorecidos. Com o intuito de avaliar esse efeito da priori sobre a seleção do modelo, foi conduzido o estudo apresentado abaixo.

Comparação de Modelos de Regressão Para Dados de Contagem Inflacionados de Zeros Por Meio de Simulações

Maicon Michael Fridrich Gottselig [1]

Juliana Sena de Souza[2]

Silvana Schneider[3]

Resumo: Modelos inflacionados de zeros são ferramentas importantes no desarme de dados não identicamente distribuídos provenientes da mistura de duas populações com processos distintos. Esta classe de modelos é evidenciada por Diane Lambert (1992) que postula uma família de modelos de mistura que permite a modelagem de dados com excesso de zeros, lidando com a sobredispersão decorrente desta característica. Posto isso, este trabalho tem como foco executar por meio de simulações computacionais uma comparação de modelos de contagem sob ótica de excesso de zeros. Os seguintes modelos: ZIP (Zero-Inflated Poisson), ZIG (Zero-Inflated Geometric), ZIB (Zero-Inflated Binomial), ZINB (Zero-Inflated Negative Binomial), ZIPIG (Zero-Inflated Poisson Inverse Gaussian), ZIBB (Zero-Inflated Beta Binomial), ZIBNB (Zero-Inflated Beta Negative Binomial), ZICMP (Zero-Inflated Conway-Maxwell Poisson) e ZIDelaporte (Zero-Inflared Delaporte); São utilizados como base para simulações e ajustes cruzados afim de avaliar e testar adaptilidade de cada modelo a diferentes cenários de sobredispersão e inflação de zeros. Notou-se que modelos os modelos relativamente novos ZID e ZICMP performam muito bem e se posicionam paralelamente aos modelos ZIPIG e ZINB. Negativamente destacam-se os modelos ZIBNB, ZIB, ZIBB e ZIG que não obtiveram estimativas satisfatórias.

Palavras-chave:*Modelos de contagem*, *Inflação de zeros*, *sobresipersão*, *Comparação*, *Simulação*

1 Introdução

Frank A. Haight (1967) explica que dados de contagem são definidos como o número de sucessos de experimentos realizado num período finito. Quando existe o intuito de se modelar variáveis de contagem, afim de se inferir acerca da relação desta esperança condicionada à variáveis explicativas, é necessária a suposição de distribuições discretas sobre a variável dependende, como exemplos bastante explorados menciona-se Poisson, Binomial e Geométrica. Como tais distribuição pertencem a família exponencial de distribuições toda a construção teórica exposta em Nelder e Wedderburn (1972) estende-se de forma natural.

[1]UFRGS - Universidade Federal do Rio Grande do Sul. Email: *maiconmfg@gmail.com*
[2]UFRGS - Universidade Federal do Rio Grande do Sul. Email: *julianass.estatistica@gmail.com*
[3]UFRGS - Universidade Federal do Rio Grande do Sul. Email: *sschneider@ufrgs.br*

Na maioria dos estudos entretanto, surge o fenômeno da sobredispersão, que é caracterizada como uma variabilidade superior a qual o modelo de contagem empregado é capaz de incorporar. No caso da distribuição de Poisson que impõe equidispersão, quando é registrado $\mathbb{E}(\bar{Y}) \neq \mathbb{V}A\bar{R}(Y)$ há indícios que colocam em cheque a Regressão de Poisson. A direção desde desbalanço entre esperança e variância caracteriza sub ou sobredispersão e tem como justificativa uma grande gama de justificativas: caudas pesadas, assimetria, excesso de zeros, entre outros.

O eixo principal deste trabalho é verificar a adaptilidade de modelos de contagem à sobredispersão e excesso de zeros. Segundo proposição de Lambert (1992) que sugere mistura de distribuições de contagem com distribuição de Bernoulli afim de captação de efeitos associados ao processo de zeros. É importante ressaltar a existência de outras alternativas para ajuste de dados inflacionados de zeros, como modelos Hurdle de Ridout (1998) e modelos de zeros alterados de Heilbron (1989).

2 Modelos Inflacionados de Zero

Em seu artigo, D. Lambert (1992) discorre acerca de dados provenientes da amostragem de um conjunto de duas populações com processos distintos. Uma população contendo apenas indivíduos com valor zero e outra população cujos indivíduos se adequam a alguma distribução de contagem.

Desta maneira assumindo $Y = (y_1, y_2, ..., y_n)$ como uma amostra aleatória independente do processo acima descrito tem-se: $P(y_i \in \text{Sempre Zero}) = \pi$ e $P(y_i \notin \text{Sempre Zero}) = 1 - \pi$, o que compila em:

$$P(Y_{ZI} = y|\theta, \pi) = \begin{cases} \pi + (1-\pi)f_y(y=0|\theta), & \text{y=0.,} \\ (1-\pi)f_y(y|\theta), & \text{y > 0 .} \\ 0, & \text{caso contrário} \end{cases} \tag{1}$$

onde f_y denota a distribuição de probabilidade indexada pelo parâmetro θ do processo de contagem e π assume posição de parâmetro que define a probabilidade da contagem de zero decorrente dos indivídios da população que apenas fornece contagem zero. Lambert percebeu que sob condições ideais a contagem de falhas de soldas eram sempre zero, e quando fora de controle, o processo observada falhas que se adequavam a distribuição de Poisson. Assim, propôs assumir que os processos sob controle e fora de controle eram na verdade populações distintas.

A formulação (1) caracteriza a família de distribuições infladas de zeros e, frente à diferentes f_y, novas propriedades são observadas e diferentes fontes de sobredispersão são captadas conforme mostrado por Paula (2004). Inicialmente é necessário verificar que

$$\mathbb{E}(Y_{ZI}) = (1-\pi)\mathbb{E}(Y) \qquad \text{e} \qquad \mathbb{V}AR(Y_{ZI}) = (1-\pi)(\mathbb{V}AR(Y) + \pi\mathbb{E}(Y)^2).$$

Por meio do índice de sobredispersão proposto por Cox e Lewis (1966), e denotado como $OI(Y)$ *(overdispersion index)*, tem a fórmula denotada por $OI(Y) = \mathbb{V}AR(Y)/\mathbb{E}(Y)$. É possível verificar que a proposição de Lambert com a insersão do parâmetro π de fato há absorção de sobredispersão de ordem $\pi\mathbb{E}(Y)$ uma vez que $OI(Y_{ZI}) = OI(Y) + \pi\mathbb{E}(Y)$.

Lambert (1992) ainda propõe a modelagem via covariáveis da proporção π de zeros estruturais e da média μ do processo. Explica também que ambos os parâmetros podem ou não ser modelados pelo mesmo conjunto de covariáveis, o que os torna ou não relacionados, conforme examinado por Daniel B. Hall (2000). Como marginalmente $\mathbb{E}(Y_i) = (1 - p_i)\mu_i$ pode haver confundimento nas estimativas dos coeficientes dos dois processos, o que atribui maior variabilidade aos coeficientes associados ao processo logístico bem como acrésimo de erros padrões.

As estimativas dos coeficientes e demais parâmetros são obtidas pela maximização da verossimilhança por meio do emprego de um método computacional recursivo. Lambert (1992) demonstra como se dá a construção do algorítmo EM, o que requer o cálculo de esperanças condicionais que podem ser complexas, por isso geralmente utiliza-se o método de Ficher Scoring que é um algortimo de *hill climbing*, conforme explicitado por Sampson (1976).

A proposição de Lambert é flexível e se estende a diversas $f_y(y)$, que acaba por incorporar ao modelo inflado de zeros seus momentos e permite melhor adequamento à diferentes perfis de dados, captando sobredisperão e modelando excesso de zeros. Outra alternativa para incorporar maior sobredispersão ao modelo por meio da inclusão de novos parâmetros é via suposição de variáveis latentes $Y|W \sim P(\lambda)$ que frente a diferentes W, confere novos parâmetros e complexidade à Y . Há também a posibilidade de se assumir Y como sendo resultado de alguma função de variáveis aleatórias do tipo $Y = W + Z$, com W e Z variáveis aleatórias. Estas táticas corroboram com a construção de modelos mais fléxiveis. A Tabela abaixo traz os modelos selecionados e expõe suas construções, bem como a paralela distribuição inflacionada de zeros e o índice de sobredispersão.

Tabela 1: Tabela resumo das distribuições infladas de zeros

Descrição	**Distribuição ($\underline{\theta}$)**	**Distribuição ZI ($\underline{\theta}$)**	**OI(Y_{ZI})**
-	P(λ)	ZIP (λ, π)	$1 + \pi\lambda$
$Y\|W \sim P(\lambda), W \sim G(\alpha, \beta))$	NB(α, β)	ZINB (α, β, π)	$\frac{(\alpha+\beta+\alpha\beta\pi)}{\beta}$
-	G(p)	ZIG(p, π)	$\frac{1+p\pi}{p}$
-	Bin(k, p)	ZIB(k, p, π)	$1 - p + kp\pi$
$Y\|W \sim P(\lambda), W \sim IG(\mu, \sigma))$	PIG(μ, σ)	ZIPIG(μ, σ, π)	$e^{\mu+\sigma^2/2}[e^{\sigma^2} - 1 + \pi]$
$Y\|p \sim Bin(n, p)$,$p \sim Beta(\alpha, \beta))$	BB(n, α, β)	ZIBB(n, α, β, π)	$\frac{n\beta}{\alpha+\beta} + \pi\frac{n\alpha}{\alpha+\beta}$
$Y\|p \sim NB(r, p)$, $p \sim Beta(\alpha, \beta))$	BNB(r, α, β)	ZIBNB(r, α, β, π)	$\frac{r\beta}{\alpha-1}\left[\frac{(r+\alpha-1)(\alpha+\beta-1)}{r\beta(\alpha-2)} - \pi^2\right]$
-	CMP(λ, v)	ZICMP(λ, v, π)	$\frac{\frac{\lambda^{1/v}}{v} + \pi(\lambda^{1/v} + \frac{1-v}{2v})^2}{\lambda^{1/v} + \frac{1-v}{2v}}$
Convolução entre NB(α, β) e P(λ)	Delaporte(λ, α, β)	ZIDelaporte($\lambda, \alpha, \beta, \pi$)	$\frac{\lambda+\alpha\beta(1+\beta)+\pi(\lambda+\alpha\beta)^2}{\lambda+\alpha\beta}$

3 Metodologia e Simulações

Com a premissa de comparar a capacidade de absorção de sobredispersão dos modelos apresentados na Tabela 1 e verificar o ajuste destes frente a dados com excesso de zeros foram realizadas simulações computacionais de dados de regressão com as distribuições alvo via software R (versão 3.4.1) com auxílio dos pacotes VGAM, gamlss.dist COMPoissonReg, pscl, Delaporte e gamlss.

Foram gerados 1000 bancos de dados de cada um dos k=9 modelos abordados, cada qual com n=500. Tomando $\beta = [1, 0.5, -0.5]'$ e $\gamma = [-2, 1, -2]'$ como coeficientes regressores, além da relação $log(\mu_i) = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2}$ e $log(\frac{\pi}{1-\pi}) = \gamma_0 + \gamma_1 X_{i,2} + \gamma_2 X_{i,3}$ sendo que $x_{i,1} \sim N(3.5, 0.6)$, $x_{i,2} \sim Gamma(20, 100)$ e $x_{i,3} \sim Gamma(1, 1)$.

Com isso gerou-se $y_i \sim D_k(\mu = e^{X'_{1:2}\beta}; \pi = \dfrac{e^{X'_{2:3}\gamma}}{1 + e^{X'_{2:3}\gamma}})$, D_k expressando o k-ésimo modelo, portanto D_k é modelo de origem de Y condicionado em X. Desta forma foram observados para μ valores que se estendem de 2.95 e 70.9 e para π foram observados valores no intervalo de 0.02 a 0.17. Os demais parâmetros de sobredispersão foram setados de forma a se obter grande variedade de índices de sobredispersão, cujos valores observados se extendem de 1.87 a 29.15.

Gerados os dados, procedeu-se o ajuste dos modelos. Para cada banco foram ajustado os nove modelos inflacionados de zero abordados neste estudo, além da regressão de Poisson tradicional, o que confere a cada banco dez ajustes. Como métricas para avaliar a adaptabilidade dos modelos aos dados foram coletadas as estimativas dos coeficientes e seus erros padrões.

Já para a verificação da qualidade do ajuste foram utilizados o logarítmo da função de verossimilhança maximizada, que consta nas tabelas como *LogLik*; o critério de informação de Akaike (AIC), de Hilbe (2014) já bastante utilizado, o critério de informação de Hannan-Quinn (HQC), que é frequentemente usado como um critério para a seleção de modelos entre um conjunto finito de modelos e o critério de informação bayesiano (BIC), uma medida de ajuste que possui um termo que penaliza o número de parâmetros do modelo de uma forma mais grave que o AIC.

4 Resultados

O ajuste dos modelos e obtenção das estimativas dos coeficientes regressores foi realizada via maximização de verossimilhaça que se deu pelo método iterativo de Fisher Scoring, um algoritmo Hill Climbing com critério de convergência definido por uma diferença absoluta mínima entre as verossimilhanças de duas iterações sucessivas. Essa classe de algorítmos apesar de amplamente versátil, apresenta problemas de convergência frente a alguns cenários dentro de um número limitado de iterações. Este trabalho conforme esperado encontrou problemas de convergência em alguns bancos e modelos, conforme já exposto por Silva (2017) em sua dissertação. Globalmente obtivemos convergência em 91.18%

dos ajustes. A regressão de Poisson, ZIP e ZIB convergiram em 100% do ajustes. ZIBN, ZIG e ZIBB apresentaram convergência na casa dos 97%, já ZIBNB, ZICMP, ZIDelaporte e ZIPIG retornaram 80% de convergência.

Nota-se uma proporcionalidade entre percentual de convergência e complexidade do modelo ajustado. Já a convergência segundo o modelo do qual o dados foram gerados apresentou percentual homogêneo na casa dos 91%. Dados simulados de ZIP, ZIB e ZIG foram os com menor índice (88%), justamente os modelos mais simplistas. Ou seja, evidenciamos em nossos dados que frente uma sobreparametrização há maiores chances de se registrar uma falha na convergência do modelo.

A Tabela 2 apresentada abaixo apresenta as estimativas médias dos coeficientes de regressão de μ e de seus erros padrões relativos ao modelos convergentes.

Tabela 2: Estimativas para β_0, β_1 e β_2 e seus respectivos erros padrões dos modelos de regressão aplicados à simulações de diferentes tipos de dados inflacionados de zeros

		POIS	**ZIP**	**ZINB**	**ZIG**	**ZIB**	**ZIPIG**	**ZIBB**	**ZIBNB**	**ZICMP**	**ZID**
ZIP	$\hat{\beta}_0(\hat{\sigma}_{\beta 0})$	0.96 (0.09)	1.00 (0.09)	1.00 (0.09)	0.95 (0.35)	1.00 (0.09)	1.00 (0.10)	0.98 (0.12)	0.98 (0.01)	1.00 (0.09)	0.97 (0.08)
	$\hat{\beta}_1(\hat{\sigma}_{\beta 1})$	0.50 (0.02)	0.50 (0.02)	0.50 (0.02)	0.50 (0.08)	0.50 (0.02)	0.50 (0.02)	0.51 (0.03)	0.53 (0.00)	0.50 (0.02)	0.51 (0.02)
	$\hat{\beta}_2(\hat{\sigma}_{\beta 2})$	-0.56 (0.27)	-0.51 (0.27)	-0.51 (0.28)	-0.54 (1.06)	-0.51 (0.27)	-0.51 (0.28)	-0.53 (0.36)	-0.55 (0.04)	-0.51 (0.27)	-0.47 (0.17)
ZINB	$\hat{\beta}_0(\hat{\sigma}_{\beta 0})$	0.96 (0.09)	1.01 (0.09)	1.00 (0.16)	0.95 (0.35)	1.01 (0.09)	1.00 (0.16)	1.02 (0.16)	1.06 (0.14)	0.97 (0.16)	1.04 (0.16)
	$\hat{\beta}_1(\hat{\sigma}_{\beta 1})$	0.50 (0.02)	0.50 (0.02)	0.50 (0.04)	0.50 (0.08)	0.50 (0.02)	0.50 (0.04)	0.48 (0.04)	0.49 (0.03)	0.51 (0.04)	0.49 (0.03)
	$\hat{\beta}_2(\hat{\sigma}_{\beta 2})$	-0.55 (0.27)	-0.50 (0.27)	-0.50 (0.48)	-0.52 (1.06)	-0.50 (0.27)	-0.50 (0.48)	-0.49 (0.48)	-0.50 (0.40)	-0.51 (0.48)	-0.46 (0.46)
ZIG	$\hat{\beta}_0(\hat{\sigma}_{\beta 0})$	0.94 (0.09)	1.16 (0.09)	0.99 (0.36)	0.99 (0.36)	1.17 (0.09)	1.08 (0.38)	1.43 (0.32)	1.46 (0.33)	1.42 (0.41)	1.48 (0.33)
	$\hat{\beta}_1(\hat{\sigma}_{\beta 1})$	0.50 (0.02)	0.47 (0.02)	0.50 (0.08)	0.50 (0.08)	0.47 (0.02)	0.49 (0.08)	0.36 (0.07)	0.37 (0.07)	0.14 (0.00)	0.37 (0.07)
	$\hat{\beta}_2(\hat{\sigma}_{\beta 2})$	-0.53 (0.27)	-0.47 (0.27)	-0.49 (1.09)	-0.49 (1.10)	-0.47 (0.27)	-0.47 (1.15)	-0.35 (0.98)	-0.36 (0.99)	-0.50 (0.03)	-0.44 (1.04)
ZIB	$\hat{\beta}_0(\hat{\sigma}_{\beta 0})$	0.96 (0.09)	1.00 (0.09)	1.00 (0.13)	0.95 (0.35)	1.00 (0.09)	1.00 (0.12)	0.98 (0.14)	0.98 (0.01)	1.00 (0.07)	0.99 (0.08)
	$\hat{\beta}_1(\hat{\sigma}_{\beta 1})$	0.50 (0.02)	0.50 (0.02)	0.50 (0.03)	0.50 (0.08)	0.50 (0.02)	0.50 (0.03)	0.51 (0.03)	0.52 (0.00)	0.50 (0.02)	0.50 (0.02)
	$\hat{\beta}_2(\hat{\sigma}_{\beta 2})$	-0.57 (0.27)	-0.51 (0.27)	-0.51 (0.37)	-0.53 (1.05)	-0.51 (0.27)	-0.51 (0.35)	-0.52 (0.42)	-0.55 (0.04)	-0.51 (0.21)	-0.50 (0.19)
ZIPIG	$\hat{\beta}_0(\hat{\sigma}_{\beta 0})$	0.95 (0.09)	1.00 (0.09)	0.99 (0.18)	0.95 (0.35)	1.00 (0.09)	0.99 (0.18)	1.03 (0.18)	1.01 (0.17)	0.93 (0.18)	1.05 (0.17)
	$\hat{\beta}_1(\hat{\sigma}_{\beta 1})$	0.50 (0.02)	0.50 (0.02)	0.50 (0.04)	0.51 (0.08)	0.50 (0.02)	0.50 (0.04)	0.48 (0.04)	0.50 (0.04)	0.52 (0.04)	0.49 (0.04)
	$\hat{\beta}_2(\hat{\sigma}_{\beta 2})$	-0.55 (0.27)	-0.01 (0.27)	-0.50 (0.52)	-0.52 (1.06)	-0.50 (0.27)	-0.49 (0.53)	-0.48 (0.52)	-0.49 (0.51)	-0.50 (0.53)	-0.47 (0.50)
ZIBB	$\hat{\beta}_0(\hat{\sigma}_{\beta 0})$	0.96 (0.09)	1.09 (0.09)	1.07 (0.20)	0.96 (0.35)	1.09 (0.09)	0.88 (0.21)	1.00 (0.20)	0.92 (0.19)	0.93 (0.18)	0.92 (0.19)
	$\hat{\beta}_1(\hat{\sigma}_{\beta 1})$	0.50 (0.02)	0.48 (0.02)	0.48 (0.04)	0.50 (0.08)	0.48 (0.02)	0.54 (0.05)	0.53 (0.04)	0.52 (0.04)	0.99 (0.19)	0.52 (0.04)
	$\hat{\beta}_2(\hat{\sigma}_{\beta 2})$	-0.55 (0.27)	-0.46 (0.27)	-0.45 (0.59)	-0.50 (1.06)	-0.46 (0.27)	-0.50 (0.61)	-0.53 (0.57)	-0.50 (0.56)	-0.48 (0.56)	-0.47 (0.56)
ZIBNB	$\hat{\beta}_0(\hat{\sigma}_{\beta 0})$	0.89 (0.09)	1.02 (0.09)	0.99 (0.22)	0.90 (0.36)	1.02 (0.09)	0.99 (0.22)	1.09 (0.22)	1.01 (0.22)	0.90 (0.23)	1.09 (0.21)
	$\hat{\beta}_1(\hat{\sigma}_{\beta 1})$	0.50 (0.02)	0.50 (0.02)	0.50 (0.05)	0.51 (0.08)	0.50 (0.02)	0.50 (0.05)	0.46 (0.05)	0.50 (0.05)	0.52 (0.05)	0.47 (0.05)
	$\hat{\beta}_2(\hat{\sigma}_{\beta 2})$	-0.50 (0.28)	-0.50 (0.28)	-0.51 (0.65)	-0.53 (1.09)	-0.50 (0.27)	-0.51 (0.67)	-0.45 (0.65)	-0.50 (0.66)	-0.51 (0.67)	-0.46 (0.63)
ZICMP	$\hat{\beta}_0(\hat{\sigma}_{\beta 0})$	0.96 (0.09)	1.00 (0.09)	1.00 (0.11)	0.96 (0.35)	1.00 (0.09)	1.00 (0.11)	0.93 (0.11)	1.01 (0.01)	1.00 (0.11)	0.99 (0.09)
	$\hat{\beta}_1(\hat{\sigma}_{\beta 1})$	0.50 (0.02)	0.50 (0.02)	0.50 (0.02)	0.50 (0.08)	0.50 (0.02)	0.50 (0.02)	0.51 (0.02)	0.52 (0.00)	0.50 (0.02)	0.50 (0.02)
	$\hat{\beta}_2(\hat{\sigma}_{\beta 2})$	-0.55 (0.27)	-0.49 (0.27)	-0.49 (0.32)	-0.53 (1.06)	-0.49 (0.27)	-0.49 (0.32)	-0.51 (0.33)	-0.49 (0.04)	-0.50 (0.32)	-0.48 (0.24)
ZID	$\hat{\beta}_0(\hat{\sigma}_{\beta 0})$	0.97 (0.09)	1.01 (0.09)	1.00 (0.20)	0.96 (0.35)	1.01 (0.09)	0.93 (0.20)	0.99 (0.20)	0.97 (0.20)	0.92 (0.20)	1.00 (0.19)
	$\hat{\beta}_1(\hat{\sigma}_{\beta 1})$	0.50 (0.02)	0.50 (0.02)	0.50 (0.04)	0.50 (0.08)	0.50 (0.02)	0.52 (0.05)	0.49 (0.04)	0.51 (0.04)	0.52 (0.04)	0.50 (0.04)
	$\hat{\beta}_2(\hat{\sigma}_{\beta 2})$	-0.57 (0.27)	-0.51 (0.27)	-0.52 (0.59)	-0.55 (1.06)	-0.51 (0.27)	-0.54 (0.60)	-0.52 (0.57)	-0.53 (0.58)	-0.53 (0.60)	-0.51 (0.57)

Verifica-se que dentro de um limiar, em média as estimativas são satisfatórias e parecem pouco viesadas. Silva (2017) mostra via simulação que EM em comparação a Hill Climbing é superior e preferível, pois apresenta menor viés e melhor índice de convergência. Fica evidente ainda que vício e convergência são afetados conjuntamente pelo π e n. Este projeto por atribuir um grau baixo a moderado de zeros e um n amistoso não lida com problema de grandes viéses e raras convergências. Referente aos erros padrões percebe-se que modelos mais complexos tendem a apresentar erros padrões maiores, com excessão do ZIG, que retorna erros bastante superiores aos outros modelos.

Na Tabela 3 apresentada abaixo são expostas as estimativas médias dos coeficientes regressores associados ao processo logístico que modela a probabilidade de pertencer ao grupo sempre zero, bem como

seus erros padrões.

Tabela 3: Estimativas para γ_0, γ_1 e γ_2 e seus respectivos erros padrões dos modelos de regressão aplicados à simulações de diferentes tipos de dados inflacionados de zeros

		ZIP	**ZINB**	**ZIG**	**ZIB**	**ZIPIG**	**ZIBB**	**ZIBNB**	**ZICMP**	**ZID**
ZIP	$\hat{\gamma}_0(\hat{\sigma}_{\gamma 0})$	-2,00 (1,00)	-0,99 (1,04)	-1,75 (3,25)	-2,00 (1,00)	-1,98 (1,01)	-2,49 (1,05)	-2,04 (1,06)	-2,02 (1,00)	-2,45 (0,99)
	$\hat{\gamma}_1(\hat{\sigma}_{\gamma 1})$	0,94 (4,75)	0,89 (4,93)	0,64 (14,07)	0,94 (4,75)	0,88 (4,77)	0,73 (4,98)	0,83 (5,03)	1,10 (4,73)	1,12 (4,62)
	$\hat{\gamma}_2(\hat{\sigma}_{\gamma 2})$	-2,13 (0,64)	-2,13 (0,68)	-2,23 (6,23)	-2,13 (0,64)	-2,12 (0,64)	-2,02 (0,70)	-2,14 (0,7)	-2,12 (0,64)	-2,05 (0,62)
ZINB	$\hat{\gamma}_0(\hat{\sigma}_{\gamma 0})$	-2,02 (1,00)	-2,01 (1,02)	-1,59 (3,28)	-2,02 (1,00)	-2,01 (1,01)	-2,17 (1,02)	-2,03 (1,03)	-2,11 (1,07)	-2,06 (1,01)
	$\hat{\gamma}_1(\hat{\sigma}_{\gamma 1})$	0,92 (4,75)	0,91 (4,84)	0,56 (13,92)	0,93 (4,75)	0,91 (4,81)	0,84 (4,83)	0,76 (4,87)	1,48 (5,06)	1,27 (4,76)
	$\hat{\gamma}_2(\hat{\sigma}_{\gamma 2})$	-2,04 (0,62)	-2,13 (0,67)	-1,62 (6,33)	-2,04 (0,62)	-2,10 (0,65)	-1,93 (0,70)	-2,12 (0,67)	-2,32 (0,77)	-2,12 (0,65)
ZIG	$\hat{\gamma}_0(\hat{\sigma}_{\gamma 0})$	-1,75 (0,67)	-1,92 (1,51)	-1,87 (1,49)	-1,75 (0,67)	-1,82 (0,84)	-2,35 (1,59)	-1,91 (1,38)	-1,90 (1,09)	-1,99 (1,30)
	$\hat{\gamma}_1(\hat{\sigma}_{\gamma 1})$	1,37 (3,18)	1,41 (7,21)	1,31 (7,12)	1,37 (3,18)	1,25 (3,98)	1,09 (7,59)	1,10 (6,66)	1,14 (4,98)	1,21 (6,01)
	$\hat{\gamma}_2(\hat{\sigma}_{\gamma 2})$	-2,44 (0,20)	-1,56 (1,52)	-2,49 (1,42)	-2,44 (0,2)	-1,78 (0,38)	-2,00 (1,85)	-2,05 (1,19)	-2,44 (0,85)	-1,67 (1,11)
ZIB	$\hat{\gamma}_0(\hat{\sigma}_{\gamma 0})$	-1,99 (1,01)	1,99 (1,07)	-1,93 (3,22)	-1,99 (1,01)	-1,98 (1,06)	-1,99 (1,07)	-2,03 (1,08)	-1,94 (1,00)	-1,64 (1,00)
	$\hat{\gamma}_1(\hat{\sigma}_{\gamma 1})$	0,88 (4,77)	0,98 (5,06)	1,38 (13,39)	0,88 (4,77)	0,83 (5,00)	0,85 (5,07)	0,84 (5,11)	0,73 (4,75)	0,96 (4,63)
	$\hat{\gamma}_2(\hat{\sigma}_{\gamma 2})$	-2,14 (0,64)	-2,12 (0,70)	-1,60 (6,68)	-2,14 (0,64)	-2,13 (0,69)	-2,12 (0,71)	-2,14 (0,71)	-2,14 (0,64)	-2,02 (0,61)
ZIPIG	$\hat{\gamma}_0(\hat{\sigma}_{\gamma 0})$	-2,00 (1,00)	-1,98 (1,03)	-1,55 (3,26)	-2,00 (1,00)	-1,99 (1,02)	-1,93 (1,04)	-1,99 (1,03)	-2,14 (1,13)	-1,98 (1,01)
	$\hat{\gamma}_1(\hat{\sigma}_{\gamma 1})$	0,89 (4,74)	0,89 (4,88)	0,88 (13,78)	0,89 (4,73)	0,88 (4,82)	0,80 (4,90)	0,89 (4,86)	0,65 (5,30)	0,90 (4,76)
	$\hat{\gamma}_2(\hat{\sigma}_{\gamma 2})$	-2,06 (0,62)	-2,20 (0,69)	-2,33 (6,40)	-2,06 (0,62)	-2,15 (0,67)	-1,98 (0,73)	-2,18 (0,69)	-1,51 (0,88)	-2,11 (0,65)
ZIBB	$\hat{\gamma}_0(\hat{\sigma}_{\gamma 0})$	-2,17 (0,87)	-2,09 (0,94)	-1,63 (2,82)	-2,17 (0,87)	-2,10 (0,92)	-1,76 (0,98)	-2,05 (0,99)	-2,10 (1,05)	-2,39 (0,98)
	$\hat{\gamma}_1(\hat{\sigma}_{\gamma 1})$	1,30 (4,11)	1,31 (4,44)	0,58 (12,36)	1,30 (4,11)	1,30 (4,32)	0,97 (4,63)	1,20 (4,67)	0,59 (4,91)	0,83 (4,58)
	$\hat{\gamma}_2(\hat{\sigma}_{\gamma 2})$	-2,14 (0,38)	-1,52 (0,56)	-2,38 (5,15)	-2,14 (0,38)	-2,40 (0,50)	-1,61 (0,66)	-1,72 (0,62)	-1,94 (0,75)	-1,66 (0,62)
ZIBNB	$\hat{\gamma}_0(\hat{\sigma}_{\gamma 0})$	-2,13 (0,70)	-2,17 (0,73)	-1,98 (2,12)	-2,14 (0,70)	-2,15 (0,71)	-2,07 (0,75)	-2,16 (0,72)	-1,54 (0,84)	-2,26 (0,72)
	$\hat{\gamma}_1(\hat{\sigma}_{\gamma 1})$	0,87 (3,33)	0,78 (3,52)	0,57 (10,17)	0,87 (3,33)	0,82 (3,41)	0,83 (3,60)	0,80 (3,47)	0,97 (3,96)	1,38 (3,41)
	$\hat{\gamma}_2(\hat{\sigma}_{\gamma 2})$	-2,02 (0,15)	-2,03 (0,16)	-1,80 (0,96)	-2,02 (0,15)	-2,02 (0,16)	-2,01 (0,17)	-2,02 (0,16)	-2,04 (0,19)	-2,03 (0,16)
ZICMP	$\hat{\gamma}_0(\hat{\sigma}_{\gamma 0})$	-2,04 (1,01)	-2,04 (1,01)	-1,87 (3,24)	-2,04 (1,01)	-2,04 (1,01)	-2,15 (1,00)	-2,11 (1,08)	-2,04 (1,01)	-2,23 (1,00)
	$\hat{\gamma}_1(\hat{\sigma}_{\gamma 1})$	1,08 (4,78)	1,08 (4,79)	1,00 (13,84)	1,08 (4,78)	1,09 (4,79)	0,98 (4,71)	1,15 (5,12)	1,11 (4,78)	1,01 (4,71)
	$\hat{\gamma}_2(\hat{\sigma}_{\gamma 2})$	-2,12 (0,64)	-2,13 (0,64)	-1,67 (6,59)	-2,12 (0,64)	-2,13 (0,64)	-1,96 (0,65)	-2,12 (0,72)	-2,14 (0,65)	-2,05 (0,62)
ZID	$\hat{\gamma}_0(\hat{\sigma}_{\gamma 0})$	-2,03 (0,99)	-2,00 (1,05)	-2,24 (3,22)	-2,03 (0,99)	-2,00 (1,02)	-1,52 (1,06)	-2,00 (1,04)	-2,21(1,20)	-2,00 (1,02)
	$\hat{\gamma}_1(\hat{\sigma}_{\gamma 1})$	1,00 (4,66)	0,97 (4,97)	0,68 (13,74)	1,00 (4,66)	0,96 (4,83)	0,82 (5,03)	0,94 (4,95)	0,90 (5,63)	0,99 (4,81)
	$\hat{\gamma}_2(\hat{\sigma}_{\gamma 2})$	-1,95 (0,59)	-2,22 (0,73)	-1,71 (6,38)	-1,95 (0,59)	-2,11 (0,67)	-1,99 (0,79)	-2,19 (0,71)	-1,65 (1,00)	-2,12 (0,68)

As estimativas médias de γ deixam de ser tão satisfatórias quanto as estimativas dos coeficientes associados à μ, o que deixa claro o vício decorrente do método de otimização e possível confundimento, uma vez que há uma covariável ($X_{i,2}$) que estabele uma interseção entre conjunto de covariáveis de μ e π, ou seja, está associada ao β_2 e γ_1. O erro padrão médio assume também novas escalas (em comparação com os erros padrões do vetor β) justamente pela interseção de covariáveis anteriormente mencionado, sendo observado valor médio máximo de 14.07 no modelo ZIG (que já apresentou erros padrões altos para o vetor $\hat{\beta}$). Da mesma maneira que nas estimativas do vetor β, o ZIB apresenta o menor erro padrão, seguido pelo modelo ZINB e ZIP, ZIPIG e ZID.

A Tabela 4 traz as médias das medidas de ajuste selecionadas para verificar qualidade do ajuste e comparar modelos. São apresentadas as medidas: o logaritmo da verossimilhança Maximizada, AIC, HQC e BIC. Estas últimas três medidas buscam por meio da verossimilhança tornar diferentes modelos comparáveis.

Os modelos que obtiveram as menores médias nos critérios de qualidade de ajuste estão evidenciados na tabela abaixo em *negrito*. Se mais de uma estimativa possui médias muito similares entre os mesmos dados, então ambas estão destacadas.

Tabela 4: Qualidade de ajuste do modelo frente aos dados inflacionado de zeros

		POIS	**ZIP**	**ZINB**	**ZIG**	**ZIB**	**ZIPIG**	**ZIBB**	**ZIBNB**	**ZICMP**	**ZID**
ZIP	LogLik	-1659,96	**-1384,91**	**-1384,98**	-1814,10	**-1384,91**	-1386,08	-2612,01	-1434,20	**-1384,83**	-1387,31
	AIC	3325,92	**2781,83**	**2783,96**	3640,20	**2781,81**	2786,16	5238,02	2884,36	**2783,66**	2790,63
	HQC	3330,88	**2791,75**	**2795,53**	3650,12	**2791,73**	2797,74	5249,60	2897,55	**2795,24**	2800,20
	BIC	3338,56	**2807,11**	**2813,46**	3665,49	**2807,10**	2815,66	5267,52	2917,98	**2813,17**	2818,13
ZINB	LogLik	-2125,73	-1851,37	**-1626,48**	-1814,55	-1859,57	**-1627,87**	-1640,95	-1630,41	-1632,34	**-1628,69**
	AIC	4257,46	3714,74	**3266,97**	3641,10	3731,15	**3269,74**	3295,91	3276,82	3278,68	**3273,39**
	HQC	4262,42	3724,66	**3278,54**	3651,02	3741,07	**3281,32**	3307,48	3290,06	3290,26	**3282,96**
	BIC	4270,11	3740,03	**3296,47**	3666,39	3756,43	**3299,24**	3325,41	3310,54	3308,19	**3300,89**
ZIG	LogLik	-4476,17	-3826,92	**-1803,02**	**-1803,60**	-3882,01	-1816,54	-1809,42	-1807,83	-1828,78	-1808,24
	AIC	8958,35	7665,83	**3620,05**	**3619,20**	7776,02	3647,08	3632,84	3631,66	3671,57	3632,48
	HQC	8963,31	7675,75	**3631,62**	**3629,12**	7785,94	3658,65	3644,41	3644,89	3683,14	3642,06
	BIC	8970,99	7691,12	**3649,55**	**3644,49**	7801,31	3676,58	3662,34	3665,38	3701,07	3659,98
ZIB	LogLik	-1565,17	-1291,84	-1292,31	-1814,55	-1290,25	-1293,25	-2517,81	-1388,65	**-1262,28**	-1293,52
	AIC	3136,34	2595,67	2598,62	3641,11	2592,50	2600,49	5049,63	2793,30	**2538,56**	2603,04
	HQC	3141,30	2605,59	2610,20	3651,03	2602,42	2612,07	5061,20	2806,53	**2550,14**	2612,62
	BIC	3148,99	2620,96	2628,13	3666,39	2617,79	2629,99	5079,13	2827,02	**2568,06**	2630,54
ZIPIG	LogLik	-2274,02	-1998,31	-1659,07	-1814,22	-2009,82	**-1656,95**	-1673,09	-1666,11	-1671,27	**-1658,77**
	AIC	4554,03	4008,61	3332,13	3640,44	4031,65	**3327,90**	3360,18	3348,21	3356,55	**3333,54**
	HQC	4559,00	4018,54	3343,71	3650,37	4041,57	**3339,48**	3371,75	3361,44	3368,12	**3343,11**
	BIC	4566,68	4033,90	3361,63	3665,73	4056,93	**3357,40**	3389,68	3381,93	3386,05	**3361,04**
ZIBB	LogLik	-2416,31	-2079,16	-1708,62	-1812,52	-2088,75	-1721,87	-1699,84	**-1695,36**	-1689,54	**-1694,78**
	AIC	4838,61	4170,32	3431,24	3637,03	4189,49	3457,75	3413,69	**3406,72**	3393,08	**3405,56**
	HQC	4843,58	4180,24	3442,82	3646,96	4199,42	3469,33	3425,26	**3419,95**	3404,66	**3415,14**
	BIC	4851,26	4195,61	3460,74	3662,32	4214,78	3487,25	3443,19	**3440,44**	3422,58	**3433,07**
ZIBNB	LogLik	-2870,55	-2314,99	-1710,29	-1785,20	-2332,88	**-1709,35**	-1725,21	**-1708,09**	-1720,44	-1711,33
	AIC	5747,11	4641,97	3434,58	3582,40	4677,75	**3432,70**	3464,43	**3432,18**	3454,88	3438,67
	HQC	5752,07	4651,90	3446,16	3592,33	4687,67	**3444,28**	3476,01	**3445,41**	3466,45	3448,24
	BIC	5759,75	4667,26	3464,08	3607,69	4703,04	**3462,21**	3493,93	**3465,90**	3484,38	3466,17
ZICMP	LogLik	-1750,53	-1479,49	-1464,98	-1815,12	-1480,93	-1465,39	-1478,03	-1504,38	**-1463,99**	**-1464,12**
	AIC	3507,07	2970,98	2943,96	3642,23	2973,87	2944,77	2970,05	3024,76	**2941,97**	**2944,24**
	HQC	3512,03	2980,91	2955,54	3652,16	2983,79	2956,35	2981,63	3037,99	**2953,55**	**2953,82**
	BIC	3519,71	2996,27	2973,46	3667,52	2999,16	2974,28	2999,56	3058,47	**2971,48**	**2971,75**
ZID	LogLik	-2464,62	-2182,17	-1701,07	-1813,56	-2196,32	-1698,73	-1707,18	**-1698,44**	-1708,60	**-1696,34**
	AIC	4935,23	4376,33	3416,13	3639,11	4404,63	3411,46	3428,35	**3412,88**	3431,19	**3408,68**
	HQC	4940,20	4386,25	3427,71	3649,04	4414,56	3423,03	3439,93	**3426,11**	3442,77	**3418,26**
	BIC	4947,88	4401,62	3445,63	3664,40	4429,92	3440,96	3457,85	**3446,60**	3460,70	**3436,19**

Espera-se que a diagonal apresente sempre indicativos de um bom ajuste, uma vez que representa a situação onde o modelo correto foi ajustado. Ou seja, o modelo originário é o mesmo que o ajustado. Com isso em mente percebe-se que únicas situações onde a diagonal não pertence ao grupo dos bons ajustes são os modelos ZIB e ZIBB, ambos provenientes da distribuição binomial.

Na contramão, o modelo que mais recebeu indicação de melhor ajuste foi o ZID, que além de ter tido a menor média dos critérios de qualidade de ajuste para os dados provenientes dessa mesma distribuição, também obteve a menor média com os dados simulados pelas distribuições ZINB,ZIPIG ,ZIBB e ZICMP.

Em seguida com três indicações de melhor ajuste surgem os modelos ZINB, ZIPIG, ZIBNB e ZICMP, todos modelos flexíveis com grande cobertura de índice de sobredispersão.

Na Tabela 5 abaixo estão expostas as proporções de vezes que dentre os 1000 ajustes o modelo empregado é dono da menor medida de HQC segundo a origem dos dados. Desta maneira espera-se similarmente que a diagonal *(em negrito)* contenha as maiores proporções, uma vez que é aguardado que o ajuste do modelo correto forneça uma alta taxa de melhor adequamento.

Casos onde a diagonal não representa a maior proporção de HQC mínimo estão indicados em vermelho. Importante notar que ZICMP apresenta três indicações de maior proporção de menor HQC. Outros modelos que conseguem apresentar uma proporção superior ao modelo de origem são ZINB, ZIB e ZIPIG.

Verifica-se que mais uma vez os modelos mais flexíveis conseguem se adequar bem a dados que são oriundos de distribuições simples, por exemplo ZINB, ZIPIG e ZICMP representam mais de 50% dos ajustes de menor HQC, ao passo que ZIP forneceu o melhor ajuste apenas 4% das vezes e Poisson 0%.

Tabela 5: Porcentagem de modelos com HQC mínimo

	POIS	**ZIP**	**ZINB**	**ZIG**	**ZIB**	**ZIPIG**	**ZIBB**	**ZIBNB**	**ZICMP**	**ZID**
ZIP	0,00	**0,41**	0,00	0,00	**0,49**	0,00	0,00	0,04	0,02	0,02
ZINB	0,00	0,00	**0,64**	0,00	0,00	0,14	0,03	0,08	0,02	0,09
ZIG	0,00	0,00	0,07	**0,90**	0,00	0,00	0,01	0,00	0,00	0,02
ZIB	0,00	0,00	0,00	0,00	**0,11**	0,00	0,00	0,00	**0,88**	0,00
ZIPIG	0,00	0,00	0,12	0,00	0,00	**0,65**	0,01	0,01	0,00	0,21
ZIBB	0,00	0,00	0,00	0,00	0,00	0,00	**0,14**	0,01	**0,84**	0,01
ZIBNB	0,00	0,00	**0,34**	0,00	0,00	**0,51**	0,01	**0,05**	0,00	0,09
ZICMP	0,00	0,00	0,10	0,00	0,00	0,03	0,23	0,28	**0,31**	0,05
ZID	0,00	0,00	0,05	0,00	0,00	0,15	0,08	0,06	0,00	**0,67**

5 Conclusões

As simulações evidenciam a flexibilidade de cada modelo, tanto na simulação de dados como no ajuste destes. Obviamente espera-se melhores resultados quando se ajusta aos dados o modelo que de fato gere o processo de contagem da população, mas esse modelo é tão desconhecido quanto os próprios parâmetros, logo o processo de ajuste, antes de passar pela estimação de parâmentros, requer a definição de modelo apropriado.

Dos modelos aqui abordados e brevemente testados se fortificam evidências da sobrepujança de alguns sobre outros. É possível dizer que a inclusão da inflação de zeros no modelo de fato é crucial quando se está lidando com dados desta natureza, bem como atenção a sobredispersão é fundamental. É recomendado testes mais exaustivos, trabalhando com diferentes graus de sobredispersão em cada modelo e buscar táticas mais adequedas para obtenção de estimativas, pois conforme mostrado por Silva (2017), para certas configurações de parâmetros e tamanho amostral, a maximização sem emprego do EM representa um grande risco.

Chama-se a atenção para modelos tradicionais que incorporam sobredispersão como PIG e NB, que ao serem inflados de zero passam não somente a absorver a sobredispersão comum, mas também a modelar os zeros estruturais que reduzem a média e elevam a variância dos dados, o que os torna ainda mais versáteis.

Houveram boas surpresas com as distribuições discretas Delaporte e sobretudo CMP, que neste estudo apresentaram desempenho tão bom quanto, ou melhor, que NB e PIG frente a casos de sobredispersão. Suas derivações infladas de zeros ZID e ZICMP performaram muito bem mais uma vez e se posicionaram paralelamente ou a frente dos modelos ZINB e ZIPIG. Mais estudos e simulações são requeridos, pois a utilização destas distribuições para análise de dados de contagem é escassa.

Negativamente menciona-se a distribuição ZIBNB, que além de superparametrizada, apresentou resultados similares ao ZIB e ZIBB. Além disso menciona-se peculiaridade do modelo ZIG cujos dados não foram bem ajustados por nenhum modelo além do próprio ZIG, sendo estes dados um dos poucos que o modelo ajusta de maneira satisfatória. À cerca do ZIP, conforme esperado ele apenas desempenhou bem com dados que além da inflação de zeros eram equidispersos, já o modelo Poisson tradicional desempenhou mal em todos os cenários, o que corrobora com a atenção necessária que devemos ter com a sobredispersão e inflação de zeros, uma vez que regressão de Poisson costuma ser tomada como procedimento padrão frente a dados de contagem.

6 Referências

CONWAY, R.W. and MAXWELL, W.L.*A queuing model with state dependent service rates*. J. Ind. Eng. 12, 132–136.1962.

COX, D. R.,LEWIS, P. A. W.*The Statistical Analysis of Series of Events*. 1966

DELAPORTE, P.J. *Quelques problèmes de statistiques mathématiques poses par l'Assurance Automobile et le Bonus pour non sinistre [Some problems of mathematical statistics as related to automobile insurance and no-claims bonus]*. Bulletin Trimestriel de l'Institut des Actuaires Français (in French). 1960. 87–102 p.

HAIGHT, F.A. *Handbook of the Poisson Distribution*. New York: John Wiley & Sons, 1967.

HAL, D.B. , *Zero-Inflated Poisson and Binomial Regression with Random Effects: A Case Study*, Department of Statistics, University of Georgia. 2000

HEILBRON, D.C. *Generalized linear models for altered zero probabilities and overdispersion in count data*. SIMS Technical Report 9, Department of Epidemiology and Biostatistics, University of California, San Francisco. 1989.

JENNRICH, R. I., and SAMPSON, P.F. *Newton-Raphson and related algorithms for maximum likelihood variance component estimation*. Technometrics, 18. 1976. 11-17 p.

LAMBERT, D. *Zero-Inflated Poisson Regression, With An Application to Defects in Manufacturing*, 1992.

NELDER, J.A. and WEDDERBURN, R.W.M. *Generalized Linear Models*. Journal of the Royal Statistical Society. 1972.

PAULA, G.P. *Modelos de Regressão com apoio Computacional*

RIDOUT, M. S., DEMÉTRIO, C.G.B. and HINDE, J.P. *Models for count data with many zeros*. 1998.

SELLERS, K.F. and RAIM, A. *A flexible zero-inflated model to address data dispersion*, Computational Statistics and Data Analysis. 2016.

SELLERS, K.F., SHMUELI, G. and BORLE, S. *The COM-Poisson model for count data: a survey of methods and applications*. Appl. Stoch. Models Bus. Ind. 28. 2011. 104–116 p.

SILVA, J.G. *Zero-Inflated Mixed Poisson Regression Models*. 2017

SIN, C. Y., WHITE, H. *Information criteria for selecting possibly misspecified parametric models*. Journal of Econometrics, 71(1), 1996. 207-225.

WANG, Z. *One mixed negative binomial distribution with application*. Journal of Statistical Planning and Inference. 2011.

WILLMOT, G.E. *The Poisson-Inverse Gaussian distribution as an alternative to the negative binomial*, Scandinavian Actuarial Journal, DOI: 10.1080/03461238.1987.10413823. 1987.

VIEIRA, A. M. C., HINDE, J. P., DEMETRIO, C. G. B. *Zero-inlated proportion data models applied to a biological control assay*. Journal of Applied Statistics 27(3), 2000. 373-389.

RIDOUT, M., HINDE, J., DEMETRIO, C. G. B. *A score test for testing a zero-inflated Poisson regression model against zero-inlated negative binomial alternatives*. Biometrics 57(1), 2001. 219-223.

Inferência Estatística para Classificação de Sinais Cardíacos

Mikaela Baldasso[1]

Marcio Valk [2]

Resumo: Doenças cardiovasculares são responsáveis por milhões de mortes anualmente, segundo a Organização Mundial da Saúde e, dado isso, várias são as iniciativas, em todo o mundo, que visam estimular o desenvolvimento de novas técnicas que permitam diagnosticar e prevenir essas enfermidades. Diferentes técnicas de diagnósticos são utilizadas para detectar e prevenir esses desfechos, em que busca-se, principalmente, utilizar métodos não invasivos, baratos e que resultem em respostas rápidas e confiáveis, como por exemplo, aqueles baseados em Eletrocardiogramas e Fonocardiogramas. A partir disso, nosso objetivo nesse trabalho é utilizar a estatística para fazer inferência sobre classificação, ou seja, mensurar a confiabilidade de uma técnica de diagnóstico, em particular testar o método baseado em U-estatística para classificação e agrupamento de dados.

Palavras-chave: *Doenças Cardíacas, Classificação, Inferência.*

2 Introdução

As doenças cardiovasculares (DCV) continuam sendo a principal causa de morbidade e mortalidade no mundo todo, de acordo com Liu et al. (2016). Estima-se que 17,5 milhões de pessoas morreram de DCV em 2012, representando 31% de todas as mortes globais (OMS 2015). Um dos primeiros passos na avaliação do sistema cardiovascular é o exame físico: a auscultação dos sons do coração é parte essencial do exame e pode fornecer importantes pistas iniciais na avaliação da doença, servindo de guia para um exame diagnóstico posterior.

A análise automatizada do som cardíaco nas aplicações clínicas geralmente consiste em três passos; Pré-processamento, segmentação e classificação. Nas últimas décadas, métodos para segmentação automatizada e classificação de sons cardíacos foram amplamente estudados. Muitos métodos demonstraram potencial para detectar com precisão patologias em aplicações clínicas. Infelizmente, as comparações entre técnicas foram dificultadas pela falta de bases de dados de alta qualidade, rigorosamente validadas e padronizadas de sons cardíacos obtidos a partir de uma variedade de condições saudáveis e patológicas. Em muitos casos, ambos os dados experimentais

[1]UFRGS - Universidade Federal do Rio Grande do Sul. Email: *mikaelabaldasso@gmail.com*
[2]UFRGS - Universidade Federal do Rio Grande do Sul. Email: *marcio.valk@ufrgs.br*

e clínicos são coletados a custos consideráveis, mas apenas analisados uma vez por seus colecionadores e, em seguida, arquivados indefinidamente por variados motivos, como mencionado em Liu et al. (2016).

Algoritmos baseados em aprendizado supervisionado são amplamente utilizados na classificação de dados, como *Support Vector Machine* (SVM, citeScholkopf2001, Scholkopf2002) ou *support vector data description* (SVDD, Tax e Duin (2004)). Outras abordagens concentram-se na estimativa de densidade paramétrica. Essas metodologias também podem ser aplicadas na detecção de novidades ou *outliers*, que são dois tópicos importantes em estatística e aprendizado de máquina, devido a sua relevância prática em cenários do mundo real. A detecção de novidade é a tarefa de classificar os dados que diferem em alguns aspectos dos dados usados durante o treinamento Pimentel et al. (2014). A detecção de anomalias, também chamada de análise outlier, é a tarefa de identificar dados que se desviam de algum comportamento esperado Chandola et al. (2009).

Com base nessa estrutura, Cybis et al. (2018) propõe um teste para avaliar a significância estatística no problema da classificação de um elemento. A abordagem baseada na U-estatística é apresentada e uma extensão de uma U-estatística de teste crucial é proposta. Para a utilização dessa técnica no contexto de séries temporais é necessário transformar os dados de alguma maneira. Para isso utilizamos o periodograma, que é uma estimativa da densidade espectral do sinal, ou seja, é uma medida que descreve como a força do sistema se comporta conforme a variação da frequência, que pode ser aplicado em análise e processamento dos eletrocardiogramas. Em termos gerais, uma maneira de estimar essa densidade espectral é encontrar a transformada de Fourier de tempo discreto das amostras do processo e apropriadamente e calcular a distância euclideana entre esses resultados.

3 Sinais cardíacos e seus padrões

Nosso objeto de estudo são sinais cardíacos provenientes de diferentes fontes; podem ser eletrocardiogramas (ECG's) ou fonocardiogramas (PCG's). Para esse trabalho, escolhemos alguns sinais do banco de dados *MIT-BIH Arrhythmia DataBase*, Goldberger et al. (2000), que foi disponibilizado, como material de teste padrão para avaliação de detectores de arritmia, em 1980. O conjunto contém 48 trechos de meia hora de registros obtidos de 47 indivíduos e as gravações foram digitalizadas com resolução de 11 bits em uma faixa de 10 mV. Na figura 3, apresentamos alguns sinais desse banco de dados. Existem 3 grupos de sinais: os sinais normais, sem qualquer tipo de anomalia; os sinais com algum tipo de arritmia considerada comum; e os sinais com arritmias não tão comuns. Dois sinais de cada grupo são apresentados no gráfico juntamente com

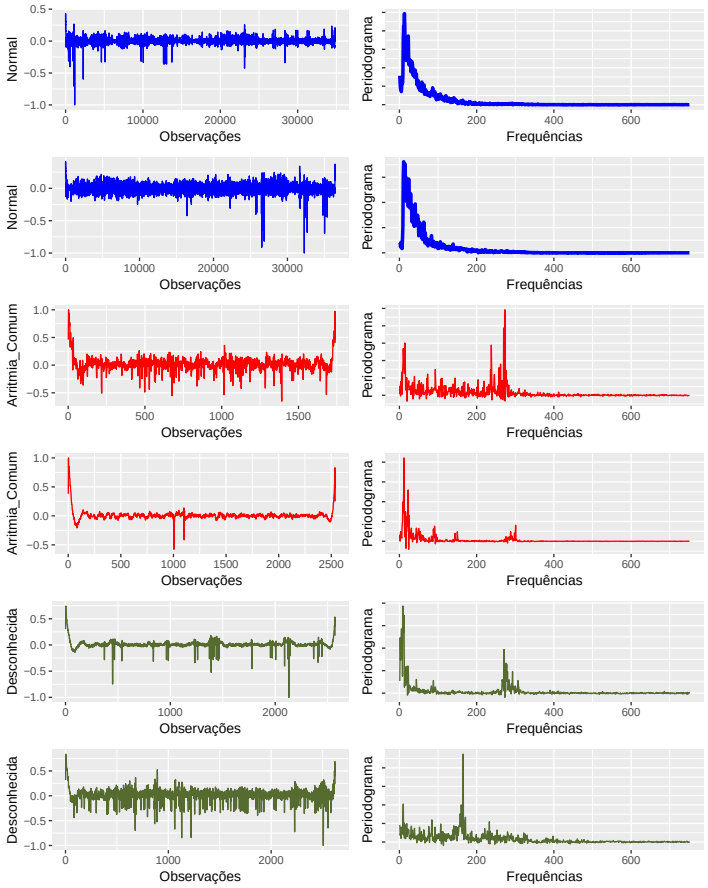


Figura 1: Sinais cardíacos com arritmia e sem arritmia com seus respectivos periodogramas que são transformações dos dados usadas na busca por padrões.

os seus respectivos periodogramas. Podemos observar que essa transformação dos dados captura padrões importantes. Isso se repete na maioria dos sinais observados. No grupo de sinais normais podemos observar um único pico no periodograma. Nos sinais com arritmias comuns podemos observar dois picos e nas arritmias não comuns o que prevalece é a não necessária existências de picos. É claro que há exceções a essa análise eurística e por isso a necessidade de um método estatístico para ajudar a decidir sabendo-se a probabilidade de errar.

Neste trabalho, temos por objetivo mensurar a confiabilidade do método baseado em u-estatísticas e, a partir disso, avaliar ECG's como séries temporais, em que a técnica de classificação e agrupamento pode ser aplicada.

O método de *clustering* é um conjunto de técnicas computacionais cujo propósito consiste em separar objetos em grupos distintos de acordo com as características que eles apresentam. De forma geral, a técnica consiste em colocar elementos similares em um mesmo grupo de acordo com algum critério já estipulado.

Uhclust - Método baseado em U-estatísticas

Dada uma amostra $X = (X_1, ..., X_n)$ de n vetores L-dimensionais dividida em dois grupos G_1 e G_2 de tamanhos n_1 e n_2 respectivamente onde $n = n_1 + n_2$. Sejam $X_1^{(g)}, ..., X_{n_g}^{(g)}$ as observações do g-ésimo grupo, independentes e com distribuição F_g. Define a distância funcional $\theta(F_1, F_2)$ por

$$\theta(F_1, F_2) = \int\int \phi(F_1, F_2) dF_1(x_1) dF_2(x_2)$$

onde $x_1, x_2 \in \mathbb{R}^L$.

Da teoria das U-estatísticas segue que um estimador não-viesado deste funcional para um mesmo grupo é uma estatística generalizada, com kernel $\phi(.,.)$ dada por

$$U_{n_g}^{(g)} = \binom{n_g}{2}^{-1} \sum_{1 \le i < j \le n_g} \phi(X_i^{(g)}, X_j^{(g)}).$$

Analogamente, o estimador para dois grupos diferentes é dado por

$$U_{n_1,n_2}^{(1,2)} = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \phi(X_i^{(1)}, X_j^{(2)}).$$

Note que a U-estatística pode ser decomposta por

$$\begin{aligned} U_n &= \binom{n}{2}^{-1} \sum_{1 \le i < j \le n_g} \phi(X_i, X_j) \\ &= \sum_{g=1}^{2} \frac{n_g}{n} U_{n_g}^{(g)} + \frac{n_1 n_2}{n(n-1)} (2U_{n_1 n_2}^{(1,2)} - U_{n_1}^{(1)} - U_{n_2}^{(2)}) \\ &= W_n + B_n. \end{aligned}$$

Assim, o teste, proposto por Cybis et al. (2018), consiste em verificar se G_1 e G_2 constituem grupos separados ou se derivam da mesma distribuição. Basicamente, quando os grupos derivam da mesma distribuição temos $F_1 = F_2$ e portanto $\mathbb{E}(B_n) = 0$, e quando os grupos diferem temos $\mathbb{E}(B_n) > 0$.

Para evitar maiores complicações computacionais, o problema resume-se em minimizar a função

$$f(G_1, G_2) = -\frac{B_n}{\sqrt{(Var(B_n))}},$$

que também caracteriza o menor p-valor que a configuração pode assumir. De certa forma, se

esse p-valor for menor que um certo nível de significância α então há uma certa "confiança" na conclusão a respeito da separabilidade dos grupos.

3.1 Extensão da estatística de teste para grupos de tamanho 1

Valk e Cybis (2018) propõe explorar o método de *clustering* apresentado em Cybis et al. (2018) para construir um algoritmo de detecção de outliers. Contudo, o método de *clustering* hierárquico não deve ser restrito a *clusters* com tamanhos $g_i \geq 2$. Essa restrição de tamanho de grupo é uma consequência da definição da B_n de um argumento de decomposibilidade de um subgrupo, resultando em somas ponderadas de distâncias entre e dentro de *clusters*.

Para construir um algoritmo de *clustering* que considere grupos de tamanho 1, é proposta uma extensão das estatísticas de teste B_n. Define-se

$$B_n = \begin{cases} \frac{n-1}{n(n-1)}(U^{(1,2)}_{1,n-1} - U^{(2)}_{n-1}) & \text{if } n_1 = 1, \\ \frac{n_1 n_2}{n(n-1)}(2U^{(1,2)}_{n_1 n_2} - U^{(1)}_{n_1} - U^{(2)}_{n_2}) & \text{if } 2 \leq n_1 \leq n-2, \\ \frac{n-1}{n(n-1)}(U^{(1,2)}_{1,n-1} - U^{(1)}_{n-1}) & \text{if } n_1 = n-1, \end{cases} \tag{1}$$

Primeiro notamos que a decomposição apresentada na expressão ainda é válida para o B_n estendida com um grupo de tamanho $2 \leq n_1 \leq n-2$, bem como a decomposição de Hoeffding e a teoria sobre convergência.

O método de Valk e Cybis (2018) está implementado no pacote *uhclust* o qual foi utilizado para as simulações. Cabe ressaltar que nenhuma abordagem a séries temporais foi proposta ainda utilizando esse método.

3.2 Simulações de Monte Carlo

Para verificar o desempenho do método de agrupamento *uclust* quando utilizado em um contexto de séries temporais, propomos um estudo de simulação em que os cenários são controlados. Nesse estudo, sabemos quem são os verdadeiros *clusters* e então podemos verificar a qualidade do método em encontrá-los e também a capacidade de detectar diferença entre os mesmos, quando ela existe.

Assim, utilizamos os processos autorregressivos de ordem 1 (AR(1)) para gerar os grupos. O processo é definido por $Y_t = \phi y_{t-1} + \varepsilon_t$, em que o parâmetro ϕ deve satisfazer $|\phi| < 1$ e ε_t é um ruído branco gaussiano.

Na Tabela 1, as $n_1 = 10$ séries temporais que compõem o grupo 1 (G1) são geradas com

$\phi = 0.3$ (conforme coluna do ϕ_1) e as $n_2 = 7$ séries que compõem o grupo 2 (G2) são geradas a partir de diferentes valores para ϕ (conforme a coluna do ϕ_2). Os resultados mostram a proporção de rejeição em 100 replicações de cada cenário, além de uma medida de "qualidade de cluster"(ARI) proposta por Rand (1971). Dessa forma, a partir do cálculo da ARI, comparamos a qualidade do nosso método com o método clássico de agrupamento hierárquico *hclust* "complete linkage", do pacote *stats* do R.

Sob a hipótese de homogeneidade de grupos, ou seja, que todos os componentes tenham mesma distribuição, que nesse contexto pode ser traduzido para mesmo processo gerador, espera-se que o método não encontre mais do que α% de rejeição, onde α é o nível de significância. Neste estudo, usamos $\alpha = 5\%$ e podemos observar que quando os parâmetros ϕ_1 e ϕ_2 são iguais a 0.3, a proporção de rejeição é muito próxima a 5%, o que indica que o método está bem "calibrado", não rejeitando mais do que α. A medida em que ϕ_1 se diferencia de ϕ_2, a proporção de rejeição aumenta, indicando que o método detecta dois grupos.

Além disso, é importante ressaltar que, quando $n_1 = 10$ e $n_2 = 7$, o ARI do método *uhclust* é melhor que o tradicional *hclust*. No entanto, em um segundo cenário em que $n_1 = 10$ e $n_2 = 1$, o ARI do método tradicional *hclust* é mais satisfatório, como mostra a Tabela 2.

$n_1 = 10$ e $n_2 = 7$				
ϕ_1	ϕ_2	Proporção de Rejeição	ARI *hclust*	ARI *uhclust*
0.30	−0.20	1.00	0.99	1.00
0.30	−0.10	1.00	0.77	0.99
0.30	0.00	0.97	0.46	0.88
0.30	0.10	0.24	0.11	0.42
0.30	0.20	0.05	0.02	0.06
0.30	0.30	0.04		
0.30	0.40	0.08	0.02	0.06
0.30	0.50	0.53	0.17	0.61
0.30	0.70	1.00	0.99	1

Tabela 1: Proporção de rejeição do *uhclust* e ARI do *uhclust* e *hclust*

$n_1 = 10$ e $n_2 = 1$				
ϕ_1	ϕ_2	Proporção de Rejeição	ARI *hclust*	ARI *uhclust*
0.30	−0.20	0.18	0.77	0.35
0.30	−0.10	0.1	0.54	0.20
0.30	0.00	0.08	0.26	0.05
0.30	0.10	0.07	0.17	0.04
0.30	0.20	0.03	0.02	0.01
0.30	0.30	0.06		
0.30	0.40	0.03	0.05	0.01
0.30	0.50	0.06	0.14	0.04
0.30	0.70	0.41	0.87	0.61

Tabela 2: Proporção de rejeição do *uhclust* e ARI do *uhclust* e *hclust*

4 Resultados

Durante a realização do presente trabalho, exploramos vários bancos de dados de diferentes fontes e características, e neles aplicamos diversas transformações na busca por padrões. Simulações de Monte Carlo foram realizadas em um contexto controlado e sugerem que o método *uhclust* pode ser usado para caracterizar sinais com dinâmicas diferentes desde que a métrica correta seja utilizada. Os próximos passos serão na direção da aplicação aos dados reais apresentados nesse trabalho.

Referências

Chandola, V., Banerjee, A., e Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):15.

Cybis, G. B., Valk, M., e Lopes, S. R. (2018). Clustering and classification problems in genetics through u-statistics. *Journal of Statistical Computation and Simulation*, pages 1–21.

Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P., Mark, R., Mietus, J., Moody, G., Peng, C.-K., Stanley, H., PhysioBank, PhysioToolkit, e PhysioNet (2000). Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):215.

Liu, C., Springer, D., Li, Q., Moody, B., Juan, R. A., Chorro, F. J., Castells, F., Roig, J. M., Silva, I., Johnson, A. E. W., Syed, Z., Schmidt, S. E., Papadaniil, C. D., Hadjileontiadis, L., Naseri, H., Moukadem, A., Dieterlen, A., Brandt, C., Tang, H., Samieinasab, M., Samieinasab, M. R., Sameni, R., Mark, R. G., e Clifford, G. D. (2016). An open access database for the evaluation of heart sound algorithms. *Physiological Measurement*, 37(12):2181.

Pimentel, M. A., Clifton, D. A., Clifton, L., e Tarassenko, L. (2014). A review of novelty detection. *Signal Processing*, 99(Supplement C):215 – 249.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850.

Tax, D. M. e Duin, R. P. (2004). Support vector data description. *Machine learning*, 54(1):45–66.

Valk, M. e Cybis, G. B. (2018). U-statistical inference for hierarchical clustering. *arXiv preprint arXiv:1805.12179*.

Um novo modelo probabilístico para dados restritos ao intervalo unitário

Tatiane Fontana Ribeiro [1] [3]

Renata Rojas Guerra [2] [3]

Fernando Arturo Peña - Ramírez [3] [3]

Pierre Louis Termidor [4] [4]

Resumo: São inúmeras as situações nas quais o objeto de estudo consiste em variáveis com suporte no intervalo unitário. Dentre as quais citam-se: taxas, proporções e índices. Embora possa ser utilizado, nesses casos, o modelo clássico: distribuição beta e outros já existentes na literatura, é importante dispor de outros modelos probabilísticos alternativos. Neste contexto, objetiva-se propor uma nova distribuição de probabilidade unitária, bem como estudar algumas de suas características estatísticas e matemáticas e estimar seus parâmetros via máxima verossimilhança. Para tanto, propõe-se uma transformação em uma dada variável aleatória que limita a imagem da nova variável obtida ao intervalo $(0, 1)$. Foi avaliado o desempenho dos estimadores de máxima verossimilhança em amostras de tamanho finito através de simulações de Monte Carlo. Obtiveram-se resultados razoáveis em termos de acurácia e precisão das estimativas, mesmo para amostras de tamanho 20.

Palavras-chave: *Distribuição Bur XII, Distribuições Unitárias, Estimação de máxima verossimilhança, Simulação de Monte Carlo*.

1 Introdução

A distribuição Burr XII (BXII) faz parte de um sistema de distribuições derivadas por Burr [1]. Por ser um modelo com suporte nos reais positivos, esta distribuição tem sido amplamente utilizada no contexto de economia como uma alternativa na modelagem de dados associados à renda. Algumas aplicações desenvolvidas nesse contexto foram apresentadas nos estudos empíricos de Kleiber e Kotz [5], as quais foram realizadas principalmente na segunda metade do século XX.

Recentemente, muitos pesquisadores utilizam a distribuição BXII em diferentes campos da ciência, sendo a maioria com ênfase em situações-modelos caracterizadas pelo comportamento das leis de potência. Além disso, Paranaíba [6] destaca que esta distribuição possui flexibilidade no ajuste de dados

[1]UFSM - Universidade Federal de Santa Maria. Email: *tatianefontanaribeiro@gmail.com*
[2]UFSM - Universidade Federal de Santa Maria. Email: *renata.r.guerra@ufsm.br*
[3]UFSM - Universidade Federal de Santa Maria. Email: *fernando.p.ramirez@ufsm.br*
[4]UFSM - Universidade Federal de Santa Maria. Email: *pierrelouis.termidor@gmail.com*
[4]Agradecimento: FIPE - CCNE

já que apresenta alguns casos particulares como as distribuições: normal, log-normal, gama, logística, valor extremo tipo I e outras.

Uma parametrização alternativa à distribuição Burr XII é tomar o parâmetro de escala igual a um. Esta parametrização é empregada por vários autores, da qual obtém-se a distribuição Burr XII biparamétrica, mais conveniente em algumas aplicações livres de escala [2].

Uma variável aleatória X contínua e positiva, segue a distribuição BXII biparamétrica com parâmetros $c > 0$, $d > 0$ se sua função densidade de probabilidade (fdp) é dada por

$$f_X\left(x \mid c, d\right) = cd\frac{x^{c-1}}{\left[1 + x^c\right]^{d+1}}, \quad x > 0, \tag{1}$$

em que c e d são parâmetros de forma. Nesse caso, a função de distribuição acumulada (fda) de (1) é dada por

$$F_X\left(x \mid c, d\right) = 1 - \left[1 + x^c\right]^{-d}. \tag{2}$$

A distribuição BXII biparamétrica também acomada outras distribuições de probabilidade para valores particulares dos parâmetros c e d. Para $c = 1$, tem-se a distribuição Pareto Tipo II e quando $d = 1$ tem-se um caso particular da distribuição *Champernowne*. Esta distribuição também pertence à família Weibull extendida proposta por Gurvich, DiBenedettos e Ranad [4].

O suporte de (1) é os reais positivos. Contudo, há inúmeros casos nos quais a variável aleatória de interesse só pode assumir valores pertencentes ao intervalo unitário, tais como variáveis relacionadas a taxas, proporções e índices. Na modelagem de dados deste tipo, a distribuição beta é a mais utilizada. Contudo, é necessário dispor de distribuições de probabilidade unitárias alternativas que podem se ajustar melhor em determinadas situações.

Com intuito de possibilitar flexibilidade à modelagem de variáveis aleatórias com suporte no intervalo $(0, 1)$, neste trabalho propõe-se uma nova distribuição de probabilidade: distribuição Burr XII unitária ($\mathcal{UBXII}$). O novo modelo é obtido a partir de uma transformação em uma variável aleatória que segue distribuição BXII, sem a necessidade da acrescentar novos parâmetros ao modelo base. São apresentadas algumas propriedades estatísticas e matemáticas da nova distribuição. Além disso, são obtidos os estimadores de máxima verossimilhança (EMVs) através da log-verossimilhança perfilada. Também é realizado um estudo de simulação para avaliar o desempenho dos EMVs em amostras de tamanho finito.

2 Materiais e Métodos

Seja X a variável aleatória que segue uma distribuição BXII biparamátrica, com fdp e fda dadas por (1) e (2), respectivamente. Considera-se a transformação $Y = \mathrm{e}^{-X}$ da qual deriva-se a nova distribuição unitária. Desta forma, a fda do modelo $\mathcal{UBXII}$ é dada por

$$F_Y\left(y \mid c, d\right) = \left[1 + \left(-\log y\right)^c\right]^{-d}, \qquad 0 < y < 1. \tag{3}$$

Assim como no modelo base, tem-se que $c, d > 0$ são parâmetros de forma. Derivando (3) obtém-se a fdp dada por

$$f_Y\left(y \mid c, d\right) = \frac{cd\left(-\log y\right)^{c-1}}{y\left[1 + \left(-\log y\right)^c\right]^{d+1}}. \tag{4}$$

Na Figura 1 são expressos gráficos da fdp (4) para alguns valores de c e d.

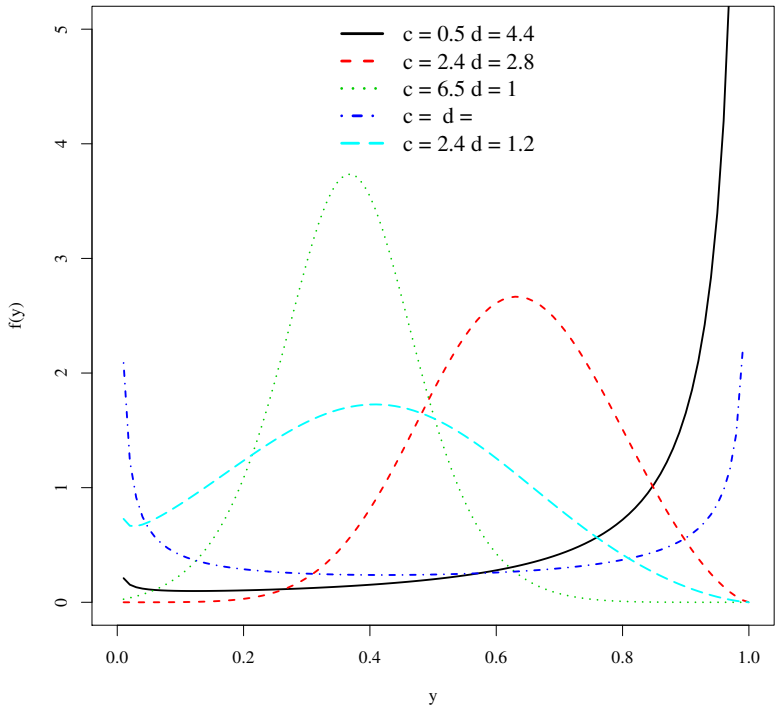


Figura 1: Gráficos da fdp do modelo $\mathcal{UBXII}$

A densidade da distribuição $\mathcal{UBXII}$ pode tomar diversas formas. Conforme a Figura 1, a fdp (4) pode ser assimétrica à esquerda ou a direita, unimodal, possuir formato de J ou de U. Consequentemente, pode-se dizer que o modelo proposto consiste em uma distribuição de probabilidade flexível, capaz de

acomodar diversos formatos de variáveis com suporte unitário.

3 Resultados e Discussões

Nesta seção são apresentados os principais resultados do presente trabalho. Apresentam-se algumas quantidades estatísticas e matemáticas da distribuição $\mathcal{UBXII}$ proposta, tais como função quantílica e momentos ordinários. Também são obtidos os os EMVs, cujo desempenho para amostras finitas é avaliado em sete diferentes cenários via simulação de Monte Carlo.

3.1 Função quantílica

A função quantílica é obtida tomando-se a inversa da função (3). Assim, é dada por

$$Q_Y(u) = \exp\left\{-\left(u^{-\frac{1}{d}} - 1\right)^{\frac{1}{c}}\right\}. \tag{5}$$

Os quantis da distribuição $\mathcal{UBXII}$ podem ser determinados a partir de (5) susbtituindo-se adequadamente os valores de u. [2]. Em particular, tomando $u = 0,5$ obtém-se a mediana deste modelo. Os coeficientes de assimetria e curtose também podem ser obtidos de (5). Além disso, por meio do método da inversão, é possível gerar ocorrências pseudo-aleatórias desta distribuição. Para isso, considera-se que se tenha um bom gerador de uniformes, em que U é uma variável aleatória contínua pertencente ao intervalo $(0, 1)$. Avaliando (5) em U tem-se $X = Q(U)$ que segue uma distribuição $\mathcal{UBXII}$.

3.2 Momentos ordinários

O h-ésimo momento ordinário de Y é determinado por

$$E\left(Y^h\right) = cd\int_0^1 y^{h-1}(-\log y)^{c-1}\left[1 + (-\log y)^c\right]^{-d-1}dy. \tag{6}$$

Considerando a troca de variáveis $u = -\log y$. A integral (6) pode ser escrita como

$$E\left(Y^h\right) = cd\int_0^\infty e^{-uh}u^{c-1}\left(1 + u^c\right)^{-d-1}du.$$

Usando a expansão binomial, tem-se que o h-ésimo momento é dado por

$$E\left(Y^h\right) = cd\sum_{k=0}^{-d-1}\binom{-d-1}{k}h^{-c(k+1)}\Gamma\left[c\left(k+1\right)\right]. \tag{7}$$

De (7) são obtidas a esperança e a variância de Y, respectivamente, dadas por

$$E\left(Y\right) = cd \sum_{k=0}^{-d-1} \binom{-d-1}{k} \Gamma\left[c\left(k+1\right)\right]$$

e

$$Var\left(Y\right) = cd \sum_{k=0}^{-d-1} \binom{-d-1}{k} (2)^{-c(k+1)} \Gamma\left[c\left(k+1\right)\right] - \left\{ cd \sum_{k=0}^{-d-1} \binom{-d-1}{k} \Gamma\left[c\left(k+1\right)\right] \right\}^2.$$

3.3 Estimação via máxima verossimilhança

Seja $Y_1, \cdots, Y_n$ uma amostra aleatória de tamanho n da distribuição $\mathcal{UBXII}\ (c, d)$, em que o vetor de parâmetros é: $\boldsymbol{\theta} = (c, d)^T$. A funçao log-verossimilhança é expressa por

$$\ell\left(\boldsymbol{\theta} \mid \mathbf{y}\right) = n \log\left(cd\right) - \sum_{i=1}^{n} \log y_i + (c-1) \sum_{i=1}^{n} \log\left(-\log y_i\right) - (d+1) \sum_{i=1}^{n} \log\left[1 + (-\log y_i)^c\right]. \tag{8}$$

É possível obter os EMVs maximizando, diretamente, a função (8). Todavia, de forma alternativa, pode-se obter os vetores escores igualá-los a zero e solucionar o sistema de equações decorrente, obtendo a expressão para cada estimador que torna ambas as equações simultaneamente verdadeiras. Deste modo, os componentes do vetor escore $\mathbf{U}\left(\boldsymbol{\theta}\right)$ são dados por

$$U_c\left(\boldsymbol{\theta}\right) = \frac{n}{c} + \sum_{i=1}^{n} \log\left(-\log y_i\right) - \frac{(d+1) \sum_{i=1}^{n} \left(-\log y_i\right)^c \log\left(-\log y_i\right)}{n + \sum_{i=1}^{n} \left(-\log y_i\right)^c}$$

e

$$U_d\left(\boldsymbol{\theta}\right) = \frac{n}{d} - \sum_{i=1}^{n} \log\left[1 + (-\log y_i)^c\right].$$

Verifica-se que nenhum dos EMVs possui forma fechada. Mas é fácil notar que, para c fixo, tem-se a forma semi-fechada do EMV do parâmetro d, dada por

$$\hat{d}\left(\hat{c}\right) = \frac{n}{\sum_{i=1}^{n} \log\left[1 + \left(-\log y_i\right)^{\hat{c}}\right]}. \tag{9}$$

Substituindo (9) em (8) obtém a função log-verossimilhança perfilada dada por

$$\ell(c \mid \mathbf{y}) = n\log(nc) - \sum_{i=1}^{n}\log(y_i) + (c-1)\sum_{i=1}^{n}\log(-\log y_i) - \sum_{i=1}^{n}\log\left[1+(-\log y_i)^c\right] - n\log\left(\sum_{i=1}^{n}\log\left[1+(-\log y_i)^c\right]\right) - n. \quad (10)$$

3.4 Simulação de Monte Carlo

Nesta seção são apresentados os resultados da Simulação de Monte Carlo realizada para avaliar o desempenho dos estimadores do novo modelo unitário proposto. As simulações foram realizadas no software R. Optou-se por maximizar a função log-verossimilhança perfilada dada em (10). Para tanto, utilizou-se a rotina *optim* com o algortimo de otimização não linear BFGS quasi-Newton.

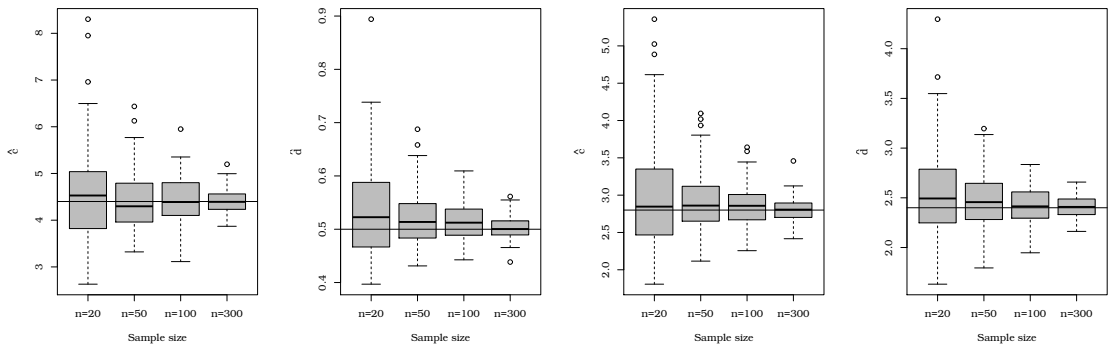
As ocorrências $y_1, \cdots, y_n$ da distribuição $\mathcal{UBXII}$ foram obtidas pelo método da inversão, utilizando (5). Foram simuladas 10.000 réplicas de Monte Carlo para amostras de tamanho 20, 50, 100 e 300 e para sete combinações diferentes do vetor de parâmetros $\boldsymbol{\theta}$, escolhidas de modo a acomodar vários formatos da densidade dada em (4).

Na Tabela 1 são exibidos os resultados obtidos a partir estudo de simulação. É apresentada a média, a raiz quadrada do erro quadrático médio (REQM) e o viés relativo percentual (VR%) dos EMVs da distribuição $\mathcal{UBXII}$.

A Figura 2 ilustra a convergência das estimativas dos parâmetros da distribuição $\mathcal{UBXII}$ para as 100 primeiras réplicas de Monte Carlo e os quatro tamanhos amostrais considerados. Nesta evidencia-se que a presença de observações discrepantes superestimam os verdadeiros valores dos parâmetros. Porém, a medida que o tamanho da amostra aumenta, a quantidade de *outliers* diminui e a convergência da estimativa para o verdadeiro valor do parâmetro aumenta. Assim, quanto maior o tamanho da amostra, mais precisa é esta estimativa, fato justificado pelas propriedades assintóticas dos EMVs.

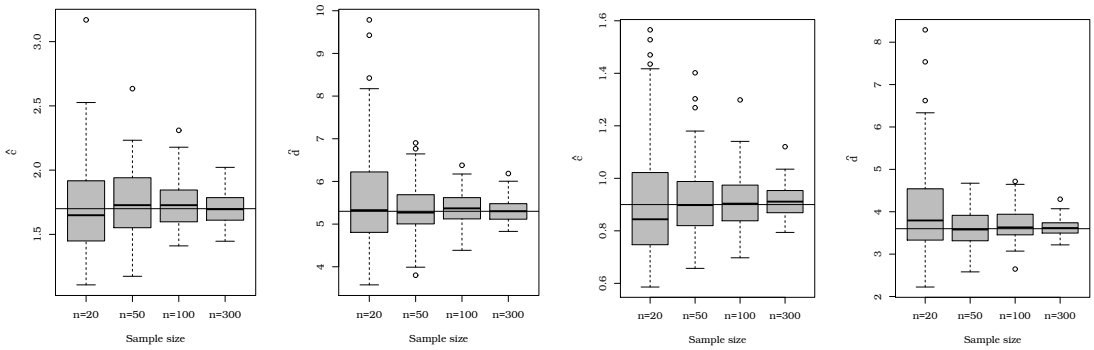
3.5 Conclusão

Destaca-se que além da obtenção de melhor precisão da estimativa via maximização de (10), o custo computacional é reduzido, uma vez que a função log-verossimilhança perfilada envolve apenas um parâmetro. Observa-se que o desempenho dos EMVs foi muito bom. Conforme esperado, à medida que o tamanho da amostra aumenta, observou-se um melhor desempenho em termos de acurácia e precisão dos estimadores de máxima verossimilhança do modelo $\mathcal{UBXII}$. O novo modelo probabilístico, portanto, pode ser utilizado na modelagem de variáveis aleatórias limitados ao intervalo unitário como alternativa às distribuições unitárias já existentes na literatura, caso se ajuste melhor ao conjunto de dados



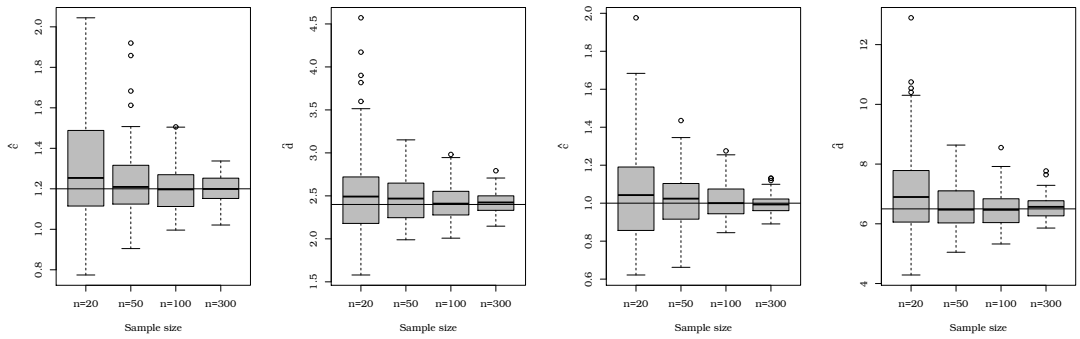
(a) Cenário 1

(b) Cenário 2



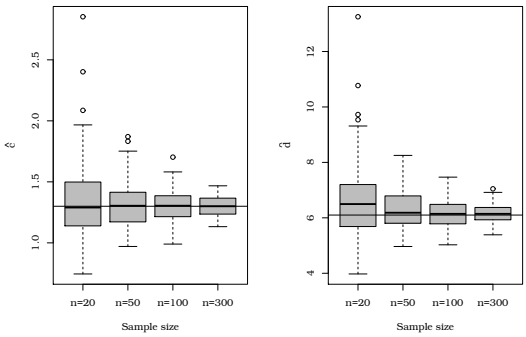
(c) Cenário 3

(d) Cenário 4



(e) Cenário 5

(f) Cenário 6



(g) Cenário 7

Figura 2: Box-plot para as estimativas dos parâmetros da $\mathcal{UBXII}$ considerando as 100 primeiras réplicas de Monte Carlo e os tamanhos amostrais $n = 20, 50, 100$ e 300 para sete cenários distintos.

Tabela 1: Resultados da simulação de Monte Carlo para o modelo $\mathcal{UBXII}$ considerando 10.000 réplicas e amostras de tamanho $n = 20, 50, 100$ e 300.

Cenário	c	d	n	Média		REQM		VR%	
				$\hat{c}$	$\hat{d}(\hat{c})$	$\hat{c}$	$\hat{d}(\hat{c})$	$\hat{c}$	$\hat{d}(\hat{c})$
1	0.5	4.4	20	0.5325	4.6180	0.0992	1.1165	6.5013	4.9546
			50	0.5119	4.4847	0.0554	0.6602	2.3806	1.9253
			100	0.5053	4.4437	0.0372	0.4538	1.0633	0.9925
			300	0.5018	4.4162	0.0209	0.2538	0.3668	0.3675
2	2.4	2.8	20	2.5475	2.9490	0.4795	0.7090	6.1444	5.3207
			50	2.4530	2.8600	0.2673	0.4189	2.2080	2.1434
			100	2.4290	2.8286	0.1843	0.2843	1.2065	1.0230
			300	2.4112	2.8091	0.1040	0.1645	0.4660	0.3258
3	5.3	1.7	20	5.6639	1.7923	1.1466	0.4352	6.8659	5.4270
			50	5.4325	1.7339	0.6477	0.2534	2.5001	1.9959
			100	5.3666	1.7189	0.4403	0.1741	1.2557	1.1130
			300	5.3178	1.7040	0.2465	0.0981	0.3361	0.2336
4	3.6	0.9	20	3.9032	0.9495	1.0248	0.2297	8.4211	5.4951
			50	3.7168	0.9189	0.5294	0.1318	3.2435	2.0945
			100	3.6642	0.9096	0.3659	0.0924	1.7824	1.0705
			300	3.6196	0.9030	0.2008	0.0521	0.5453	0.3298
5	2.4	1.2	20	2.5800	1.2620	0.5851	0.3067	7.5011	5.1666
			50	2.4679	1.2267	0.3198	0.1783	2.8304	2.2210
			100	2.4379	1.2119	0.2206	0.1237	1.5796	0.9958
			300	2.4107	1.2044	0.1217	0.0696	0.4469	0.3632
6	6.5	1.0	20	7.0382	1.0569	1.7548	0.2566	8.2801	5.6898
			50	6.6971	1.0204	0.9297	0.1477	3.0328	2.0360
			100	6.5951	1.0093	0.6211	0.1023	1.4630	0.9329
			300	6.5332	1.0037	0.3526	0.0580	0.5106	0.3740
7	6.1	1.3	20	6.5525	1.3625	1.4564	0.3239	7.4175	4.8057
			50	6.2886	1.3277	0.8157	0.1941	3.0915	2.1344
			100	6.1730	1.3112	0.5383	0.1324	1.1973	0.8635
			300	6.1261	1.3042	0.3053	0.0759	0.4281	0.3252

considerado.

Referências

[1] BURR, I. W. Cumulative frequency functions. *Annals of Mathematical Statistics*, 13, 215 - 232, 1942.

[2] GUERRA, R. R. *Some generalized BXII distributions with applications to income and lifetime data*. 2017. 119 p., Thesis, Universidade Federal de Pernambuco, Recife, 2017.

[3] GUERRA, R. R.; PEÑA-RAMÍREZ, F. A.; BOURQUIQNONB, M. The unit extended Weibull family of distributions and its applications. *Journal of Applied Statistics*. Submetido, 2018.

[4] GURVICH, M. R.; DiBENEDETTOS, A. T.; RANADE, S. V. A new statistical distribution for characterizing therandom strength of brittle materials. *Journal of Materials Science*, v. 32, p. 2559-2564, 1997.

[5] KLEIBER, C; KOTZ, S. *Statistical Size distribution in Economics and Actuarial Sciences*. John Wiley, New Jersey, 2003.

[6] PARANAÍBA, P. F. *Caracterização e extensões da distribuição Burr XII: propriedades e aplicações*. 2011. 142 p., Tese, Universidade de São Paulo, Piracicaba, 2011.

Estudo simulado envolvendo Cartas de Controle Multivariadas

Eduardo de Oliveira Correa [1]

Danilo Marcondes Filho[2]

Resumo: Processos industriais geram dados acerca de inúmeras variáveis de interesse correlacionadas. Buscando um monitoramento mais robusto de tais processos,cartas de controle baseados em técnicas estatísticas multivariadas foram desenvolvidos. Destacam-se as cartas de controle *Qui-Quadrado* (χ^2) e da *Variância Generalizada (W)*. Estas estatísticas permitem um monitoramento simultâneo do vetor de médias e da matriz de covariâncias das variáveis, respectivamente, a cada nova amostra do processo. Este trabalho apresenta um estudo por simulação para investigar o poder de detecção das cartas χ^2 e *W*. A partir de um processo simulado com 4 variáveis e uma estrutura de covariância, descontroles são impostos tanto no vetor de médias quanto na matriz de covariâncias do processo sob controle. Os resultados mostram que a sensibilidade da carta *W* aumenta para a detecção de modificações maiores na estrutura de covariância original das variáveis. Já em relação à carta χ^2, podemos notar que alterações no vetor de médias nas direções comuns de variância das variáveis (isto é, na direção das suas covariâncias) são detectadas com menos sensibilidade em relação às alterações que não estão nas suas direções de covariância.

Palavras-chave: *Cartas de Controle Multivariadas, Carta de Controle Qui-Quadrado, Carta de Controle da Variância Generalizada*.

1 Introdução

Com o avanço tecnológico e uma disputa mercadológica extremamente competitiva, tem-se aumentado o interesse das indústrias no estudo dos métodos estatísticos para controle de processos. O *Controle Estatistico do Processo* (CEP) consiste em um grupo de ferramentas desenvolvidas para monitorar o desempenho de um processo, sendo as *cartas de controle* (CCs) possivelmente a ferramenta mais sofisticada; ver (Montgomery, 2007).

As CCs foram introduzidas por Shewhart em 1924, buscando entender as causas que provocam variabilidades no processo. Segundo este autor, a variabilidade pode ocorrer por *causas comuns* (variações

[1]UFRGS - Universidade Federal do Rio Grande do Sul. Email: *eduardo.correa@ufrgs.br*
[2]UFRGS - Universidade Federal do Rio Grande do Sul. Email: *marcondes.danilo@gmail.com*

aleatórias inerentes ao processo), e por *causas especiais* (eventos destoantes no processo que prejudicam a qualidade do produto). Através das CCs busca-se monitorar a variabilidade existente nos processos, procurando detectar a possível presença de causas especiais. Eliminando as causas especiais, consegue-se obter a redução sistemática da variabilidade do processo, aprimorando a qualidade, produtividade, confiabilidade e o custo do produto. A CC é uma ferramenta gráfica onde medidas de amostras igualmente espaçadas no tempo são representadas cronologicamente e cujos limites de controle são obtidos a partir de amostras preliminares do processo sob controle estatístico (isto é, apenas causas comuns presentes). Destacam-se as tradicionais cartas de controle univariadas para monitoramento da tendência central dos dados (*Carta de Controle para a Média*) e para o monitoramento da variabilidade (*Carta de Controle para Amplitude*). O trabalho percursor de Shewhart está sumarizado em (Shewhart, 1931).

Processos mais complexos geram uma grande massa de dados acerca de inúmeras de variáveis correlacionadas, tornando inadequado o uso das cartas de controles univariadas tradicionais. Neste caso, versões multivariadas das cartas mencionadas foram desenvolvidas. As CCs *Qui-Quadrado* para minitoramento do vetor de médias e da *Variância Generalizada* para o monitoramento da matriz de covariâncias permitem o monitormanto simultâneo de um conjunto de variáveis e oferecem performance superior as suas versões univariadas.

Este trabalho apresenta um estudo simulado para avaliar o desempenho das cartas de controle multivariadas *Qui-Quadrado* e *Variância Generalizada*. Para tanto, considerando um processo com 4 variáveis sobre interesse, cenários representando descontroles impostos no vetor de médias e na matriz de covariâncias do processo serão investigados.

2 Cartas de Controle Multivariadas

As primeiras publicações na perspectiva multivariada foram feitas por Harold Hotelling [(Hotelling, 1947)], utilizando abordagem multivariada em dados contendo informações sobre bombardeios durante a Segunda Guerra Mundial. Esta seção descreve brevemente a base teórica das tradicionais cartas de controle *Qui-Quadrado* e da *Variância Generalizada*.

2.1 Carta de controle Qui-Quadrado

Considera-se p-características correlacionadas medidas simultaneamente compondo amostras p-variadas de tamanho n. Supõe-se que estas características seguem uma distribuição p-dimensional multivariada normal com vetor de médias $\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_p)$ e matriz de covariância $\boldsymbol{\Sigma}$, sendo μ_i a média para a *i-ésima* característica e $\boldsymbol{\Sigma}$ uma matriz consistindo de variâncias e covariâncias das p-características, onde os elementos da diagonal principal são as variâncias de x's e os elementos fora da diagonal principal representam as covariâncias. O monitoramento futuro de vetores p-variados de tamanho n é dado por:

$$\chi_p^2 = n(\overline{\mathbf{x}} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\overline{\mathbf{x}} - \boldsymbol{\mu}) \ \sim \ \chi_p^2 \tag{2.1}$$

onde χ_p^2 segue uma distribuição *Qui-Quadrado* com p graus de liberdade e representa a distância quadrada padronizada, p-dimensional, entre um vetor de observações $\overline{\mathbf{x}}$ e o vetor de médias do processo $\boldsymbol{\mu}$. A raiz quadrada χ_p é conhecida como *distância de Mahalanobis*. Na prática, é necessário estimar $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$ a partir de amostrar preliminares, tomadas quando se assume que o processo está sob controle. $\overline{\mathbf{x}}$ e $\mathbf{S}$ representam, respectivamente, as estimativas para o vetor das médias e a matriz de covariância do processo. Entretanto, neste trabalho não necessitamos estimar tais parâmetros, visto que será utilizado um estudo de caso simulado, onde conhecemos inteiramente não só a distribuição geradora dos dados, como o vetor de médias e a matriz de covariâncias populacionais. Mais detalhes sobre a construção da carta de controle utilizando as estimativas da média e da covariância podem ser encontrados em (Johnson & Wichern, 2007).

2.2 Carta de controle para variabilidade do processo

Considere novamente um conjunto de observações p-variadas geradas de uma distribuição normal p-variada com vetor de médias $\boldsymbol{\mu}$ e matriz de covariâncias $\boldsymbol{\Sigma}$. A carta de controle da *Variância Generalizada* é empregada para detectar mudanças na estrutura de covariâncias dos dados. Esta carta se constitui numa extensão multivariada da carta de controle univariada S^2. (Montgomery, 2007) descreve a estatística para monitorar a estrutura de covariâncias de futuras amostras baseada no determinante de matrizes de covariância, como se segue:

$$W_i = -pn + pnln(n) - nln(|\boldsymbol{A}_i|/|\boldsymbol{\Sigma}|) + tr(\boldsymbol{\Sigma^{-1}} \boldsymbol{A}_i) \ \sim \ \chi^2_{p(p+1)/2} \tag{2.2}$$

onde W_i segue uma distribuição *Qui-Quadrado* com $p(p+1)/2$ graus de liberdade, $\boldsymbol{\Sigma}$ é a matriz de covariância populacional, $\boldsymbol{A_i} = (n-1)\boldsymbol{S}_i$, $\boldsymbol{S}_i$ é a matriz de covariância da *i-ésima* amostra de tamanho *n* e *tr* é o operador de traço.

3 Metodologia

Consideramos um processo simulado sob controle com quatro variáveis. Suponha que os dados seguem uma distribuição normal 4-variada com vetor de médias $\boldsymbol{\mu}$ e a matriz das covariâncias $\boldsymbol{\Sigma}$, dados por:

$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.9 & 0.07 & 0.05 \\ 0.9 & 1 & 0.04 & 0.03 \\ 0.07 & 0.04 & 1 & 0.9 \\ 0.05 & 0.03 & 0.9 & 1 \end{bmatrix} \quad (3.1)$$

Consideraremos no monitoramento vetores de médias amostrais 4-variados de tamanho $n = 50$ observações. Para avaliação do desempenho da carta de controle χ^2, em cada cenário simulado de descontrole são gerados 100 vetores de médias e replicados 500 vezes. Novas amostras são descontroladas no vetor de médias. Para cada amostra os escores referentes à estatística χ^2 são comparados ao limite de controle correspondente [equação (2.1)].

Considere descontroles realizados para três casos distintos: **(I)** monitoramento do processo simulando o descontrole no vetor de médias para duas variáveis correlacionadas, mantendo as outras duas variáveis com média fixa em 0. Escolhemos deslocar para esse caso a média da primeira (μ_1) e segunda (μ_2) variável, dado a alta correlação entre elas ($\rho_{12} = 0.9$); **(II)** monitoramento do processo simulando o descontrole para duas variáveis fracamente correlacionadas, com médias fixas em 0 para as outras duas variáveis. Escolhemos a média da primeira (μ_1) e quarta (μ_4) variável, com correlação $\rho_{14} = 0.05$; **(III)** monitoramento simulando o descontrole para duas variáveis correlacionadas conjuntamente com outra variável que apresenta fraca correlação com as duas, sendo mantida a outra variável com média 0. Escolhemos a primeira e segunda, dado a alta correlação entre elas, juntamente com a quarta variável, que apresenta fraca correlação com as duas primeiras ($\rho_{14} = 0.05$ e $\rho_{24} = 0.03$, respectivamente). Temos assim:

Caso **(I)** Caso **(II)** Caso **(III)**

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ 0 \\ 0 \end{bmatrix} \qquad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ 0 \\ 0 \\ \mu_4 \end{bmatrix} \qquad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ 0 \\ \mu_4 \end{bmatrix} \quad (3.2)$$

onde $\mu_1 = \mu_2 = \mu_3 = 0, 0.1, 0.2, 0.3, 0.4, ..., 2.9, 3$. Em cada cenário variamos na razão de 0.1 na média para cada variável descontrolada.

Para o monitoramento da variabilidade consideramos a matriz de covariância do processo sob con-

trole dado na equação (3). Para avaliar o desempenho da carta de controle W, em cada cenário simulado de descontrole, novamente foram gerados 100 vetores de médias de 50 amostras e computada a matriz de covariâncias associada a cada vetor. Os cenários foram replicados 500 vezes. Essa estrutura de covariância é descontrolada para novas amostras. Para cada amostra os escores referentes à estatística W_i são comparados ao limite de controle correspondente [equação (2.2)]. Serão escolhidos dois casos a serem simulados: **(I)** monitoramento do processo simulando o descontrole na estrutura de covariância para duas variáveis altamente correlacionadas. Escolhemos para esse caso a primeira e segunda variável, cuja correlação é de $\rho_{12} = 0.9$; **(II)** monitoramento do processo simulando o descontrole na primeira e na quarta variável, cuja a correlação é fraca ($\rho_{14} = 0.05$). Temos assim:

Caso **(I)** Caso **(II)**

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho_{12} & 0.07 & 0.05 \\ \rho_{21} & 1 & 0.04 & 0.03 \\ 0.07 & 0.04 & 1 & 0.9 \\ 0.05 & 0.03 & 0.9 & 1 \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.9 & 0.07 & \rho_{14} \\ 0.9 & 1 & 0.04 & 0.03 \\ 0.07 & 0.04 & 1 & 0.9 \\ \rho_{41} & 0.03 & 0.9 & 1 \end{bmatrix} \quad (3.3)$$

onde ρ_{12} = ρ_{21} = 0.6, 0.61, 0.63, 0.63, 0.64, . . . , 0.98, 0.99 e ρ_{14} = ρ_{41} = -0.14, -0.13, -0.12, -0.11, -0.10, . . . , 0.24, 0.25. Em cada cenário variamos na razão de 0.01 a correlação nas variáveis descontroladas.

4 Estudo de casos simulados

Nessa seção apresentamos os resultados das simulações dos cenários descritos da seção anterior. Em cada tabela apresentamos os resultados de ambas as cartas em função da média e do desvio-padrão entre replicações. O limite de controle das cartas χ^2 e *W* foram obtidos considerando probabilidade de alarme falso de $\alpha = 0.05$.

A tabela 1 apresenta os resultados das simulações para os três casos descritos na seção anterior para avaliação da carta χ^2. Observamos no caso **(I)** que a detecção de descontroles cresce conforme o tamanho do descontrole imposto. Entretanto, a carta apresenta baixa sensibilidade de detecção, pelo fato do descontrole representar uma alteração na direção comum de variabilidade destas variáveis. No caso **(II)** observamos acentuada sensibilidade na detecção de descontroles comparado ao caso **(I)**. Isto se deve a baixa correlação nas variáveis descontroladas, dado que esses descontroles estão em direções opostas as direções comuns de variabilidade. Dessa forma, mesmo pequenas alterações são detectadas com

frequência relativa alta. Observamos no caso **(III)** a boa sensibilidade da carta na detecção de descontroles . Notamos que a sensibilidade na detecção de pequenos descontroles é superior ao caso mostrado no caso **(I)** e inferior descrito no caso **(II)**. Isto se justifica plenamente pelo fato de que neste cenário simulamos descontroles simultâneos em duas variáveis fortemente correlacionadas entre si e uma terceira fracamente correlacionada com as demais. Dessa forma, este descontrole está numa direção de variabilidade próxima (não oposta) a direção comum entre as duas variáveis correlacionadas.

Tabela 1: Quantidade média (e desvio padrão) de amostras perturbadas identificadas pela carta χ^2 para cada caso.

Descontroles	caso (I)		caso (II)		caso (III)	
	média	desvio padrão	média	desvio padrão	média	desvio padrão
0.0	0.054	0.022	0.052	0.024	0.048	0.022
0.1	0.049	0.027	0.057	0.021	0.054	0.024
0.2	0.055	0.022	0.071	0.023	0.063	0.023
0.3	0.052	0.020	0.106	0.033	0.082	0.027
0.4	0.057	0.024	0.155	0.035	0.110	0.032
0.5	0.059	0.022	0.231	0.046	0.143	0.035
0.6	0.071	0.024	0.322	0.050	0.195	0.040
0.7	0.078	0.027	0.431	0.049	0.250	0.042
0.8	0.087	0.025	0.542	0.048	0.327	0.046
0.9	0.096	0.030	0.661	0.042	0.411	0.048
1.0	0.110	0.032	0.760	0.045	0.502	0.049
1.1	0.122	0.036	0.852	0.036	0.592	0.051
1.2	0.139	0.033	0.909	0.028	0.680	0.047
1.3	0.154	0.037	0.952	0.021	0.755	0.044
1.4	0.176	0.036	0.977	0.016	0.825	0.039
1.5	0.201	0.039	0.988	0.011	0.880	0.033
1.6	0.222	0.039	0.995	0.007	0.920	0.028
1.7	0.252	0.047	0.998	0.004	0.949	0.023
1.8	0.275	0.043	0.999	0.003	0.970	0.017
1.9	0.302	0.044	1	0	0.984	0.013
2.0	0.336	0.047	1	0	0.991	0.010
2.1	0.376	0.046	1	0	0.995	0.007
2.2	0.405	0.047	1	0	0.998	0.004
2.3	0.444	0.048	1	0	0.999	0.003
2.4	0.474	0.052	1	0	1	0
2.5	0.512	0.052	1	0	1	0
2.6	0.553	0.048	1	0	1	0
2.7	0.587	0.056	1	0	1	0
2.8	0.615	0.048	1	0	1	0
2.9	0.662	0.051	1	0	1	0
3.0	0.688	0.042	1	0	1	0

Tabela 2: quantidade média (e desvio padrão) de amostras perturbadas identificadas pela carta *W* para cada caso.

Descontroles	caso (I)		Descontroles	caso (II)	
	média	desvio padrão		média	desvio padrão
0.60	1.000	0.002	-0.14	1	0
0.61	0.999	0.003	-0.13	1	0
0.62	0.999	0.003	-0.12	1	0
0.63	0.999	0.003	-0.11	1	0
0.64	0.998	0.004	-0.10	1	0
0.65	0.998	0.004	-0.09	1	0
0.66	0.997	0.005	-0.08	0.999	0.003
0.67	0.996	0.007	-0.07	0.990	0.010
0.68	0.994	0.008	-0.06	0.955	0.022
0.69	0.991	0.010	-0.05	0.873	0.034
0.70	0.989	0.011	-0.04	0.739	0.043
0.71	0.982	0.013	-0.03	0.593	0.050
0.72	0.974	0.016	-0.02	0.456	0.053
0.73	0.965	0.018	-0.01	0.323	0.049
0.74	0.952	0.021	0.00	0.227	0.037
0.75	0.933	0.025	0.01	0.161	0.036
0.76	0.909	0.030	0.02	0.108	0.031
0.77	0.875	0.032	0.03	0.089	0.027
0.78	0.830	0.039	0.04	0.070	0.028
0.79	0.777	0.041	0.05	0.064	0.025
0.80	0.710	0.047	0.06	0.074	0.026
0.81	0.632	0.046	0.07	0.091	0.028
0.82	0.546	0.048	0.08	0.112	0.030
0.83	0.451	0.047	0.09	0.164	0.037
0.84	0.357	0.049	0.10	0.214	0.040
0.85	0.267	0.044	0.11	0.309	0.047
0.86	0.189	0.039	0.12	0.403	0.048
0.87	0.131	0.034	0.13	0.541	0.048
0.88	0.095	0.029	0.14	0.689	0.048
0.89	0.070	0.024	0.15	0.811	0.037
0.90	0.067	0.025	0.16	0.914	0.026
0.91	0.082	0.026	0.17	0.970	0.015
0.92	0.114	0.031	0.18	0.994	0.007
0.93	0.193	0.039	0.19	1	0
0.94	0.354	0.048	0.20	1	0
0.95	0.613	0.050	0.21	1	0
0.96	0.884	0.030	0.22	1	0
0.97	0.994	0.007	0.23	1	0
0.98	1	0	0.24	1	0
0.99	1	0	0.25	1	0

A tabela 2 apresenta os resultados das simulações para os dois casos descritos na seção anterior para avaliação da carta W. Observamos no caso **(I)** uma alta sensibilidade na detecção dos descontroles, visto que pequenas alterações em relação a correlação de referência já são detectadas em 100% das amostras. O caso **(II)** apresenta resultados semelhantes aos do caso **(I)**, isto é, alta sensibilidade nas detecções dos descontroles impostos. Dessa forma, verificamos o bom desempenho da carta *W* independente do grau da correlação das variáveis no processo sob controle estatístico.

5 Considerações Finais

Este trabalho apresentou um estudo do desempenho das cartas de controle χ^2 e *W*. As cartas χ^2 e *W* são abordagens multivariadas clássicas para o monitoramento de médias e covariâncias, respectivamente.

Através de um estudo simulado utilizando quatro variáveis apresentando uma estrutura de covariância, exibindo correlações fortes e fracas, diferentes cenários foram investigados incluindo diversos descontroles impostos no vetor de médias e na matriz de covariâncias.

Em relação a carta χ^2 verificamos que descontroles impostos nas direções comuns de variabilidade são detectados com menos sensibilidade quando comparados aos descontroles impostos fora das direções comuns. Já em relação a carta *W* verificamos a boa sensibilidade na detecção de descontroles independente do tamanho da correlação entre as variáveis no processo sob controle.

5.1 Bibliografia

Referências

Hotelling, H. (1947). Multivariate quality control. *Techniques of statistical analysis*.

Johnson, R., & Wichern, D. (2007). Applied multivariate statistical analysis. *INC., New Jersey*.

Montgomery, D. C. (2007). *Introduction to statistical quality control*. John Wiley & Sons.

Shewhart, W. A. (1931). *Economic control of quality of manufactured product*. ASQ Quality Press.

Estudo de Simulações na Estimação de Parâmetros dos Processos k-Factor GARMA$(p, u, \lambda, q)_S\alpha S$

Cleber Bisognin [1]

Sílvia R.C. Lopes[2]

Leticia Menegotto[3]

Resumo: Neste trabalho estamos interessados em estudar séries temporais com as características de longa dependência, sazonalidade e alta variabilidade. Os processos k-Factor GARMA $(p, \boldsymbol{u}, \boldsymbol{\lambda}, q)$ com inovações α-estáveis simétricas, denotados por k-Factor GARMA $(p, \boldsymbol{u}, \boldsymbol{\lambda}, q)\_S\alpha S$, nos permitem trabalhar com tais séries temporais. Séries de agregados monetários e rendimentos financeiros são exemplos para aplicações destes processos. O principal objetivo é verificar as condições de estacionariedade, invertibilidade e propor estimadores para os parâmetros destes processos. Para tanto, estendemos o estimador para os processos SARFIMA$(p, d, q) \times (P, D, Q)_s_S\alpha S$, proposto por Ndongo et al. [2010], para os processos k-Factor GARMA $(p, \boldsymbol{u}, \boldsymbol{\lambda}, q)\_S\alpha S$. Neste estimador utilizamos as funções periodograma normalizado suavizado e periodograma suavizado de correlações como estimadores da função poder de transferência [Stein, 2012]. Foram realizadas simulações de Monte Carlo para verificar a acurácia das estimativas dos parâmetros e para tal foram analisados o vício, o erro quadrático médio (EQM) e a variância (Var) das estimativas. Constatamos que ambos os estimadores propostos, apresentaram boas estimativas, no sentido de baixos vício, erro quadrático médio e variância para todos os parâmetros na maioria dos casos analisados. Verificou-se também que quanto menor o valor do $0 < \alpha < 2$ (parâmetro relacionado a variabilidade dos dados, quanto menor α maior a variabilidade da série temporal) menor é a acurácia das estimativas para o parâmetro λ do processo.

Palavras chave: Longa Dependência, Estimação de Parâmetros, Distribuições α-estáveis.

1. Introdução

Em muitas aplicações práticas, pesquisadores têm estudado séries temporais que apresentam longa dependência e sazonalidade. Esse fenômeno ocorre em séries de rendimentos financeiros, agregados monetários e taxa de inflação, por exemplo. Desta forma, vários métodos estatísticos foram propostos para modelar estas séries, dentre eles, os processos Gegenbauer (u, λ) e GARMA (p, u, λ, q). Giraitis e Leipus [1995] e, depois, Woodward et al. [1998] estendem os modelos Gegenbauer e GARMA, respectivamente, aos modelos k-Factor Gegenbauer $(\boldsymbol{u}, \boldsymbol{\lambda})$ e k-Factor GARMA $(p, \boldsymbol{u}, \boldsymbol{\lambda}, q)$, para os quais a função densidade espectral é ilimitada para um número finito k de frequências, chamadas de *frequências de Gegenbauer*.

Há também o interesse em modelar séries temporais com alta variabilidade. Inicialmente, para estudar séries temporais com as propriedades de longa dependência e alta variabilidade, foram propostos, por Kokoszka e Taqqu [1995] os processos ARFIMA(p, d, q) com inovações α-estáveis, denotados por ARFIMA$(p, d, q)_S\alpha S$.

[1]UFSM - Universidade Federal de Santa Maria. Email: *cleber.bisognin@ufsm.br*
[2]Programa de Pós-Graduação em Matemática - UFRGS. Email: *silvia.lopes@ufrgs.br*
[3]UFRGS - Universidade Federal do Rio Grande do Sul. Email: *leticia.menegotto@gmail.com*

Kokoszka e Taqqu [1999] definem os processos ARFIMA$(p, d, q)_S\alpha S$, apresentam a função poder de transferência dos mesmos e demonstram as propriedades de longa dependência e estacionariedade, além de propor um estimador para os parâmetros dos mesmos.

Diongue et al. [2008] apresentam os processos SARFIMA$(p, d, q) \times (P, D, Q)_s$ com variância infinita, denotados por SARFIMA$(p, d, q) \times (P, D, Q)_s_S\alpha S$. Ademais, demonstram algumas propriedades como estacionariedade e invertibilidade, além de proporem um estimador para os parâmetros destes processos. Tais processos, quando $P = 0 = Q$, são um caso particular dos processos k-Factor GARMA $(p, \boldsymbol{u}, \boldsymbol{\lambda}, q)\_S\alpha S$.

Neste trabalho estendemos o estimador proposto por Ndongo et al. [2010] para os processos SARFIMA $(p, d, q) \times (P, D, Q)_s\_S\alpha S$, o qual utiliza o algoritmo de Metropolis-Hastings e a função periodograma normalizado, para os processos k-Factor GARMA $(p, \boldsymbol{u}, \boldsymbol{\lambda}, q)\_S\alpha S$. Utilizamos a função periodograma normalizado suavizado e a função periodograma suavizado de correlação em substituição a função periodograma normalizado como estimadores da função poder de transferência. Tal substituição deve-se ao fato das funções periodograma normalizado suavizado e periodograma suavizado de correlação serem estimadores consistentes da função poder de transferência. Foram testadas várias janelas espectrais e de suavização. Neste estudo apresentamos a janela espectral e de suavização de Bartlett (ver Bartlett [1950]).

2. Processos k-Factor GARMA $(p, u, \lambda, q)\_S\alpha S$

Os processos ARFIMA(p, d, q), onde $d \in (-0.5, 0.5)$, podem ser tratados como uma generalização dos processos ARIMA(p, d, q), onde $d \in \mathbb{N}$, para modelar dados com a propriedade de longa dependência, isto é, quando a função densidade espectral é ilimitada na frequência zero. Similarmente, os processos GARMA(p, u, λ, q) são tratados como uma generalização dos processos ARFIMA(p, d, q), na qual a sua função densidade espectral torna-se ilimitada em alguma frequência G no intervalo $(0, \pi]$, não necessariamente a frequência zero. Contudo, uma limitação dos processos ARFIMA(p, d, q) e do processo mais geral GARMA(p, u, λ, q) é que as suas funções densidade espectral tornam-se ilimitadas em apenas uma frequência do intervalo $(0, \pi]$. Por este motivo, Giraitis e Leipus [1995] e, depois, Woodward et al. [1998] estendem os modelos Gegenbauer e GARMA, respectivamente, aos modelos k-Factor Gegenbauer $(\boldsymbol{u}, \boldsymbol{\lambda})$ e k-Factor GARMA $(p, \boldsymbol{u}, \boldsymbol{\lambda}, q)$, para os quais a função densidade espectral é ilimitada para um número finito k de frequências, chamadas de *frequências de Gegenbauer* (ou frequências G), no intervalo $(0, \pi]$. Na Definição 1 apresentamos os processos k-Factor GARMA $(p, \boldsymbol{u}, \boldsymbol{\lambda}, q)$. Maiores detalhes a respeito destes processos podem ser encontrados em Giraitis e Leipus [1995] e Woodward et al. [1998].

Definição 1. Seja $\{X_t\}_{t\in\mathbb{Z}}$ um processo estocástico que satisfaz a equação

$$\phi(\mathcal{B}) \prod_{j=1}^{k} (1 - 2u_j\mathcal{B} + \mathcal{B}^2)^{\lambda_j} (X_t - \mu) = \theta(\mathcal{B})\varepsilon_t, \tag{1}$$

onde k é um número inteiro, $|u_j| \leqslant 1$, λ_j é um número fracionário, para $j = 1, \cdots, k$, μ é a média do processo, $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ é um processo ruído branco e $\phi(\cdot)$ e $\theta(\cdot)$ são os polinômios de grau p e q dados, respectivamente, por

$$\phi(z) = \sum_{\ell=0}^{p} (-\phi_\ell)\, z^\ell \quad \text{e} \quad \theta(z) = \sum_{m=0}^{q} (-\theta_m)\, z^m, \tag{2}$$

com ϕ_ℓ, $1 \leqslant \ell \leqslant p$, e θ_m, $1 \leqslant m \leqslant q$, constantes reais e $\phi_0 = -1 = \theta_0$.

Então, $\{X_t\}_{t\in\mathbb{Z}}$ é um *processo auto-regressivo de média móvel k-Factor Gegenbauer* de ordem $(p, \boldsymbol{u}, \boldsymbol{\lambda}, q)$, denotado por k-Factor GARMA$(p, \boldsymbol{u}, \boldsymbol{\lambda}, q)$, onde $\boldsymbol{u} = (u_1, \cdots, u_k)'$ e $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_k)'$.

Neste trabalho estamos interessados em estudar os processos k-Factor GARMA $(p, \boldsymbol{u}, \boldsymbol{\lambda}, q)$, apresentados na Definição 1, onde $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ é um processo ruído branco onde suas variáveis aleatórias possuem distribuição α-estável simétrica. Denotaremos estes processos por *k*-Factor GARMA$(p, \boldsymbol{u}, \boldsymbol{\lambda}, q)_S\alpha S$.

Definição 2. Seja X uma variável aleatória que segue distribuição α-estável simétrica. Então, sua função característica é dada por

$$\varphi_X(t) = \mathbb{E}\left(e^{\imath tX}\right) = e^{-\sigma^\alpha |t|^\alpha}, \quad t \in \mathbb{R}, \tag{3}$$

onde $0 < \alpha \leqslant 2$ é o índice de estabilidade e $\sigma > 0$ é o parâmetro de escala.

Se $\alpha = 2$, a variável aleatória X possui distribuição Gaussiana com $\mathbb{E}(X) = 0$ e $\mathrm{Var}(X) = 2\sigma^2$.

Proposição 1. Seja $\{X_t\}_{t\in\mathbb{Z}}$ um processo *k*-Factor GARMA$(p, \boldsymbol{u}, \boldsymbol{\lambda}, q)_S\alpha S$. Então as seguintes afirmações são verdadeiras.

(i) O processo $\{X_t\}_{t\in\mathbb{Z}}$ é estacionário se todas as raízes da equação $\phi(z) = 0$ estão fora do círculo unitário. Além disso, $\lambda_j < 1 - \frac{1}{\alpha}$, quando $|u_j| < 1$, e $\lambda_j < \frac{1}{2}(1 - \frac{1}{\alpha})$, quando $|u_j| = 1$, para $j = 1, \cdots, k$;

(ii) O processo $\{X_t\}_{t\in\mathbb{Z}}$ é invertível se todas as raízes da equação $\theta(z) = 0$ estão fora do círculo unitário. Além disso, $\lambda_j > -1 + \frac{1}{\alpha}$, quando $|u_j| < 1$, e $\lambda_j > -\frac{1}{2}(1 - \frac{1}{\alpha})$, quando $|u_j| = 1$, para $j = 1, \cdots, k$;

(iii) Sob as condições dos itens (i) e (ii) as representações MA(∞) e AR(∞), respectivamente, são dadas por

$$\psi(z) = \sum_{\ell\geqslant 0} \psi_\ell z^\ell = \frac{\theta(z)}{\phi(z)} \prod_{j=1}^{k} (1 - 2u_j z + z^2)^{-\lambda_j}. \tag{4}$$

e

$$\pi(z) = \sum_{l\geqslant 0} \pi_l z^l = \frac{\phi(z)}{\theta(z)} \prod_{j=1}^{k} (1 - 2u_j z + z^2)^{\lambda_j}. \tag{5}$$

(iv) Seja $\{X_t\}_{t\in\mathbb{Z}}$ um processo *k*-Factor GARMA$(p, \boldsymbol{u}, \boldsymbol{\lambda}, q)_S\alpha S$ estacionário. Então a função poder de transferência do processo $\{X_t\}_{t\in\mathbb{Z}}$ é dada por

$$f_X(\omega) = \left| \sum_{\ell\geqslant 0} \psi_\ell e^{-\imath\ell\omega} \right|^2 = \frac{|\theta(e^{-\imath\omega})|}{|\phi(e^{-\imath\omega})|} \prod_{j=1}^{k} [2(\cos(\omega) - u_j)]^{-2\lambda_j}, \tag{6}$$

onde $0 < \omega \leqslant \pi$ e $G_j = \cos^{-1}(u_j)$ são chamadas frequências de Gegenbauer.

3. Estimação dos Parâmetros

Nos estudos de séries temporais, temos como um dos principais objetivos a estimação dos parâmetros dos processos que são utilizados para modelar os dados. Neste trabalho, a fim de realizar a estimação dos parâmetros dos processos, estendemos o estimador proposto, para os processos SARFIMA $(0, d, 0)\times\ (0, D, 0)_s\_S\alpha S$ (ver Ndongo et al. [2010]), agora para os processos *k*-Factor GARMA$(p, \boldsymbol{u}, \boldsymbol{\lambda}, q)_S\alpha S$ (ver Definição 1). O método

que estamos propondo consiste em estimar os parâmetros do modelo utilizando o algoritmo de Metropolis-Hastings que é baseado nas funções periodograma normalizado suavizado e periodograma suavizado de correlações. Estes são estimadores consistentes da função poder de transferência, com janela espectral e de suavização de Bartlett.

Estimador MCMCPS - Este estimador é obtido substituindo-se a função periodograma normalizado pela função periodograma normalizado suavizado, pois esta última função é um estimador consistente para a função poder de transferência. Para mais detalhes ver teorema 2.1 de Klüppelberg e Mikosch [1994]. Assim, o estimador do vetor de parâmetros $\boldsymbol{\eta} = (\boldsymbol{\phi}, \boldsymbol{u}, \boldsymbol{\lambda}, \boldsymbol{\theta})$, denotado por $\widehat{\boldsymbol{\eta}}$, é o valor que minimiza $\sigma_T^2(\boldsymbol{\eta})$, dada por

$$\widehat{\sigma}_T^2(\boldsymbol{\eta}) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\tilde{T}_n(\omega)}{f_{_X}(\omega, \boldsymbol{\eta})} d\omega, \tag{7}$$

onde $f_{_X}(\cdot, \boldsymbol{\eta})$ é a função poder de transferência dada pela equação (6). O numerador do integrando da expressão (7) é a função periodograma normalizado suavisado dado por

$$\tilde{T}_n(\omega) = \sum_{|k| \leqslant m} W_n(k)\tilde{I}_n(\omega_k), \tag{8}$$

onde $W(\cdot)$ é a *janela espectral* com $\omega_k = \omega + \frac{k}{n}$, para $|k| \leqslant m,$, $m = m(n)$ é uma sequência em $\mathbb{N}$ tal que

$$m \to \infty, \quad \text{e} \quad \frac{m}{n} \to 0, \quad n \to \infty,$$

e $(W_n)_{n\in\mathbb{N}}$ é uma sequência de pesos que satisfazem as seguintes condições

$$W_n(k) = W_n(-k), \quad W_n(k) \geqslant 0, \quad \text{para todo} \quad h \in \mathbb{N}, \tag{9}$$

$$\sum_{|k|\leqslant m} W_n(k) = 1, \quad \sum_{|k|\leqslant m} W_n^2(k) = o(1), \quad n \to \infty. \tag{10}$$

Estimador MCMCPSC - este estimador é obtido substituindo-se a função periodograma normalizado pela função periodograma suavizado de correlações. Isso decorre do fato da função periodograma suavizado de correlações ser um estimador consistente para a função poder de transferência. Para maiores detalhes ver teorema 2.8 de Stein [2012]. Assim, o estimador do vetor de parâmetros $\boldsymbol{\eta} = (\boldsymbol{\phi}, \boldsymbol{u}, \boldsymbol{\lambda}, \boldsymbol{\theta})$, denotado por $\widehat{\boldsymbol{\eta}}$, é o valor que minimiza $\sigma_K^2(\boldsymbol{\eta})$, dada por

$$\widehat{\sigma}_K^2(\boldsymbol{\eta}) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\tilde{K}_n(\omega)}{f_{_X}(\omega, \boldsymbol{\eta})} d\omega. \tag{11}$$

onde $f_{_X}(\cdot, \boldsymbol{\eta})$ é a função poder de transferência (ver equação (6)),

$$\widetilde{K}_n(\omega) = \sum_{|h|<m_n} \mathcal{W}(h/m_n)\widehat{\rho}_{_X}(h)e^{-\imath\omega h}, \quad \text{para} \quad \omega \in [-\pi, \pi], \tag{12}$$

é a função periodograma suavizado de correlações. Segundo Brockwell e Davis [2013], página 358, a função $\mathcal{W}(\cdot)$ é chamada de *lag window* ou *janela de suavização* e é uma função par, contínua por partes e satisfaz as condições: $\mathcal{W}(0) = 1$, $|\mathcal{W}(x)| \leqslant 1$, para todo $x \in \mathbb{R}$ e $\mathcal{W}(x) = 0$, para $|x| > 1$.

Segundo Brockwell e Davis [2013], página 358, para processos estacionários com inovações Gaussianas ($\alpha = 2$), m_n é uma função em $\mathbb{N}$ tal que $m_n \to \infty$ e $\frac{m_n}{n} \to 0$, quando $n \to \infty$. No caso dos processos satisfazendo as condições do Teorema 10.4.1 (página 351) e a condição para m_n, o periodograma suavizado de covariância [ver equação 10.4.8 Brockwell e Davis, 2013], é um estimador consistente para a função densidade espectral.

O procedimento para encontrar o vetor $\widehat{\boldsymbol{\eta}}$ que minimiza a equação (7) ou (11) é baseado no algoritmo de Metropolis-Hastings. Maiores de detalhes podem ser encontrados em Ndongo et al. [2010] e Bisognin e Menegotto [2017].

Neste trabalho utilizamos como janelas espectral e de suavização de Bartlett, as quais são baseadas na função triangular dada por

$$\mathtt{w}(x) = \begin{cases} 1 - |x|, & \text{se } |x| \leqslant 1; \\ \\ 0, & \text{se } |x| > 1. \end{cases} \tag{13}$$

A seguir definimos as janelas espectral e de suavização de Bartlett (ver Bartlett [1950]).

Definição 3. A janela espectral de Bartlett é dada por

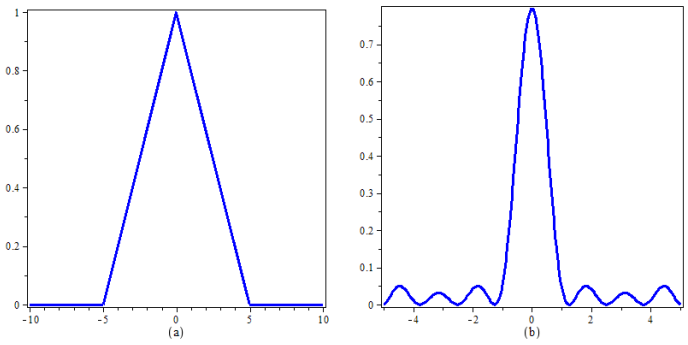
$$W_n(\omega) = \frac{1}{2\pi m}\left[\frac{\text{sen}(\frac{\omega m}{2})}{\text{sen}(\frac{\omega}{2})}\right]. \tag{14}$$

A janela de suavização de Bartlett é dada por

$$\mathcal{W}(h/m_n) = \begin{cases} 1 - \frac{|h|}{m_n}, & \text{se } |h| \leqslant m_n; \\ \\ 0, & \text{se } |h| > m_n, \end{cases} \tag{15}$$

onde m_n é o ponto de truncamento que depende do tamanho da amostra.

A Figura 1 apresenta o gráfico da janela de suavização de Bartlett e sua correspondente janela espectral.



(a) (b)

Figura 1: Janelas de suavização e espectral de Bartlett. (a) Janela de Suavização de Bartlett $\mathcal{W}(\cdot)$, com $m_n = 5$. (b) Janela Espectral de Bartlett $W_n(\cdot)$, com $m = 5$.

Fonte: Os Autores.

4. Simulações de Monte Carlo

Para gerarmos realizações dos processos k-Factor GARMA$(p, \boldsymbol{u}, \boldsymbol{\lambda}, q)_S\alpha S$ utilizamos a representação média móvel infinita (ver equação (4)) com apropriado ponto de truncamento. Por ser um processo complexo, este ponto de truncamento da representação média móvel infinita deve ser consideravelmente grande. Gray et al. (1989) utilizam a representação média móvel infinita dos processos Gegenbauer (quando $k = 1$ e $p = 0 = q$) para gerar realizações dos mesmos, truncando a representação em 290.000 valores. Esta forma de gerar as realizações de um processo estocástico consome muito tempo computacional e a precisão depende de quão rápido os coeficientes da representação média móvel infinita convergem à zero. Neste trabalho truncamos a representação média móvel infinita em 5000.

A seguir, descrevemos o procedimento utilizado para gerar as realizações de um processo k-Factor GARMA $(p, \boldsymbol{u}, \boldsymbol{\lambda}, q)\_S\alpha S$.

1. Calculamos 5000 coeficientes da representação média móvel infinita.

2. Geramos um processo cujas variáveis aleatórias tem distribuição α-estável simétrica, dada pela Definição 2, quando $\alpha \in \{0,3; 0,5; 0,7; 0,9; 1,3; 1,5; 1,7; 1,9\}$ e parâmetro de escala $\sigma = 1$;

3. Para cada $t \in \{1, \cdots, n\}$, os valores X_t são calculados através da convolução entre os coeficientes da representação média móvel infinita e o processo α-estável simétrico.

A seguir, apresentamos alguns resultados sobre estimação dos parâmetros dos processos k-Factor GARMA $(p, \boldsymbol{u}, \boldsymbol{\lambda}, q)\_S\alpha S$ gerados a partir do procedimento mencionado anteriormente. Os parâmetros foram estimados utilizando os estimadores MCMCPS e MCMCPSC descritos na Seção 3. Para a função periodograma suavizado, usamos $m \in \{1, 2, 3, 4\}$ e para a função periodograma suavizado de correlação usamos $m_n = n^\beta$, com $\beta \in \{0,8; 0,85; 0,9; 0,95\}$.

Tabela 1: Estimador MCMCPS - Resultados das simulações de Monte Carlo para o processo k-Factor GARMA$(p, \boldsymbol{u}, \boldsymbol{\lambda}, q)_S\alpha S$ quando $p = 0 = q$, $k = 1$, $u_1 = 0,2$, $\lambda_1 = 0,4$, $\alpha \in \{1,3; 1,5; 1,7; 1,9\}$, $m \in \{1, 2, 3, 4\}$, com $n = 1000$, utilizando a janela espectral de Bartlett.

$\alpha = 1,3$								
	$m = 1$		$m = 2$		$m = 3$		$m = 4$	
	$\widehat{u}_1$	$\widehat{\lambda}_1$	$\widehat{u}_1$	$\widehat{\lambda}_1$	$\widehat{u}_1$	$\widehat{\lambda}_1$	$\widehat{u}_1$	$\widehat{\lambda}_1$
Média	0,2000	0,4097	0,2001	0,4023	0,2003	0,4048	0,2002	0,4032
Vício	0,0000	0,0097	0,0001	0,0023	0,0003	0,0048	0,0002	0,0032
EQM	0,0001	0,0025	0,0001	0,0028	0,0001	0,0027	0,0001	0,0026
Var	0,0001	0,0024	0,0001	0,0028	0,0001	0,0027	0,0001	0,0026
$\alpha = 1,5$								
Média	0,1999	0,4051	0,2004	0,3958	0,2004	0,3978	0,1996	0,3984
Vício	-0,0001	0,0051	0,0004	-0,0042	0,0004	-0,0022	-0,0004	-0,0016
EQM	0,0001	0,0023	0,0003	0,0029	0,0001	0,0026	0,0001	0,0026
Var	0,0001	0,0022	0,0003	0,0028	0,0001	0,0026	0,0001	0,0026
$\alpha = 1,7$								
Média	0,2003	0,4045	0,2000	0,3946	0,1996	0,3951	0,2002	0,3933
Vício	0,0003	0,0045	0,0000	-0,0054	-0,0004	-0,0049	0,0002	-0,0067
EQM	0,0001	0,0022	0,0001	0,0030	0,0001	0,0026	0,0001	0,0027
Var	0,0001	0,0022	0,0001	0,0029	0,0001	0,0026	0,0001	0,0026
$\alpha = 1,9$								
Média	0,1996	0,4022	0,2007	0,3934	0,2007	0,3924	0,2006	0,3926
Vício	-0,0004	0,0022	0,0007	-0,0066	0,0007	-0,0076	0,0006	-0,0074
EQM	0,0001	0,0022	0,0002	0,0025	0,0002	0,0028	0,0001	0,0026
Var	0,0001	0,0022	0,0002	0,0025	0,0002	0,0028	0,0001	0,0026

Fonte: Os Autores.

Tabela 2: Estimador MCMCPS - Resultados das simulações de Monte Carlo para o processo *k*-Factor GARMA$(p, \boldsymbol{u}, \boldsymbol{\lambda}, q)_S\alpha S$ quando $p = 0 = q$, $k = 1$, $u_1 = 0,2$, $\lambda_1 \in \{-2,4; -1,1; -0,45; -0,10\}$, $\alpha \in \{0,3; 0,5; 0,7; 0,9\}$, $m \in \{1,2,3,4\}$, com $n = 1000$, utilizando a janela espectral de Bartlett.

$\lambda_1 = -2,4$ e $\alpha = 0,3$								
	$m = 1$		$m = 2$		$m = 3$		$m = 4$	
	$\widehat{u}_1$	$\widehat{\lambda}_1$	$\widehat{u}_1$	$\widehat{\lambda}_1$	$\widehat{u}_1$	$\widehat{\lambda}_1$	$\widehat{u}_1$	$\widehat{\lambda}_1$
Média	0,2011	-2,9953	0,1940	-3,0896	0,1971	-3,1006	0,1961	-3,1030
Vício	0,0011	-0,5953	-0,0060	-0,6896	-0,0029	-0,7006	-0,0039	-0,7030
EQM	0,0089	0,5400	0,0072	0,6248	0,0074	0,6333	0,0072	0,6349
Var	0,0089	0,1857	0,0072	0,1495	0,0074	0,1426	0,0072	0,1408
$\lambda_1 = -1,1$ e $\alpha = 0,5$								
Média	0,1988	-1,3292	0,1972	-1,3757	0,1963	-1,3715	0,1997	-1,3586
Vício	-0,0012	-0,2292	-0,0028	-0,2757	-0,0037	-0,2715	-0,0003	-0,2586
EQM	0,0040	0,1164	0,0021	0,1420	0,0024	0,1391	0,0022	0,1320
Var	0,0040	0,0640	0,0021	0,0661	0,0024	0,0655	0,0022	0,0652
$\lambda_1 = -0,45$ e $\alpha = 0,7$								
Média	0,2030	-0,4938	0,2008	-0,4983	0,2016	-0,4973	0,2020	-0,4968
Vício	0,0030	-0,0438	0,0008	-0,0483	0,0016	-0,0473	0,0020	-0,0468
EQM	0,0029	0,0063	0,0032	0,0080	0,0030	0,0069	0,0031	0,0067
Var	0,0029	0,0043	0,0032	0,0057	0,0030	0,0046	0,0031	0,0045
$\lambda_1 = -0,10$ e $\alpha = 0,9$								
Média	0,2175	-0,1288	0,2218	-0,1295	0,2219	-0,1288	0,2229	-0,1287
Vício	0,0175	-0,0288	0,0218	-0,0295	0,0219	-0,0288	0,0229	-0,0287
EQM	0,0138	0,0032	0,0182	0,0018	0,0137	0,0018	0,0149	0,0015
Var	0,0135	0,0024	0,0177	0,0010	0,0132	0,0010	0,0143	0,0007

Fonte: Os Autores.

Tabela 3: Estimador MCMCPSC - Resultados das simulações de Monte Carlo para o processo *k*-Factor GARMA$(p, \boldsymbol{u}, \boldsymbol{\lambda}, q)_S\alpha S$ quando $p = 0 = q$, $k = 1$, $u_1 = 0,2$, $\lambda_1 = 0,4$, $\alpha \in \{1,3; 1,5; 1,7; 1,9\}$, $m_n = n^{\beta}$, sendo $n = 1000$ e $\beta \in \{0,8; 0,85; 0,9; 0,95\}$ para a janela de suavização de Bartlett.

$\alpha = 1,3$								
	$\beta = 0,8$		$\beta = 0,85$		$\beta = 0,9$		$\beta = 0,95$	
	$\widehat{u}_1$	$\widehat{\lambda}_1$	$\widehat{u}_1$	$\widehat{\lambda}_1$	$\widehat{u}_1$	$\widehat{\lambda}_1$	$\widehat{u}_1$	$\widehat{\lambda}_1$
Média	0,1998	0,4194	0,2003	0,4164	0,2001	0,4149	0,2002	0,4090
Vício	-0,0002	0,0194	0,0003	0,0164	0,0001	0,0149	0,0002	0,0090
EQM	0,0001	0,0025	0,0001	0,0028	0,0001	0,0026	0,0001	0,0024
Var	0,0001	0,0021	0,0001	0,0025	0,0001	0,0023	0,0001	0,0023
$\alpha = 1,5$								
Média	0,1995	0,4172	0,2003	0,4111	0,2000	0,4124	0,1996	0,4105
Vício	-0,0005	0,0172	0,0003	0,0111	0,0000	0,0124	-0,0004	0,0105
EQM	0,0001	0,0024	0,0001	0,0022	0,0001	0,0022	0,0001	0,0023
Var	0,0001	0,0021	0,0001	0,0021	0,0001	0,0020	0,0001	0,0022
$\alpha = 1,7$								
Média	0,2000	0,4120	0,1999	0,4102	0,1997	0,4086	0,2003	0,4050
Vício	0,0000	0,0120	-0,0001	0,0102	-0,0003	0,0086	0,0003	0,0050
EQM	0,0001	0,0020	0,0001	0,0022	0,0001	0,0021	0,0001	0,0024
Var	0,0001	0,0018	0,0001	0,0021	0,0001	0,0020	0,0001	0,0023
$\alpha = 1,9$								
Média	0,2006	0,4113	0,2002	0,4073	0,2001	0,4079	0,2001	0,4015
Vício	0,0006	0,0113	0,0002	0,0073	0,0001	0,0079	0,0001	0,0015
EQM	0,0001	0,0022	0,0001	0,0021	0,0001	0,0021	0,0001	0,0021
Var	0,0001	0,0020	0,0001	0,0020	0,0001	0,0020	0,0001	0,0021

Fonte: Os Autores.

A análise das Tabelas 1 a 3 é apresentada na seção de Conclusão.

5. Conclusão

Neste trabalho estendemos o estimador de Ndongo et al. [2010], proposto inicialmente para os processos SARFIMA $(p, d, q) \times (P, D, Q)_s\_S\alpha S$, para os processos *k*-Factor GARMA$(p, \boldsymbol{u}, \boldsymbol{\lambda}, q)_S\alpha S$. Além disso, utilizamos as funções periodograma normalizado suavizado e periodograma suavizado de correlação como estimadores da função poder de transferência. Para os processos SARFIMA $(p, d, q) \times (P, D, Q)_s\_S\alpha S$, Ndongo et al. [2010] utilizaram apenas a função periodograma normalizado. Os estimadores foram denotados por MCMCPC e MCMCPSC, respectivamente.

Estamos interessados em estimar os parâmetros dos processos *k*-Factor GARMA$(p, \boldsymbol{u}, \boldsymbol{\lambda}, q)_S\alpha S$ estacionários. Pela Proposição 1, item (i), as suposições para a estacionariedade destes processos são: todas as raízes da equação $\phi(z) = 0$ devem estar fora do círculo unitário e $\lambda_j < 1 - \frac{1}{\alpha}$, quando $|u_j| < 1$, e $\lambda_j < \frac{1}{2}(1 - \frac{1}{\alpha})$, quando $|u_j| = 1$, para $j = 1, \cdots, k$. Desta forma para $\alpha \in \{1,3; 1,5; 1,7; 1,9\}$, fixamos $u_1 = 0,2$ e $\lambda_1 = 0,4$. Quando $\alpha \in \{0,3; 0,5; 0,7; 0,9\}$, para o processo ser estacionário é preciso que $\lambda_1 \in \{-2,4; -1,2; -0,45; -0,10\}$, respectivamente.

Para $\alpha \in \{1,3; 1,5; 1,7; 1,9\}$ é possível verificar através das Tabelas 1 e 3 que as estimativas, utilizando ambos os estimadores, possuem baixo vício, erro quadrático médio (EQM) e variância (Var).

As estimativas obtidas através do estimador MCMCPS, para o parâmetro u_1, apresentam baixo vício e permanece praticamente inalterado a medida que o α cresce. Já o vício das estimativas para o parâmetro λ_1 decresce a medida que α cresce. Podemos destacar também que o vício, o erro quadrático médio e a variância das estimativas de u_1 são menores que as mesmas estatísticas para o parâmetro λ_1. Analisando o vício das estimativas, quanto a variação de m, percebemos que o vício para o parâmetro u_1 permanece praticamente inalterado, enquanto o vício do parâmetro λ_1 decresce, a medida que m cresce.

O estimador MCMCPSC, apresentou menor vício nas estimativas do parâmetro u do que para o parâmetro λ_1. Destacamos também que o vício, o erro quadrático médio e a variância das estimativas de u_1 são menores que as mesmas estatísticas para o parâmetro λ_1. O erro quadrático médio e a variância das estimativas de ambos os parâmetros, independente do valor de α, permanecem praticamente inalterados a medida que β cresce. Quando β cresce, o vício das estimativas de u_1 permanece quase inalterado enquanto que o vício de λ_1 decresce, para todos os valores de α analisados.

Para $\alpha \in \{0,3; 0,5; 0,7; 0,9\}$, o estimador MCMCPS apresenta menor vício nas estimativas do parâmetro u_1, mas na maioria dos casos estudados ocorre um aumento do vício a medida que m cresce. O estimador também apresenta baixos erro quadrático médio (EQM) e variância (Var) nas estimativas de u_1. O maior valor ocorre quando $m = 4$. Nas estimativas de λ_1, o estimador MCMCPS, apresenta maior vício do que nas estimativas de u_1. Além disso, o vício aumenta a medida que o m cresce. Quando λ_1 cresce, também ocorre um aumento no vício, no erro quadrático médio (EQM) e na variância (Var).

Como futuros trabalhos devemos considerar outros valores de tamanho amostral, $k > 1, p \neq 0, q \neq 0$, além de estudar métodos de previsão utilizando os processos *k*-Factor GARMA$(p, \boldsymbol{u}, \boldsymbol{\lambda}, q)_S\alpha S$.

Referências

Bartlett, M. S. (1950). Periodogram analysis and continuous spectra. *Biometrika*, 37(1/2):1–16.

Bisognin, C. e Menegotto, L. (2017). Previsão utilizando processos sarfima com estimação dos parâmetros via mcmc. *Encontro de Modelagem Estatística (1.: 2017: Maringá, PR).[Anais]. Maringá: UEM, 2017.*

Brockwell, P. J. e Davis, R. A. (2013). *Time series: theory and methods*. Springer Science & Business Media.

Diongue, A. K., Diop, A., e Ndongo, M. (2008). Seasonal fractional arima with stable innovations. *Statistics & Probability Letters*, 78(12):1404–1411.

Giraitis, L. e Leipus, R. (1995). A generalized fractionally differencing approach in long-memory modeling. *Lithuanian Mathematical Journal*, 35(1):53–65.

Klüppelberg, C. e Mikosch, T. (1994). Some limit theory for the self-normalised periodogram of stable processes. *Scandinavian Journal of Statistics*, 21(4):485–491.

Kokoszka, P. S. e Taqqu, M. S. (1995). Fractional arima with stable innovations. *Stochastic processes and their applications*, 60(1):19–47.

Kokoszka, P. S. e Taqqu, M. S. (1999). Discrete time parametric models with long memory and infinite variance. *Mathematical and computer modelling*, 29(10):203–215.

Ndongo, M., Diongue, A. K., Diop, A., e Dossou-Gbété, S. (2010). Estimation of long-memory parameters for seasonal fractional arima with stable innovations. *Statistical Methodology*, 7(2):141–151.

Stein, J. (2012). Estimação em processos com longa dependência, sazonalidade e inovações normais ou α-estáveis. *Dissertação de Mestrado. Porto Alegre: UFRGS.*

Woodward, W. A., Cheng, Q. C., e Gray, H. L. (1998). A k-factor garma long-memory model. *Journal of time series analysis*, 19(4):485–504.

Estudo da Sensibilidade do Bayes Factor para seleção de modelos

Lauren Alves Vieira [1]

Gabriela Bettella Cybis[2]

Resumo: Métodos bayesianos filogenéticos são uma ferramenta central na biologia evolutiva. Dentre estes o Modelo de Variável Latente estima correlações entre características fenotípicas (contínuas e categóricas ordinais ou nominais), controlando para história evolutiva entre os indivíduos amostrados. Nas aplicações deste modelo é comum a escolha de prioris pouco informativas, geralmente adotando a distribuição conjugada Wishart Inversa para matriz de covariâncias do modelo.

Nossos resultados prévios evidenciaram uma possível sensibilidade do método de seleção de modelos quanto a escolha da priori, de modo que modelos com maior número de graus de liberdade (**gl**), pareciam ser favorecidos. Com o intuito de avaliar esse efeito da priori sobre a seleção do modelo, foi conduzido o estudo apresentado abaixo.

Palavras-chave: *Variável latente, Bayes Factor, Priori*.

1 Introdução

O estudo de correlações evolutivas é um dos grandes focos da biologia evolutiva, com aplicações nas mais diversas áreas. Neste contexto, está a estimação de correlações nos processos evolutivos de traços fenotípicos. Entretanto para estimar adequadamente estas correlações devemos separá-las das correlações induzidas pela história evolutiva compartilhada entre os indivíduos, que pode ser inferida através de dados. O modelo Filogenético de Variável Latente (Cybis et al 2015) mostra-se como uma opção para estas análises, já que pode ser usado para estimar correlações entre diferentes tipos de dados fenotípico enquanto controla para história evolutiva compartilhada dos indivíduos ou espécies em estudo.

A diferenciação entre correlações inerentes ao processo de evolução dos fenótipos e correlações geradas pela história evolutiva é necessária para identificação de dois fenômenos de interesse biológico: ligação gênica e seleção natural. O estudo da evolução da resistência bacteriana a diferentes antibióticos é um exemplo de problema de interesse epidemiológico em que correlações na evolução de fenótipos

[1]UFRGS - Universidade Federal do Rio Grande do Sul. Email: *laurendiasalves@gmail.com*
[2]UFRGS - Universidade Federal do Rio Grande do Sul. Email: *gabriela.cybis@ufrgs.br*

são um indício de ligação gênica. De modo similar, pressões seletivas entre características como hábitos alimentares e traços morfológicos em grupos de mamíferos também podem ser estudadas por meio de correlações evolutivas.

Para estimação deste tipo de correlação é comum o uso de uma abordagem Bayesiana. Dentre outros modelos para este tipo de relação o Modelo de Variável Latente utiliza uma transformação bijetora que relaciona uma variável latente a uma variável observável. A forma entre essas variáveis depende do tipo de variável em estudo (Contínua, Binária, Categórica Ordinal ou Nominal). Em alguns casos, como no estudo de hábitos alimentares de morcegos, é difícil se escolher um modelo para os dados, uma vez que não existam informações prévias ou mesmo indícios que estes dados são Ordinais ou Nominais. Para verificação do ajuste do modelo aos dados é comum o uso do método de Bayes Factor (Gelman et al.2003), que compara pares de modelos quanto ao seu ajuste a um mesmo conjunto de dados.

Em trabalhos prévios que visavam avaliar as propriedades estatísticas do Modelo de Variável Latente, obtivemos resultados que evidenciavam que o método de Bayes Factor é afetado pela escolha da priori. Neste estudo consideramos amostras de características Ordinais, para as quais utilizamos o Modelo de Variável Latente, considerando um modelo para os dados ora ordinal e ora nominal, comparando os resultados através de Bayes Factor. Se percebeu uma provável sensibilidade do método a escolha de priori. Para melhor compreender este comportamento realizamos o breve estudo descrito neste trabalho.

2 Metodologia

Modelo Filogenético de Variável Latente

A história evolutiva de conjuntos de indivíduos pode ser representada através de uma árvore filogenética (ou filogênia) τ, que nada mais é que um grafo acíclico onde os N nós externos (vértices de grau 1, também chamados folhas) representam os indivíduos da amostra no tempo atual, também possui apenas um nó de grau 2 chamado raiz, que representa o ancestral comum mais recente a todos os indivíduos. Esta estrutura conta ainda com $N - 2$ nós internos (vértices de grau 3), que descrevem as bifurcações evolutivas decorrentes da separação das diferentes linhagens. As arestas (ou galhos) que ligam estes nós representam o tempo evolutivo decorrido até a ocorrência de uma bifurcação, de modo que a o tamanho das arestas é proporcional a esta quantidade. É possível modelar a evolução de variáveis fenotípicas através de um processo estocástico que inicia na raiz da filogenia e evolui ao longo dos galhos da árvore até as folhas onde os valores foram observados. A figura 1 apresenta um exemplo de filogênia.

O modelo filogenético de variável latente descreve a evolução de uma variável observável Y sobre uma filogênia τ, determinada por uma variável X não observável, chamada de variável latente cuja evolução temporal ao longo de τ segue o modelo de movimento browniano. Assim ao final deste processo

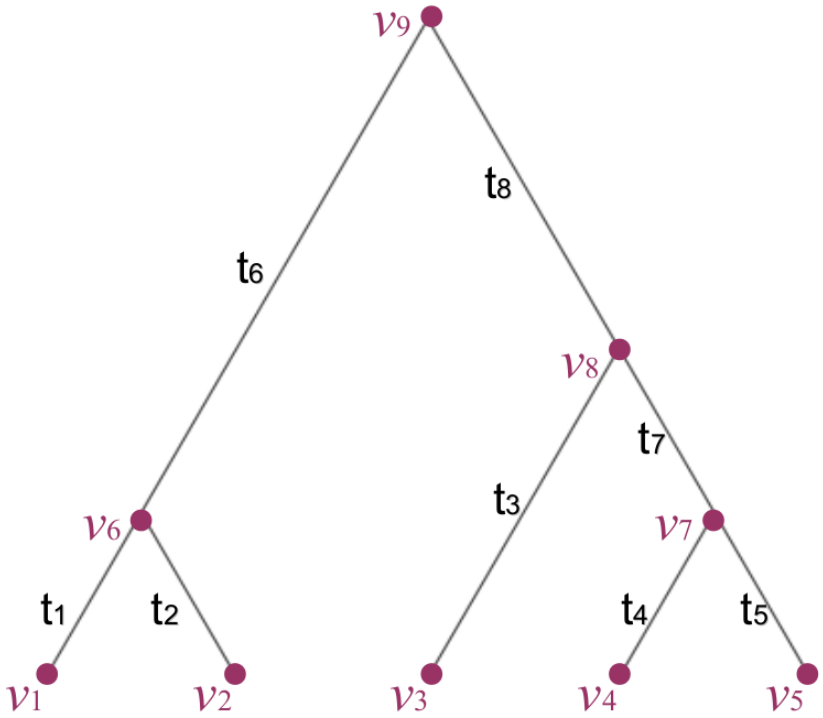


Figura 1: Exemplo de Árvore filogenética com $N = 5$.

a variável Y é determinada por meio de uma função de ligação $g(X)$, a partir dos valores de X. Quando, por exemplo, a variável Y é binária seu valor é determinado pela posição de X em relação a um limiar, já quando Y é contínua temos $Y = X$. No caso de Y multivariado, com estados não ordenados, cada componente de Y é determinada por mais de uma componente de X, porém se Y é multivariado e seus k estados possuem algum tipo de ordenamento, então seus valores são determinados pela posição de X quanto a $k-1$ limiares. Este modelo foi inspirado pelo modelo limiar filogenético. A matriz de precisão Σ^{-1} do movimento browniano multivariado que descreve a evolução de X é utilizada como um proxi para estimar a correlação evolutiva entre as variáveis componentes de Y (Felsenstein 2005).

Para o cálculo da função de verossimilhança deste modelo, consideramos uma extensão dos dados Z, tal que $\mathbf{Z} = (\mathbf{Y}, \mathbf{X})$, onde $\mathbf{Y} = (Y_0, ..., Y_N)$ são os valores observados da variável D-dimensional de interesse Y nos N indivíduos da amostra (folhas da filogênia), e $\mathbf{X} = (X_0, ..., X_N)$ são os valores da variável latente D-dimensional X nos mesmos nós. O movimento browniano ao longo da árvore τ que descreve a evolução de X é um processo já longamente explorado na literatura (Felsenstein, 1988), e sua densidade $P(\mathbf{X}|\Sigma^{-1}, \tau)$ pode ser calculada por meio de um algoritmo iterativo que computa uma série de convoluções de distribuições normais D-variadas ao longo das arestas de τ. Desse modo, temos

$$P(X, Y|\tau, \Sigma^{-1}) = P(X|\tau, \Sigma^{-1})P(Y|X).$$

Se Y é uma variável binária, definimos $P(X|Y)$ como

$$P(Y|X) = \prod_{i=1}^{N}\prod_{j=1}^{D}\left(\mathbf{I}(y_{i,j} = 1)\mathbf{I}(x_{i,j} > 0) + \mathbf{I}(y_{i,j} = 0)\mathbf{I}(x_{i,j} \leq 0)\right),$$

em que $\mathbf{I}(A)$ é a função indicadora de A, e $x_{i,j}$ e $y_{i,j}$ são a j-ésima componente das respectivas variáveis no nó i. Logo, em cada coordenada, temos $Y = 1$ se a variável latente é maior do que zero, e $Y = 0$ caso contrário. Quando Y é contínuo, tomamos $Y = X$, fixando o valor da variável latente nos nos externos. Se Y é uma variável categórica com k estados ordenados suas entradas são determinadas de acordo com k intervalos definidos na variável latente X a partir de $k-1$ limiares independentes. Já se estas categorias não são ordenadas, então a cada entrada de Y correspondem $k-1$ variáveis latentes em X. O valor observado $y_{i,j}$ na componente j da observação i é determinado pela maior das variáveis latentes correspondentes$\{x_{i,j'}, \cdots x_{i,j'+k-2}\}$ de modo que a função link é dada neste caso por

$$y_{ij} = g(x_{i,j'}, \cdots x_{i,j'+k-2}) = \begin{cases} s_1 \quad se \quad 0 = sup(0, x_{i,j'}, \cdots x_{i,j'+k-2}) \\ s_l \quad se \quad x_{i,l} = sup(0, x_{i,j'}, \cdots x_{i,j'+k-2}), \end{cases}$$

em que, sem perda de generalidade, tomamos o primeiro estado s_1 como o estado de referência. Neste caso

$$P(Y|X) = \prod_{i=1}^{N} \prod_{j=1}^{D} \left(\mathbf{I}(y_{i,j} = g(x_{i,j'}, \cdots x_{i,j'+k-2}) \right).$$

Também podemos naturalmente considerar a extensão em que alguns componentes de Y são discretos e outros contínuos.

Neste modelo a inferência é feita em uma perspectiva Bayesiana, de modo que calculamos a distribuição à posteriori como

$$P(\Sigma|X, Y, \tau) \propto P(X, Y|\tau, \Sigma^{-1})P(\Sigma) = P(Y|X)P(X|\tau, \Sigma^{-1})P(\Sigma),$$

na qual utilizamos a distribuição conjugada Whishart para distribuição à priori $P(\Sigma)$. Para fazer inferência baseada nesse modelo utilizamos um algorítimo de MCMC.

Bayes Factor

O método de Bayes Factor compara duas hipóteses independentes, aplicadas a modelos como

$$\text{O modelo } \boldsymbol{M}_1 \text{ é o correto } \times \text{ O modelo } \boldsymbol{M}_2 \text{ é o correto,}$$

avaliando qual hipótese é mais verossímil, dada a amostra observada. Para isto calcula-se a razão entre as verossimilhanças marginais das hipóteses

$$BF = \frac{L(\boldsymbol{M}_1|\boldsymbol{Y}, \Sigma^{-}, \tau)}{L(\boldsymbol{M}_2|\boldsymbol{Y}, \Sigma^{-}, \tau)}.$$

Assim valores mais próximos de zero são indicativos de que se deve rejeitar a hipótese nula (Gelman 2003).

Para aplicar tal método a estimação da verossimilhança marginal é usualmente feita através de métodos numéricos como *Stepping Stone Sampling* que estima a log-verossimilhança marginal de um modelo (Xie et al 2011), por exemplo

$$\mathbf{P}(\boldsymbol{M}) \propto \int \mathbf{P}(X, Y|\Sigma, \tau, \boldsymbol{M}) \cdot \mathbf{P}(\Sigma) d\Sigma.$$

Esta estimação é feita através de uma integral de linha entre a distribuição a priori e a distribuição a posteriori da hipótese de interesse

$$L(\boldsymbol{M}|\boldsymbol{Y}) = \sum P(\boldsymbol{M}|\boldsymbol{Y}, \Sigma^{-1}, \tau)^{\beta} \cdot P(\boldsymbol{M}|\Sigma^{-1}, \tau),$$

Tabela 1: Resultados de Bayes Factor comparando modelos com diferentes ordenamentos e número de graus de liberdade da priori conjugada Wishart.

	Comparações					
	$N_3 \times O_2$	$N_3 \times O_3$	$N_3 \times O_6$	$O_2 \times O_3$	$O_2 \times O_6$	$O_3 \times O_6$
Média	-13.055	-12.99	-16.55	0.06159	-3.496	-3.577
Mediana	-12.506	-12.57	-16.1	0.02981	-3.561	-3.608
Máximo	-8.502	-8.2	-11.79	2.1348	-1.576	-2.525
Mínimo	-20.876	-21.13	-24.10	-0.7767	-4.047	-4.345
Desvio	2.8691	2.8421	2.7431	0.4619	0.4274	0.4041

onde $\beta \in [0, 1]$

Estudo de Simulação

Para este pequeno estudo foram geradas amostras de tamanho $n = 10$ de uma variável contínua e de uma variável discreta com $k = 3$ estados não ordenados e com raiz fixa. A partir destas amostras foi feita inferência da verossimilhança marginal de cada modelo condicionado aos dados observados, foram feitas $Re = 50$ repetições do experimento.

Para os dados gerados foi assumido que os estados eram independentes enquanto as correlação entre o primeiro estado e a variável contínua era $r = 0.5$ a correlação entre o segundo estado e a variável contínua era seu simétrico, $r = -0.5$.

Em seguida a estimação da log-verossimilhança marginal dos diferentes modelos, foi feita através do método de Stepping Stone Sampler usando MCMC e comparadas através do método de Bayes Factor

3 Resultados preliminares e Conclusões

A variabilidade das estimativas de verossimilhança marginal obtidas com o modelo nominal (sd = 1.998) é maior que as obtida a partir do modelo ordinal (sd ≤ 1.398). O que reflete o fato de que o espaço paramétrico do modelo ordinal é menor que o do modelo nominal com o mesmo número de categorias, indiferente aos dados observados.

Os resultados na Tabela 1 mostram que o BF favorece o modelo ordenado em todos casos, o que talvez seja associado ao pequeno tamanho de amostra.

Se percebe ainda que o modelo ordenado com maior número de graus de liberdade tende a ser favorecido.

4 Conclusões

Podemos concluir por este pequeno estudo que os resultados não se mantém, quando utilizados dados nominais.

Referências

[1] Cybis, G.B., Sinsheimer, J.S., Bedford, T., Mather, A.E., Lemey, P. and Suchard, M.A., Assessing phenotypic correlation through the multivariate phylogenetic latent liability model. The Annals of Applied Statistics, 9(2): 969-991. 2015.

[2] Drummond AJ, Rambaut A. Beast: Bayesian evolutionary analysis by sampling trees. BMC Evolutionary Biology, 7:214. 2007.

[3] Felsenstein J. Phylogenies and quantitative characters. Annual Review of Ecology and Systematics, 1:445-71. 1988.

[4] Felsenstein J. Using the Quantitative Genetics Threshold Model for Inferences Within and Between Species. Philosophical Transactions of the Royal Society B, 360:1427-1434. 2005.

[5] Kingman, J. F. C. The coalescent. Stochastic processes and their applications, 13(3):235-248. 1982.

[6] Xie, Wangang, Lewis, Paul O., Fan, Yu, Kuo, Lynn and Chen, Ming-Hui, Improving marginal likelihood estimation for bayesian phylogenetic model selection. Systematic Biology, 60(2): 150-160. 2011.

[7] Gelman, Andrew, Carlin, John B., Stern, Hal S. and Rubin, Donald B., Bayesian data analysis (2d ed). Chapman and Hall/CRC, 2003