



ENCOLHIMENTO BAYESIANO DE COEFICIENTES DE ONDALETAS SOB PRIORI DE CHAMPERNOWNE COM APLICAÇÕES

BAYESIAN WAVELET SHRINKAGE BASED ON CHAMPERNOWNE PRIOR WITH APPLICATIONS

Alex Rodrigo dos Santos Sousa¹

Submetido: 23/02/2021

Aprovado: 11/03/2021

RESUMO

Métodos estatísticos bayesianos para encolhimento de coeficientes de ondaletas (*wavelet shrinkage*) têm sido amplamente utilizados em diversas áreas para redução de ruídos em dados. Neste trabalho, propomos uma mistura da função delta de Dirac com a distribuição de Champernowne como distribuição probabilística a priori para os coeficientes das ondaletas em um problema de regressão não paramétrica. A regra bayesiana de encolhimento associada possui parâmetros de fácil interpretação e seu desempenho em estudos de simulação foi superior aos métodos disponíveis na literatura utilizados para comparação na maioria dos cenários considerados. Aplicações do método em dados de potenciais de ação neuronais e do índice da bolsa de valores de São Paulo (IBOVESPA) são feitas.

PALAVRAS-CHAVE: Estatística. Ondaletas. Distribuição de Champernowne. Regressão Não Paramétrica.

ABSTRACT

Bayesian wavelet shrinkage have been widely used in several areas to reduce noise in data analysis. In this work, we propose a mixture of a Dirac function with the Champernowne distribution as prior probabilistic distribution to wavelet coefficients in a nonparametric regression problem. The associated bayesian shrinkage rule has parameters that are easily interpreted and its performance in simulation studies was superior in most of the considered scenarios against methods available in the literature and used for comparison. Applications of the method to neuronal action potentials and the São Paulo Stock Market Index (IBOVESPA) datasets are made.

KEYWORDS: Statistics. Wavelets. Champernowne Distribution. Nonparametric Regression.

1. INTRODUÇÃO

Métodos baseados em ondaletas (*wavelets*) têm sido extremamente aplicados em diversas áreas da matemática, em processamento de sinais, compressão de imagens, entre outras. Na estatística, a possibilidade de analisar uma série de dados no domínio das ondaletas por meio de sua transformada discreta (DWT, sigla em inglês) permite uma representação esparsa das observações, uma vez que tipicamente os coeficientes das ondaletas são significativamente diferentes de zero apenas em localizações que representam características importantes dos dados, como mudanças de comportamento, picos, descontinuidades, entre outras. Além disso, é possível visualizar os dados no

¹ Professor de Estatística - Centro Universitário Campo Limpo Paulista - alex.sousa89@usp.br - <https://orcid.org/0000-0001-5887-3638>



domínio das ondaletas em níveis de resolução, de acordo com os princípios da Análise de Multirresolução de Mallat. Para uma completa descrição teórica sobre ondaletas com aplicações em estatística, ver VIDA KOVIC (1999), MALLAT (2008) e MORETTIN (2014).

Em análise de dados, porém, é muito comum a presença de ruídos aleatórios que contaminam seus valores, o que dificulta a estimação de parâmetros de interesse ou até mesmo de funções de interesse, como será o foco deste artigo. E a presença de ruído nos dados permanece no domínio das ondaletas, o que pode ser notado ao verificar uma quantidade significativa de coeficientes das ondaletas, denominados empíricos (observados), não nulos, porém muito próximos de zero. Quando isto ocorre, a aplicação de métodos de encolhimento de coeficiente de ondaletas (*wavelet shrinkage*) ou de escolha de limiar (*thresholding*) é essencial na análise, por reduzir o efeito do ruído nos dados por meio da redução da magnitude dos coeficientes empíricos de ondaletas. Existem, de fato, diversos métodos de encolhimento ou de escolha de limiar propostos na literatura desde os anos 1990, ver DONOHO e JOHNSTONE (1994,1995), DONOHO (1995,1996) e MORETTIN (2014).

Métodos bayesianos de encolhimento de coeficientes de ondaletas (*wavelet shrinkage*) têm tido destaque na área de encolhimento de coeficientes de ondaletas pois permitem incorporar informações prévias sobre os coeficientes da função geradora dos dados desconhecida (a ser estimada) transformada para o domínio das ondaletas, tais como suavidade (esparsidade), suporte (se forem limitados), características locais como picos, descontinuidades ou oscilações por exemplo, entre outros.

Vários procedimentos de encolhimento bayesiano foram estudados e propostos nos últimos anos na Estatística. Alguns deles são encontrados em LIAN (2011), BERGER et al. (2012), KARAGIANNIS et al. (2015), GRIFFIN e BROWN (2017) e TORKAMANI e SADEGHZADEH (2017). Nesse sentido, a escolha da distribuição de probabilidade a priori para os coeficientes de ondaletas é fundamental para se ter bons resultados na análise bayesiana. Diversas distribuições a priori foram propostas com sucesso, algumas delas podem ser vistas em CHIPMAN et al (1997), ABRAMOVICH, SAPATINAS e SILVERMAN (1998), VIDA KOVIC (1998), VIDA KOVIC e RUGGERI (2001), ANGELINI e VIDA KOVIC (2004), JOHNSTONE e SILVERMAN (2004, 2005), REMÉNYI e VIDA KOVIC (2015), BHATTARCHARYA et al. (2015) entre outros, porém, como em qualquer campo de pesquisa, o estudo e proposição de novas distribuições à priori é sempre bem-vindo. Desta forma, propomos neste trabalho a utilização de uma mistura de uma função delta de Dirac com a distribuição de *Champernowne* como distribuição a priori para os coeficientes de ondaletas.

A distribuição de *Champernowne* foi proposta por CHAMPERNOWNE (1953) com o objetivo de modelar dados econométricos e seu uso é muito interessante no contexto da estimação dos coeficientes de ondaletas, uma vez que é simétrica (em torno de zero em nossa proposta) e seu parâmetro de escala é facilmente relacionado com o grau de encolhimento da regra bayesiana de encolhimento associada, o que pode ser útil para determiná-lo em situações práticas. Além disso, possui como casos particulares distribuições de probabilidade bastante conhecidas, como as



distribuições logística e de Burr. Por fim, os desempenhos em termos de erro quadrático médio (EQM) da regra de encolhimento sob priori *Champernowne* foram melhores do que alguns dos métodos de encolhimento/limiar mais usados atualmente.

Este artigo está organizado da seguinte forma: a Seção 2 define o modelo estatístico de regressão não paramétrica utilizado e a distribuição a priori com base na distribuição de *Champernowne* proposta para os coeficientes de ondaletas, a Seção 3 estabelece a regra de encolhimento associada e discute como determinar os parâmetros e hiperparâmetros envolvidos. Estudos de simulação para avaliar desempenhos são feitos na Seção 4 e aplicações da regra de encolhimento proposta em um conjunto de dados de potenciais neuronais e em uma série de Índices da Bolsa de Valores de São Paulo (IBOVESPA) são feitas na Seção 5. O artigo termina com a conclusão e as considerações finais na Seção 6.

2. O MODELO ESTATÍSTICO

Considere $n = 2^J$ ($J \in \mathbb{N}$) observações y_1, \dots, y_n obtidas nos tempos x_1, \dots, x_n de acordo com o seguinte modelo de regressão não paramétrica:

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1, \dots, n \quad (1)$$

em que f é uma função desconhecida a ser estimada e $\varepsilon_1, \dots, \varepsilon_n$ são variáveis aleatórias independentes normalmente distribuídas com média igual a zero e variância constante (ruído branco), isto é, $E(\varepsilon_i) = 0$ e $Var(\varepsilon_i) = \sigma^2$, $\sigma > 0$. Podemos reescrever o modelo (1) em notação vetorial, isto é,

$$\mathbf{y} = \mathbf{f} + \boldsymbol{\varepsilon}, \quad (2)$$

para $\mathbf{y} = [y_1, \dots, y_n]'$, $\mathbf{f} = [f(x_1), \dots, f(x_n)]'$ e $\boldsymbol{\varepsilon} = [\varepsilon_1, \dots, \varepsilon_n]'$. A transformada discreta de ondaletas pode ser representada pela sua matriz de transformação \mathbf{W} , que é ortogonal e dependente da base de ondaletas utilizada. Desta forma, o modelo (2) pode ser transformado linearmente para o domínio das ondaletas fazendo:

$$\begin{aligned} \mathbf{W}\mathbf{y} &= \mathbf{W}(\mathbf{f} + \boldsymbol{\varepsilon}) \\ \mathbf{W}\mathbf{y} &= \mathbf{W}\mathbf{f} + \mathbf{W}\boldsymbol{\varepsilon} \\ \mathbf{d} &= \boldsymbol{\theta} + \boldsymbol{\varepsilon}, \quad (3) \end{aligned}$$

em que $\mathbf{d} = [d_1, \dots, d_n]'$ = $\mathbf{W}\mathbf{y}$ é o vetor de coeficientes de ondaletas empíricos (observados), $\boldsymbol{\theta} = [\theta_1, \dots, \theta_n]'$ = $\mathbf{W}\mathbf{f}$ é o vetor de coeficientes de ondaletas da função desconhecida e $\boldsymbol{\varepsilon} = [\varepsilon_1, \dots, \varepsilon_n]'$ = $\mathbf{W}\boldsymbol{\varepsilon}$ é o vetor com os ruídos transformados. Uma vez que a DWT é ortogonal, os ruídos transformados preservam a distribuição gaussiana com média igual a zero e mesma variância dos ruídos do modelo original. Portanto, o problema de estimar a função f no domínio do tempo é convertido no problema de estimar o vetor $\boldsymbol{\theta}$ dos seus coeficientes de ondaletas. Além disso, temos que:

$$d_i \sim N(\theta_i, \sigma^2), \quad (4)$$

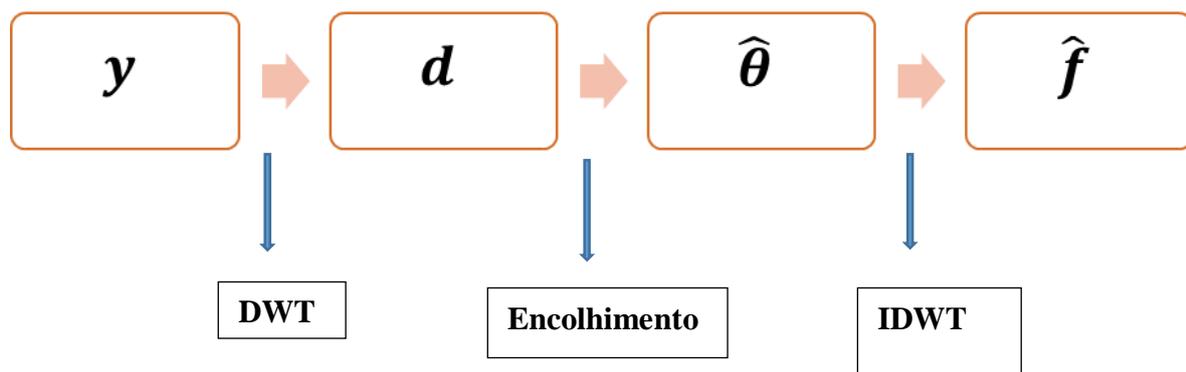


ou seja, o problema de estimar os coeficientes de ondaletas da função desconhecida é equivalente ao problema de estimar o parâmetro de locação de uma distribuição gaussiana. Como a estimação é realizada individualmente coeficiente por coeficiente e para simplificar a notação, os subíndices do modelo serão omitidos, sem perda de interpretação.

Desta forma, podemos interpretar um coeficiente empírico d como um coeficiente de ondaletas θ contaminado aditivamente por um ruído aleatório ϵ . Assim, para estimar θ tipicamente é aplicado em d uma regra de encolhimento ou de limiar $\delta(d)$ que minimiza o efeito do ruído reduzindo a magnitude de d , de modo que o estimador $\hat{\theta}$ de θ é dado por $\hat{\theta} = \delta(d)$. Em termos gerais, coeficientes empíricos próximos de zero são encolhidos a zero pelas regras de encolhimento ou limiar. O quão próximo de zero é necessário para a regra encolher um coeficiente empírico determina o grau de encolhimento da regra e pode ser controlado por seus parâmetros.

Finalmente, a partir do vetor $\hat{\theta}$ de coeficientes de ondaletas estimados (usualmente um vetor esparsos), a função f pode ser estimada pela transformada discreta de ondaletas inversa (IDWT, na sigla em inglês), isto é, $\hat{f} = W^T \hat{\theta}$. A Figura 1 apresenta esquematicamente o processo de estimação por encolhimento de coeficientes de ondaletas.

Figura 1: Processo de aplicação das transformadas discretas de ondaletas DWT e IDWT para estimação \hat{f} de pontos de uma função f .



Desta forma, propomos um processo de estimação bayesiana para θ por meio da seguinte distribuição a priori:

$$\pi(\theta; \tau, \alpha, \lambda) = \tau \delta_0(\theta) + (1 - \tau) g(\theta; \alpha, \lambda), \quad (5)$$



RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR

ENCOLHIMENTO BAYESIANO DE COEFICIENTES DE ONDALETAS SOB
PRIORI DE CHAMPERNOWNE COM APLICAÇÕES
Alex Rodrigo dos Santos Sousa

em que $\tau \in (0,1)$, $\delta_0(\cdot)$ é a função delta de Dirac centralizada em zero e $g(\cdot; \alpha, \lambda)$ é a função densidade de Champernowne, definida por:

$$g(\theta; \alpha, \lambda) = \frac{C(\alpha, \lambda)}{\cosh(\alpha\theta) + \lambda}, \quad \theta \in \mathbb{R}, \quad (6)$$

para $\alpha, \lambda > 0$ e $C(\alpha, \lambda)$ é a constante de normalização, determinada numericamente. Logo, a distribuição a priori proposta possui os hiperparâmetros τ, α e λ , que devem ser determinados previamente à análise. Veremos que tais hiperparâmetros estão fortemente relacionados com o grau de suavização da regra de Bayes (estimador) associada aos modelos probabilísticos (4), (5) e (6). Os Gráficos 1 e 2 apresentam os gráficos de densidades de Champernowne para alguns valores atribuídos para os hiperparâmetros α e λ . Observe ainda que a distribuição de Champernowne possui casos particulares conhecidos de acordo com os valores de seus parâmetros. Para $\alpha = \lambda = 1$, a distribuição de Champernowne é a distribuição logística e quando $\lambda = 1$, a distribuição é a distribuição de Burr.

A ideia principal da distribuição à priori proposta neste trabalho é que um coeficiente de ondaleta é nulo com um peso τ ou gerado pela distribuição de Champernowne simétrica em zero com peso $1 - \tau$. Desta forma, quanto maior o valor do hiperparâmetro τ , maior será o grau de encolhimento da regra bayesiana de encolhimento associada. Além disso, observe nos Gráficos 1 e 2 que tanto o hiperparâmetro α como λ controlam a variabilidade da distribuição de Champernowne. De fato, quanto maior o valor de α (pelo Gráfico 1) e menor o valor de λ (pelo Gráfico 2), mais concentrada é a distribuição de Champernowne em torno de zero. Este fato está intimamente relacionado com o grau de encolhimento da regra associada, uma vez que o grau aumenta à medida que a distribuição possui menor variabilidade, isto é, quando é mais concentrada em zero.

Como os dois hiperparâmetros da distribuição de Champernowne controlam a variabilidade da distribuição, propomos neste trabalho fixar o valor de λ para $\lambda = 1$ e estudar o controle dos valores de α no grau de encolhimento da regra bayesiana vinculada.

Gráfico 1: Densidades de Champernowne para $\lambda = 1$ e $\alpha \in \{0,5; 1; 2; 3; 4\}$.

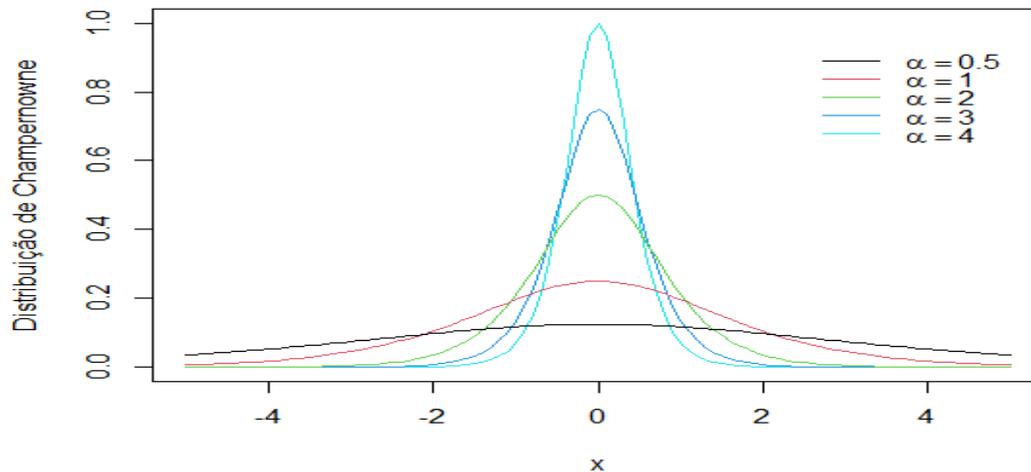
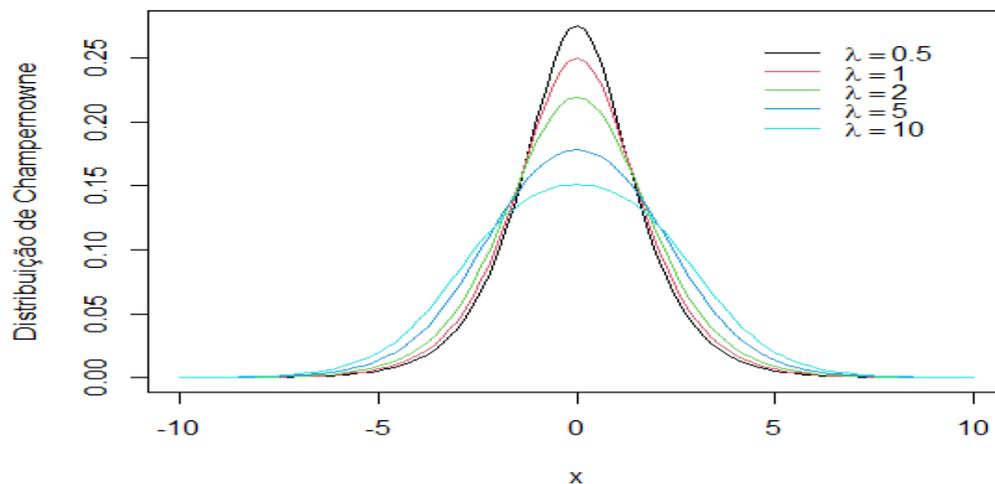


Gráfico 2: Densidades de Champernowne para $\alpha = 1$ e $\lambda \in \{0,5; 1; 2; 5; 10\}$.



3. A REGRA BAYESIANA DE ENCOLHIMENTO E DETERMINAÇÃO DOS PARÂMETROS

A regra bayesiana de encolhimento δ associada aos modelos (4), (5) e (6) a ser utilizada para estimar os coeficientes de ondaletas é determinada supondo função de perda quadrática,



$L(\delta, \theta) = (\delta - \theta)^2$. Neste caso, a regra de Bayes que minimiza tal função perda é o valor esperado a posteriori de θ , isto é:

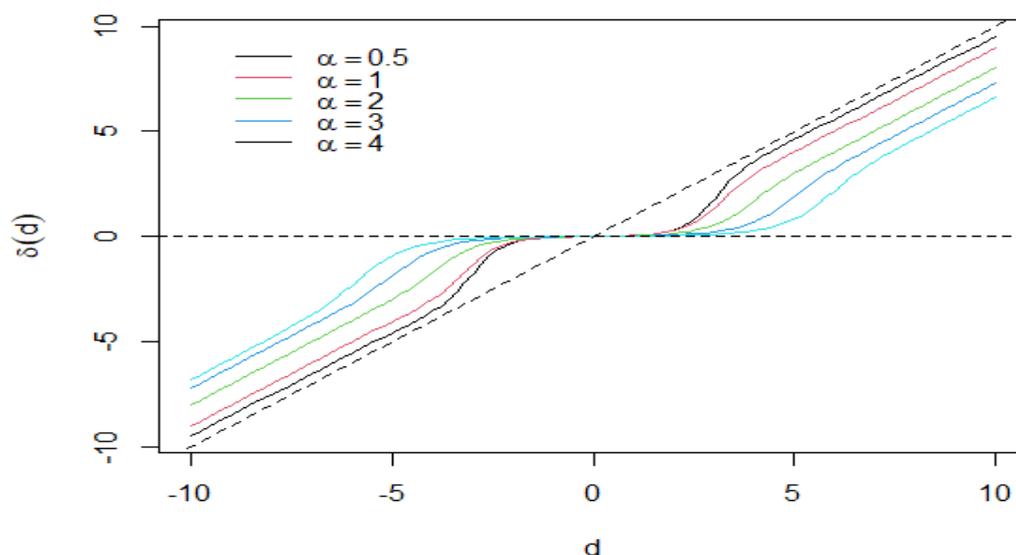
$$\delta = \delta(d) = E_{\pi}(\theta|d), \quad (7)$$

ver ROBERT (2007) para mais detalhes. Logo, utilizando os modelos (4), (5) e (6) para obter a regra bayesiana de encolhimento em (7), obtemos:

$$\delta(d) = \frac{(1 - \tau) \int_{-\infty}^{\infty} (\sigma u + d) g(\sigma u + d; \alpha, \lambda) \phi(u) du}{\frac{\tau}{\sigma} \phi\left(\frac{d}{\sigma}\right) + (1 - \tau) \int_{-\infty}^{\infty} g(\sigma u + d; \alpha, \lambda) \phi(u) du}, \quad (8)$$

em que $\phi(\cdot)$ é a função densidade da distribuição normal padrão. De fato, as integrais envolvidas em (8) não são facilmente tratáveis, o que geralmente ocorre em procedimentos bayesianos. Assim, a regra de encolhimento é obtida por meio de métodos computacionais, como métodos MCMC (*Markov Chain Monte Carlo*), ver ROBERT e CASELLA (2005) para mais detalhes. Os Gráficos 3 e 4 apresentam as regras de encolhimento (8) para valores específicos dos hiperparâmetros α e τ respectivamente.

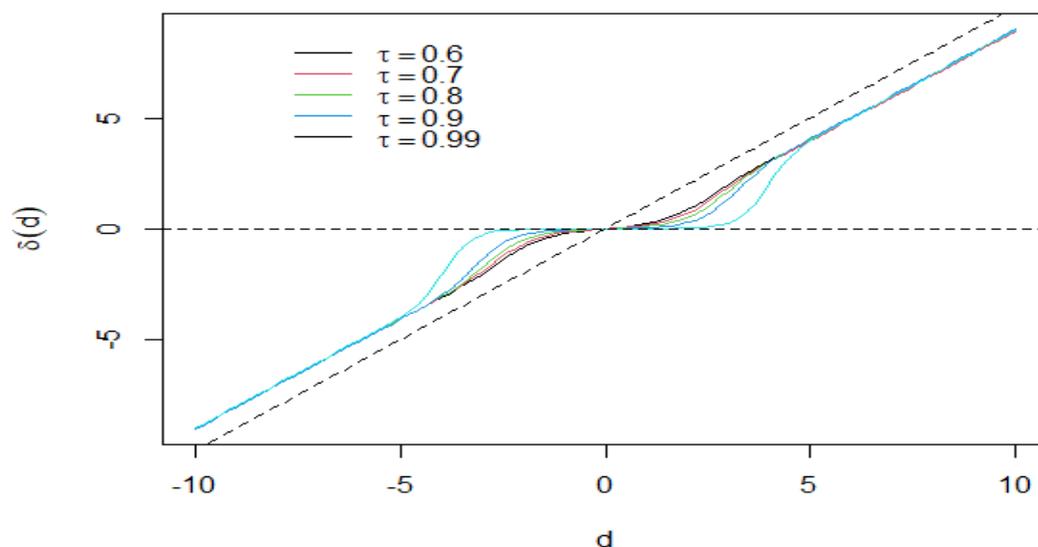
Gráfico 3: Regra de encolhimento (8) para $\tau = 0,9, \lambda = 1$ e $\alpha \in \{0,5; 1; 2; 3; 4\}$.





Observe nos Gráficos 3 e 4 que para valores do coeficiente empírico d suficientemente próximos de zero, a regra encolhe tais valores à zero. Como já esperado, o grau de encolhimento da regra aumenta à medida que α (Gráfico 3) ou τ (Gráfico 4) aumenta, o que pode ser visto pelo aumento do comprimento da faixa de valores de d passíveis de encolhimento a zero pela regra.

Gráfico 4: Regra de encolhimento (8) para $\alpha = \lambda = 1$ e $\tau \in \{0,6; 0,7; 0,8; 0,9; 0,99\}$.



Para a aplicação da regra de encolhimento para estimar os coeficientes de ondaletas, os valores dos parâmetros e hiperparâmetros devem ser previamente estabelecidos.

Para estimar o parâmetro σ , DONOHO e JOHNSTONE (1994a) sugerem a seguinte estimação robusta:



$$\hat{\sigma} = \frac{\text{mediana}\{d_{j-1,k} : k = 0, \dots, 2^{j-1}\}}{0,6745}, \quad (9)$$

para o hiperparâmetro τ , ANGELINI e VIDAKOVIC (2004) sugerem:

$$\tau = \tau(j) = 1 - \frac{1}{(j - J_0 + 1)^\gamma}, \quad (10)$$

em que $J_0 \leq j \leq J - 1$ e J_0 é o nível de resolução primária. Além disso, caso não haja informações adicionais, o valor $\gamma = 2$ pode ser adotado.

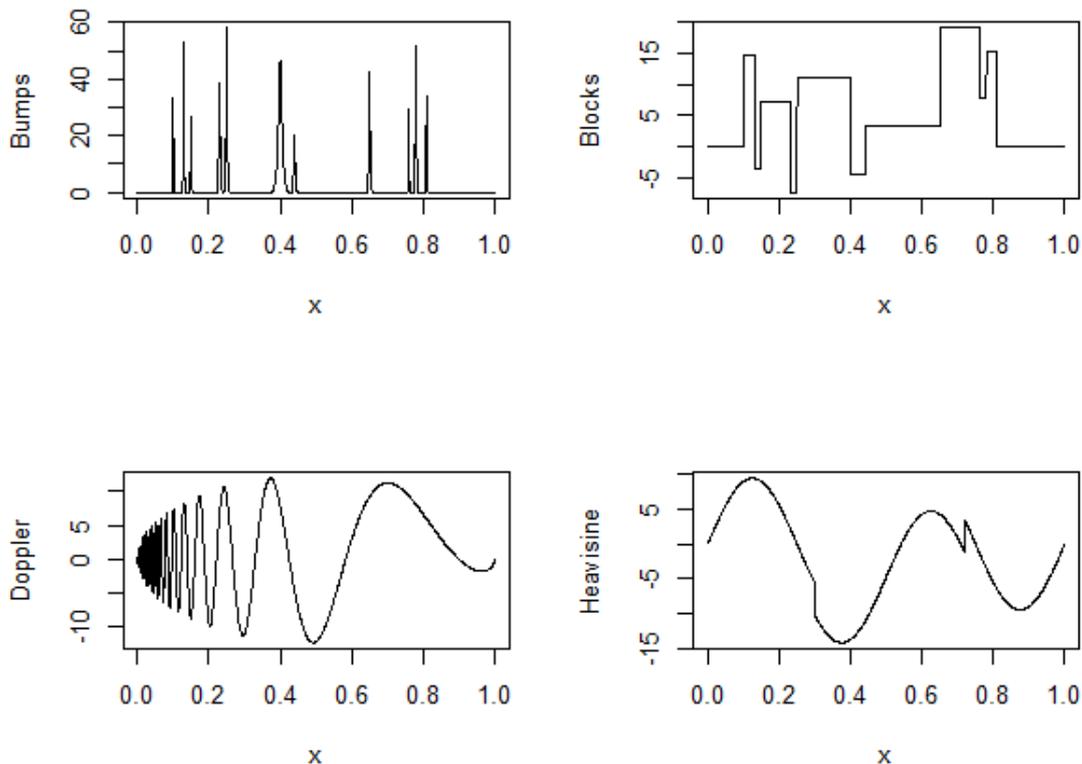
Para determinação do hiperparâmetro α , sugerimos a aplicação da regra para um conjunto de valores, $\alpha \in \{0,5; 1; 2; 3; 4; 5\}$ e verificar o desempenho da regra nos dados específicos sob análise. Como este hiperparâmetro está associado ao grau de encolhimento da regra, indicamos a utilização de valores maiores de α para aplicação da regra em dados com maior presença de ruído, isto é, com baixa razão entre sinal-ruído (SNR, sigla em inglês).

4. SIMULAÇÕES COMPUTACIONAIS

Para avaliar a performance da regra de encolhimento proposta e comparar com métodos de encolhimento e de limiar disponíveis na literatura e aplicados em problemas reais, um estudo com simulações computacionais foi conduzido. Para tanto, as funções testes de Donoho e Johnstone (D-J) foram utilizadas como geradoras dos dados simulados. As funções testes de D-J são amplamente aplicadas em estudos sobre ondaletas por possuírem características importantes a serem detectadas, como pontos de máximo e mínimo locais, picos, descontinuidades e oscilações por exemplo e são compostas por quatro funções, Bumps, Blocks, Doppler e Heavisine. O Gráfico 5 apresenta estas quatro funções.



Gráfico 5: Funções testes de Donoho e Johnstone.



Para cada função teste, $n = 512 = 2^9$ pontos igualmente espaçados foram selecionados e, em cada ponto da função foi adicionado um erro aleatório normalmente distribuído de acordo com o modelo (1). O erro aleatório foi gerado com média igual a zero e variância σ^2 de acordo com três valores de razão sinal-ruído (SNR, sigla em inglês), 3, 5 e 7. Quanto menor o valor de SNR, mais ruído está presente nos dados, de modo que a estimação da função subjacente se torna mais complicada. Após a geração do conjunto de dados e aplicação da DWT para transformá-los para o domínio das ondaletas, aplicamos a regra de encolhimento proposta e, para comparação de performance, quatro métodos de encolhimento aplicados na literatura, a saber, limiar universal (UNIV) proposto por DONOHO e JOHNSTONE (1994b), validação cruzada (CV, sigla em inglês) proposto por NASON (1996), taxa de falsa descoberta (FDR, sigla em inglês) de ABRAMOVICH e BENJAMINI (1996) e o estimador não viesado do risco de Stein (SURE, sigla em inglês) de DONOHO e JOHNSTONE (1995). Após a aplicação dos métodos de encolhimento e limiar, a transformada inversa IDWT é utilizada para enfim estimar os pontos da função geradora. Para cada função, este procedimento foi replicado $M = 200$ vezes. Em cada replicação k , o erro quadrático médio EQM_k foi calculado para cada método, em que:



RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR

ENCOLHIMENTO BAYESIANO DE COEFICIENTES DE ONDALETAS SOB
PRIORI DE CHAMPERNOWNE COM APLICAÇÕES
Alex Rodrigo dos Santos Sousa

$$EQM_k = \frac{1}{n} \sum_{i=1}^n (\hat{f}(x_i) - f(x_i))^2, \quad (11)$$

de modo que a medida de performance utilizada para comparar os métodos é a média dos erros quadráticos médios das replicações (*MEQM*),

$$MEQM = \frac{1}{M} \sum_{k=1}^M EQM_k, \quad (12)$$

e quanto menor for o *MEQM*, melhor é a performance do método.

As Tabelas 1, 2 e 3 apresentam os *MEQMs* obtidos no estudo de simulação para SNR igual a 3, 5 e 7 respectivamente com a melhor performance destacada em negrito para cada função.

Tabela 1: Média dos erros quadráticos médios (MEQM) para simulações - SNR = 3.

Método	Bumps	Blocks	Doppler	Heavisine
UNIV	11,104	6,896	2,646	0,566
CV	11,451	2,543	1,276	0,505
FDR	9,345	5,869	2,559	0,593
SURE	3,635	2,769	1,331	0,569
CHAMP	3,226	2,542	1,252	0,718

Tabela 2: Média dos erros quadráticos médios (MEQM) para simulações - SNR = 5.

Método	Bumps	Blocks	Doppler	Heavisine
UNIV	5,179	3,669	1,397	0,407
CV	9,502	1,248	0,649	0,280
FDR	4,375	2,937	1,256	0,437
SURE	1,577	1,234	0,591	0,416
CHAMP	1,121	0,962	0,485	0,300

Tabela 3: Média dos erros quadráticos médios (MEQM) para simulações - SNR = 7.

Método	Bumps	Blocks	Doppler	Heavisine
UNIV	3,031	2,256	0,885	0,302
CV	6,312	0,835	0,440	0,174
FDR	2,621	1,724	0,758	0,312
SURE	0,889	0,680	0,330	0,315
CHAMP	0,600	0,484	0,242	0,165

Os Gráficos 6, 7 e 8 mostram os boxplots dos erros quadráticos médios das replicações para SNR = 3, 5 e 7 respectivamente.



Gráfico 6: Boxplots dos erros quadráticos médios (EQM) das replicações dos estudos de simulação para SNR = 3. No eixo horizontal, 1-UNIV, 2-CV, 3-FDR, 4-SURE e 5-CHAMP.

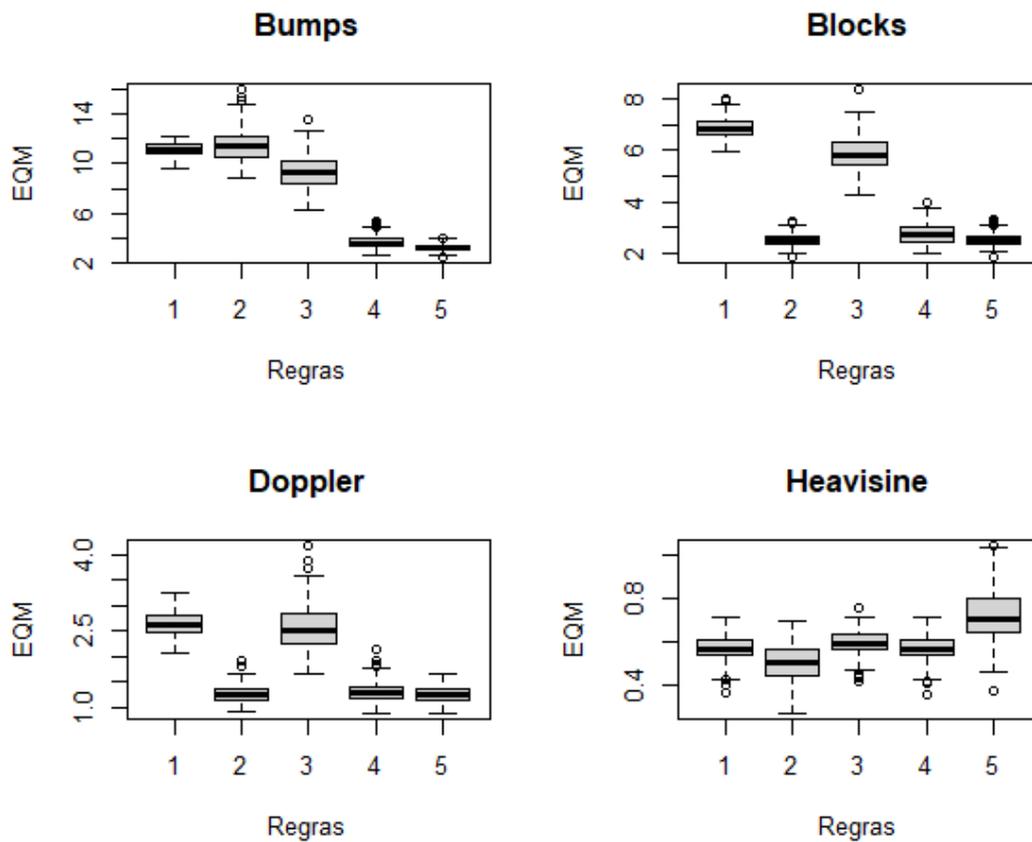
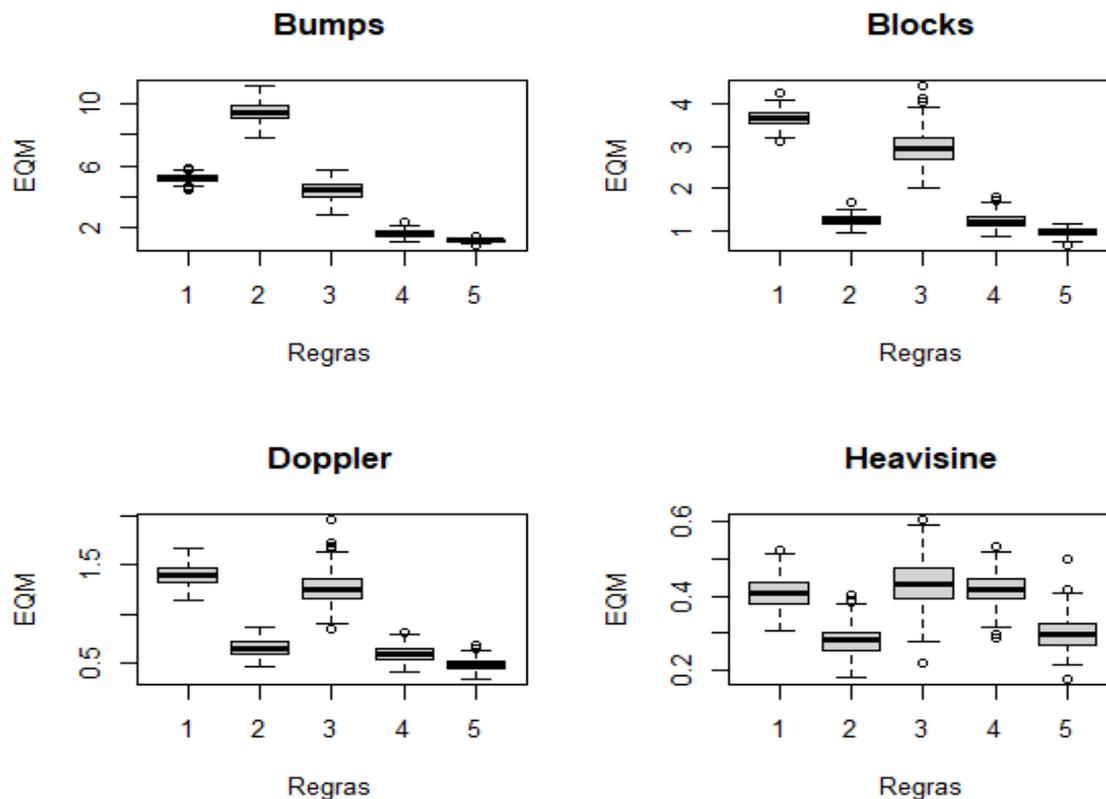


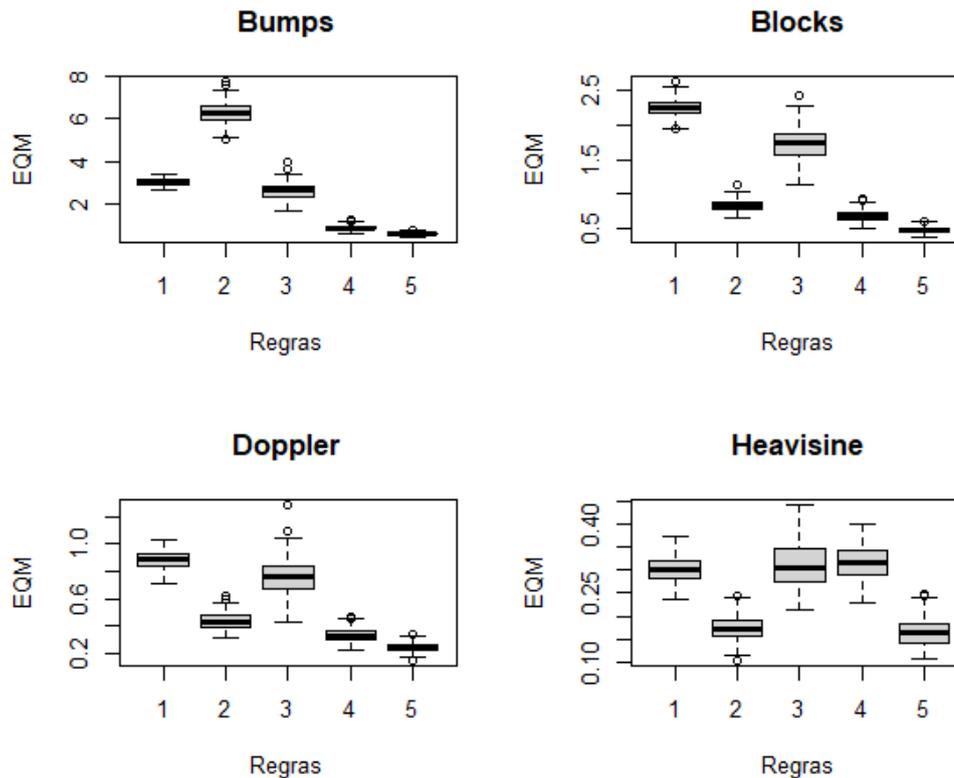
Gráfico 7: Boxplots dos erros quadráticos médios (EQM) das replicações dos estudos de simulação para SNR = 5. No eixo horizontal, 1-UNIV, 2-CV, 3-FDR, 4-SURE e 5-CHAMP.



Pelas tabelas e boxplots podemos observar que a regra de encolhimento proposta teve melhor desempenho na maioria dos cenários do estudo de simulação. Para as funções Bumps, Blocks e Doppler, a regra teve o menor *MEQM* em todos os cenários de SNR. Este fato é importante por mostrar que a regra tem bom desempenho mesmo com forte presença de ruído nos dados, representado quando $\text{SNR} = 3$. Além disso, os boxplots mostram que as variabilidades dos *EQMs* foram pequenas e com poucos valores discrepantes, o que evidencia consistência do desempenho do estimador proposto. Outro ponto importante a destacar é que, mesmo no caso em que a regra não obteve o melhor desempenho global, o que ocorreu somente com a função Heavisine, seu desempenho foi satisfatório, isto é, apesar de não ter sido o melhor, ficou entre os melhores ou próximo do melhor.

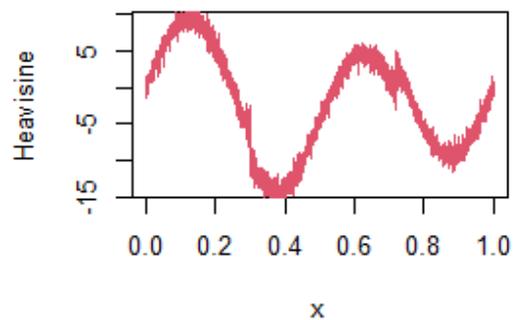
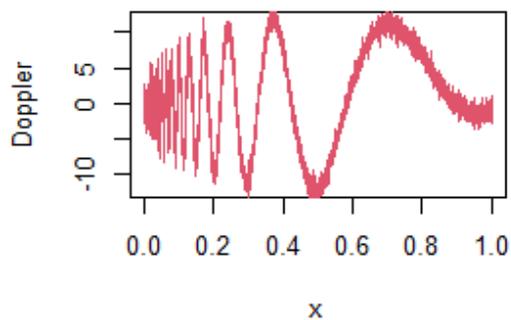
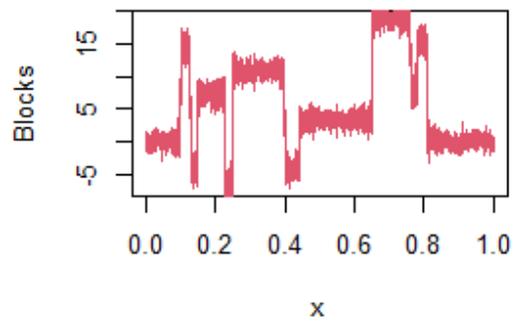
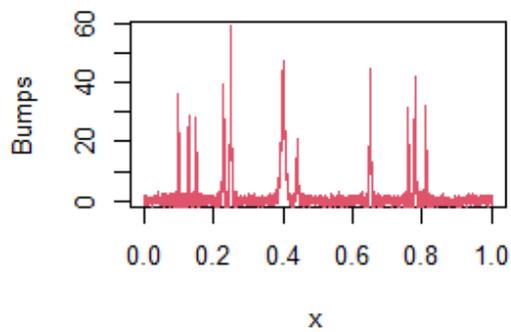


Gráfico 8: Boxplots dos erros quadráticos médios (EQM) das replicações dos estudos de simulação para SNR = 7. No eixo horizontal, 1-UNIV, 2-CV, 3-FDR, 4-SURE e 5-CHAMP.



Como ilustração das estimativas, o Gráfico 9 apresenta as curvas estimadas pela regra proposta para cada função em cada replicação para o cenário em que SNR = 7. Note que as principais características das funções subjacentes foram captadas e estimadas pela regra de encolhimento sob priori Champernowne. A generalidade típica das funções-testes de D-J em termos de características mostra que a regra pode ser bem-sucedida para aplicação em problemas de estimação de curvas com os mais variados perfis.

Gráfico 9: Curvas estimadas pela regra de encolhimento sob priori Champernowne no estudo de simulação para SNR = 7.



5. APLICAÇÕES EM DADOS REAIS

Aplicamos a regra de encolhimento sob priori Champernowne em dois conjuntos de dados reais para analisar seu potencial em problemas de interesse em outras áreas. O primeiro conjunto de dados (Aplicação 1) contém potenciais de ação neuronais coletados para classificação de pulsos cerebrais. Em neurociência, este problema de classificação é conhecido por *spike sorting*. O segundo conjunto de dados (Aplicação 2) refere-se aos índices da Bolsa de Valores de São Paulo, IBOVESPA.

5.1 APLICAÇÃO 1: POTENCIAIS DE AÇÃO NEURONAIS



RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR

ENCOLHIMENTO BAYESIANO DE COEFICIENTES DE ONDALETAS SOB
PRIORI DE CHAMPERNOWNE COM APLICAÇÕES
Alex Rodrigo dos Santos Sousa

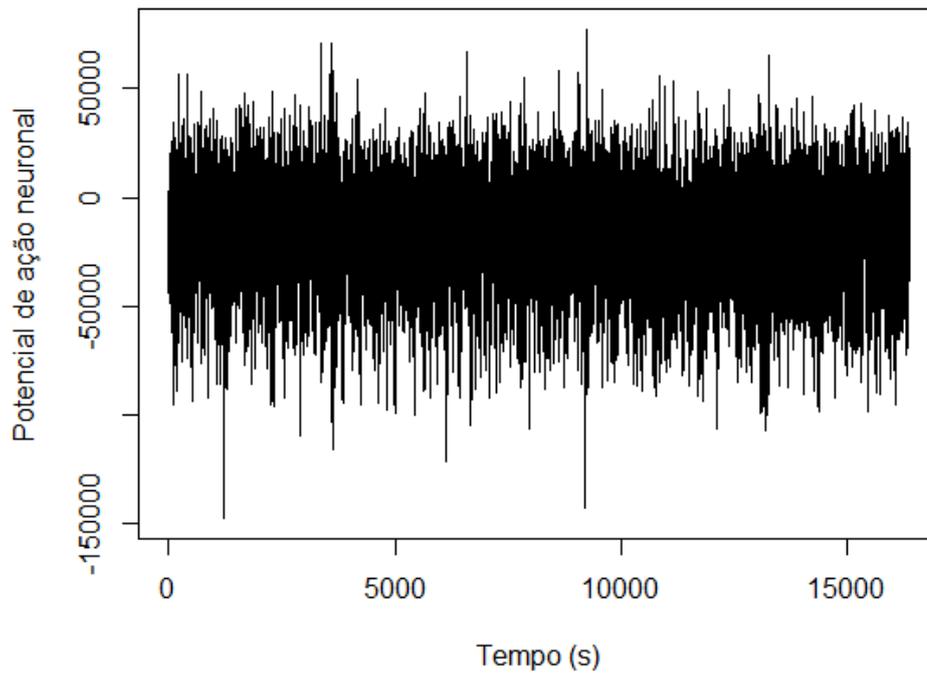
Spike sorting é um procedimento de classificação de pulsos cerebrais aplicado em neurociência com diversos objetivos, como detecção de doenças neurológicas por exemplo. Para a realização deste procedimento de forma não invasiva, eletrodos são instalados na cabeça de uma pessoa e potenciais de ação (*spikes*) são coletados por um intervalo de tempo.

Uma vez com os dados coletados, a primeira etapa do procedimento de *spike sorting* é remover o ruído presente nos dados, o que ocorre principalmente devido aos instrumentos de medição utilizados. Existem diversos métodos estatísticos para realização deste pré-processamento dos dados, como pode ser visto em LEWICKI (1998), POUZAT et al. (2002), SHOHAN et al. (2003) e EINEVOLL et al. (2012). Ondaletas, entretanto, têm sido um dos métodos mais aplicados devido às suas características de remoção de ruído, como detalhado em WEBER (2000), QUIROGA et al. (2004) e SHALCHYAN et al. (2012).

Neste sentido, aplicamos a regra de encolhimento sob priori Champernowne em um conjunto de dados de potenciais de ação neuronais com o intuito de minimizar os efeitos dos ruídos presentes nos dados. Os dados foram coletados por Kenneth Harris, do *Institute of Neurology, Faculty of Brain Sciences, University College London* e estão disponíveis publicamente em <https://ifcs.boku.ac.at/repository/data/spike-sorting/index.html>. Utilizamos $n = 16384 = 2^{14}$ potenciais de ação do conjunto de dados, que estão apresentados no Gráfico 10.



Gráfico 10: Potenciais de ação neuronais utilizados para aplicação da regra de encolhimento sob priori Champernowne para remoção de ruído.



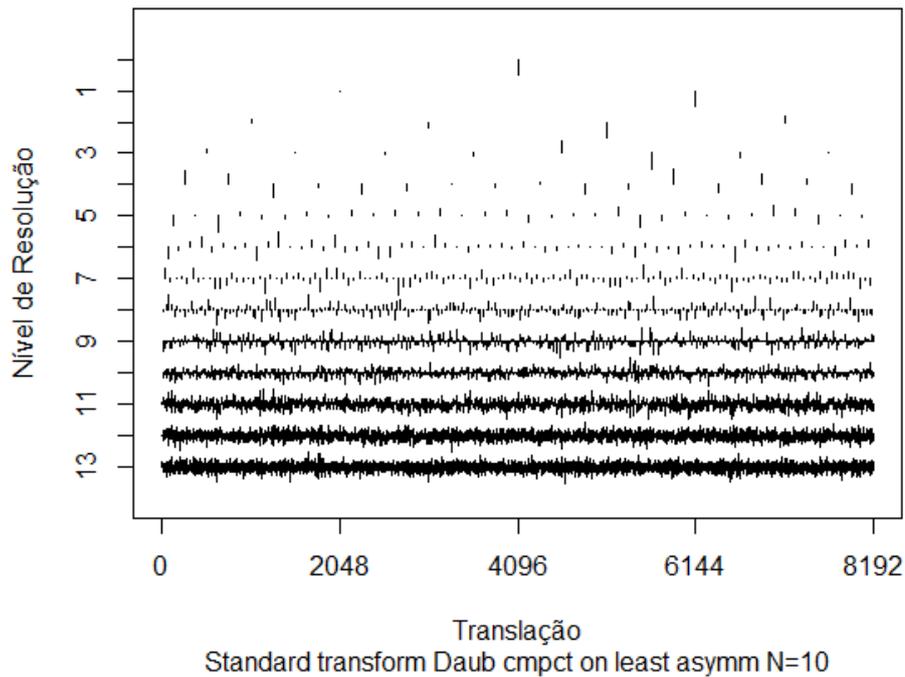
Ao conjunto de dados brutos, aplicamos uma DWT com base Daubechies com 10 momentos nulos (Daub10). O Gráfico 11 mostra os coeficientes de ondaletas empíricos obtidos por nível de resolução. Note que os coeficientes possuem considerável ruído principalmente nos níveis mais altos de resolução.



RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR

ENCOLHIMENTO BAYESIANO DE COEFICIENTES DE ONDALETAS SOB
PRIORI DE CHAMPERNOWNE COM APLICAÇÕES
Alex Rodrigo dos Santos Sousa

Gráfico 11: Coeficientes de ondaletas empíricos dos dados de potenciais de ação neuronais obtidos após aplicação de DWT sob base Daub10.



Aos coeficientes empíricos obtidos aplicamos a regra de encolhimento sob priori Champernowne, com σ e τ determinados de acordo com (9) e (10), $\alpha = 0,5$ e $\lambda = 1$. Os Gráficos 12 e 13 apresentam, respectivamente, os potenciais estimados após eliminação de ruído e os coeficientes de ondaletas estimados. De fato, após a minimização dos efeitos do ruído, os potenciais significativos (com altas amplitudes) ficaram evidentes, o que facilita o processo de classificação. Além disso, os coeficientes de ondaletas estimados são, em sua maioria, praticamente nulos, de modo que os coeficientes com maiores magnitudes representam localmente características importantes nos dados, como os spikes com amplitudes significativas.



Gráfico 12: Potenciais de ação neuronais estimados após remoção do ruído pela regra de encolhimento sob priori Champernowne.

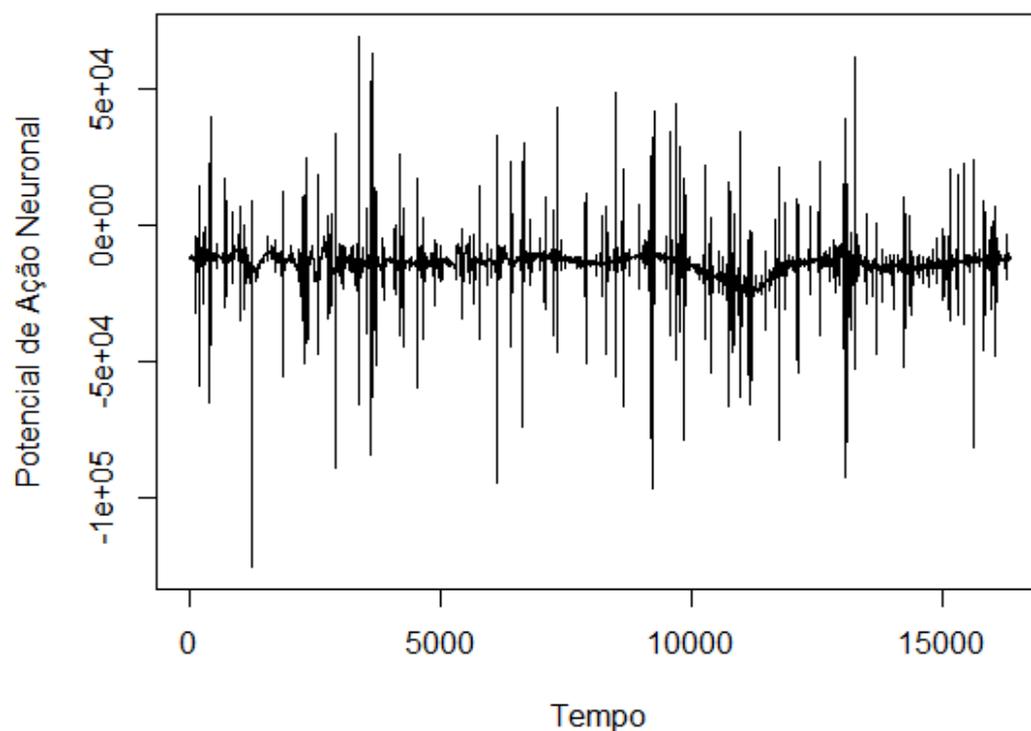
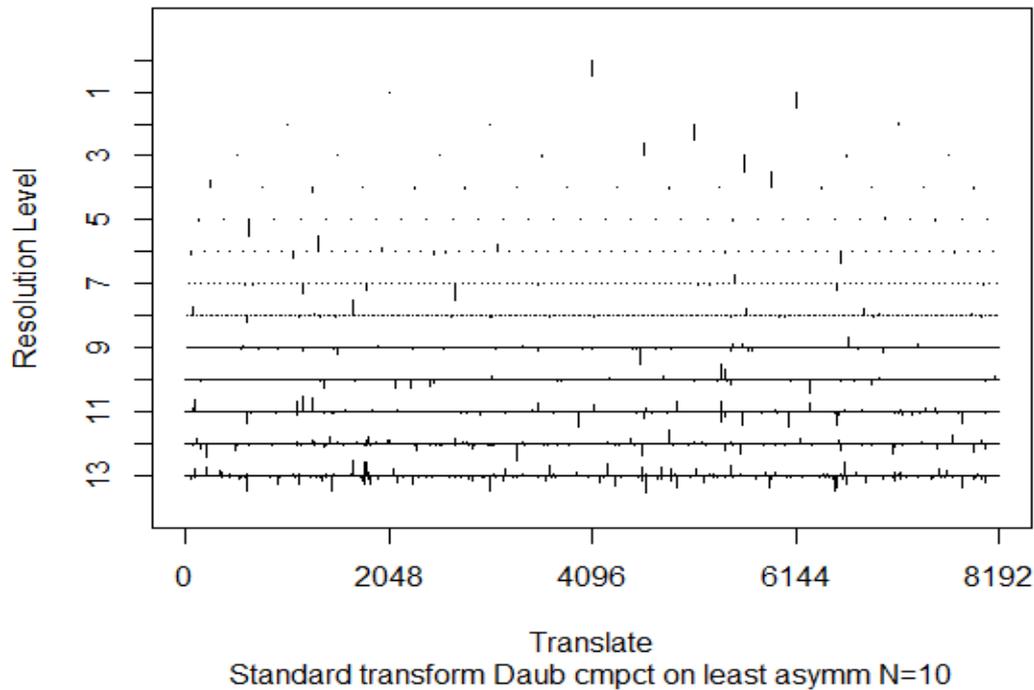




Gráfico 13: Coeficientes de ondaletas estimados pela regra de encolhimento sob priori Champernowne.

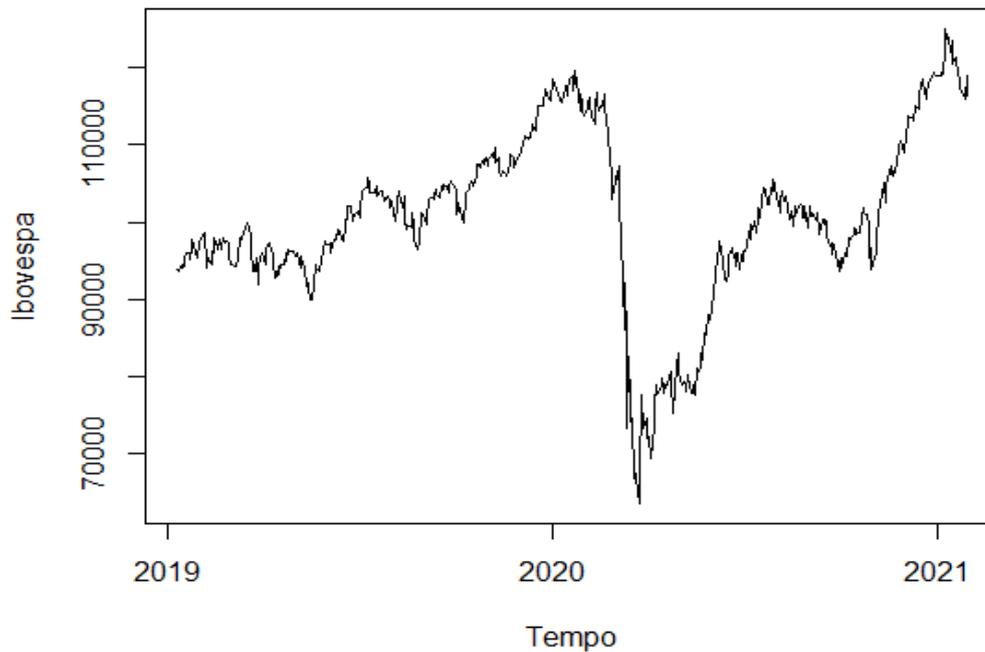


5.2 APLICAÇÃO 2: IBOVESPA

Na aplicação aos dados de potenciais de ação neuronais, o valor do hiperparâmetro α selecionado foi pequeno devido à necessidade de maior grau de encolhimento do estimador, uma vez que os dados apresentavam alto ruído. Em conjunto de dados com menor ruído, um valor maior do hiperparâmetro pode ser adotado. Considere a série temporal composta por $n = 512$ valores do Índice da Bolsa de Valores de São Paulo, o IBOVESPA, coletados diariamente entre 10 de Janeiro de 2019 e 29 de Janeiro de 2021. A série foi obtida em <https://www.infomoney.com.br/cotacoes/ibovespa/grafico/> e está representada no Gráfico 14.



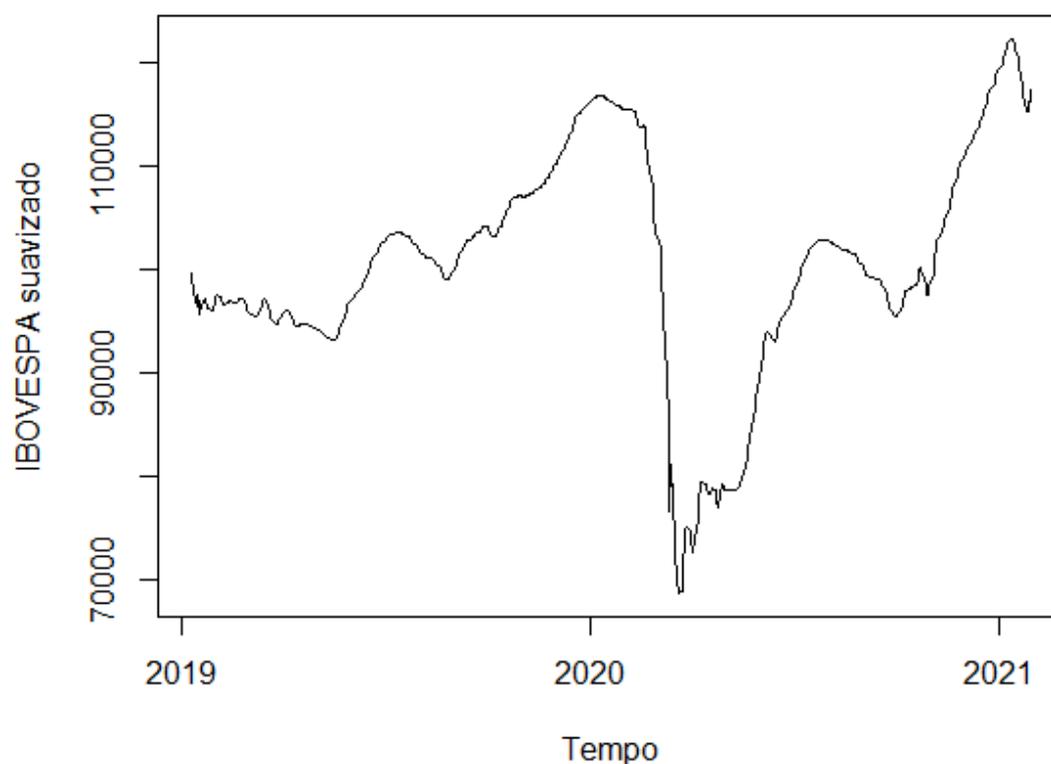
Gráfico 14: Série com os Índices da Bolsa de Valores de São Paulo entre os dias 10/01/2019 e 29/01/2021.



Observe que a série apresenta uma queda visível no início de 2020, porém em outros períodos há pequenas flutuações passíveis de serem ruídos aleatórios ou, de fato, alguma pequena mudança estrutural da série, o que seria importante de ser detectada. Desta forma, a aplicação da regra de encolhimento proposta é interessante para remoção do ruído da série para detecção de mudanças sistemáticas da mesma. O Gráfico 15 apresenta a série suavizada pela regra de encolhimento sob priori Champernowne com σ e τ determinados de acordo com (9) e (10), $\alpha = 5$ e $\lambda = 1$. Neste caso, consideramos um valor maior de α em relação ao utilizado nos dados de potenciais de ação neuronais, uma vez que a presença do ruído na série é menor, apesar de existente. Se utilizássemos um valor menor para tal hiperparâmetro, o grau de encolhimento da regra associada seria maior e os dados seriam suavizados em excesso, o que pode acarretar na falta de detecção de características estruturais relevantes da série. Para mais detalhes sobre ondaletas em séries temporais, ver MORETTIN (2014).



Gráfico 15: Série com os Índices da Bolsa de Valores de São Paulo entre os dias 10/01/2019 e 29/01/2021 suavizada pela regra de encolhimento sob priori Champernowne.



6. CONSIDERAÇÕES FINAIS

Propomos neste trabalho a utilização da distribuição a priori formada pela mistura da função delta de Dirac com a distribuição de Champernowne para os coeficientes de ondaletas em um problema de regressão não paramétrica sob abordagem bayesiana.

A distribuição de Champernowne é atraente no contexto das ondaletas por ser simétrica (em torno de zero neste trabalho), por generalizar algumas distribuições conhecidas como a logística e de Burr e por possuir hiperparâmetros que são de fácil interpretação em termos do grau de encolhimento a ser atribuído à regra bayesiana de encolhimento associada a priori.

Nosso estudo de simulações, que levou em consideração como funções subjacentes as funções testes de Donoho e Johnstone, mostrou excelente performance em termos de erro quadrático médio da regra de encolhimento sob priori de Champernowne em relação às regras de encolhimento ou de



limiar selecionadas para comparação e que são amplamente aplicadas na área. De fato, a regra proposta superou as regras de comparação na maioria dos cenários determinados nas simulações, tanto em relação às funções subjacentes como em relação aos diferentes valores da razão sinal-ruído selecionados, o que mostra flexibilidade da regra para aplicações em dados com variadas características.

Por fim, as aplicações em dados reais da regra de encolhimento associada a priori de Champernowne foram bem-sucedidas, isto é, tanto para o conjunto de dados de potenciais de ação neuronais como para os dados do IBOVESPA, a regra minimizou satisfatoriamente os ruídos aleatórios presentes nos dados, de modo a tornar mais evidente as principais características a serem detectadas, como pontos de máximos e mínimos locais, picos, quedas abruptas e descontinuidades.

Portanto, concluímos de acordo com os resultados obtidos neste trabalho que a utilização da distribuição de Champernowne como constituinte da distribuição à priori para coeficientes de ondaletas pode ser considerada como uma alternativa para aplicação em problemas de regressão não paramétrica por ondaletas sob abordagem bayesiana. Estudos sobre os impactos das escolhas da base de ondaletas e da distribuição probabilística do erro aleatório no modelo original no desempenho da regra de encolhimento são sugeridos como estudos futuros e de interesse para análise.

AGRADECIMENTOS

Esta pesquisa foi realizada sob financiamento da CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior).

REFERÊNCIAS

- ABRAMOVICH, F.; BENJAMINI, Y. (1996) Adaptive thresholding of wavelet coefficients. **Comput Stat Data Anal**, v. 22, p. 351–361, 1996.
- ABRAMOVICH, F.; SAPATINAS, T.; SILVERMAN, B. W. Wavelet thresholding via a Bayesian approach. **J R Stat Soc Ser B**, v. 60, p. 725–749, 1998.
- ANGELINI, C.; VIDAKOVIC, B. (2004) Gama-minimax wavelet shrinkage: a robust incorporation of information about energy of a signal in denoising applications. **Stat Sin**, v. 14, p. 103–125, 2004.
- BHATTACHARYA, A.; PATI, D.; PILLAI, N. *et al.* Dirichlet–Laplace priors for optimal shrinkage. **J Am Stat Soc**, v. 110, p. 1479–1490, 2015.
- BERGER, J.; JEFFERYSW, MULLER P. Bayesian nonparametric shrinkage applied to cepheid star oscillations. **Stat Sci**, v. 27, n. 1, p. 3–10, 2012.
- CHAMPERNOWNE, D. G. The Graduation of Income Distributions. **Econometrica**, v. 20, p. 591-615, 1952.



DONOHO, D. L. Nonlinear wavelet methods of recovery for signals, densities, and spectra from indirect and noisy data. *In.: Proceedings of Symposia in Applied Mathematics*. American Mathematical Society: Providence: RI, 1993a v. 47.

DONOHO, D. L. Unconditional bases are optimal bases for data compression and statistical estimation. **Appl Comput Harmonic Anal**, v. 1, p. 100–115, 1993b.

DONOHO, D. L. De-noising by soft-thresholding. **IEEE Trans Inf Theory**, v. 41, p. 613–627, 1995a
DONOHO, D. L. Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition.

Appl Comput Harmonic Anal, v. 2, p. 101–26, 1995b.

DONOHO, D. L.; JOHNSTONE, I. M. Ideal denoising in an orthonormal basis chosen from a library of bases. **Compt Rend Acad Sci Paris A**, v. 319, p. 1317–1322, 1994a.

DONOHO, D. L.; JOHNSTONE, I. M. Ideal spatial adaptation by wavelet shrinkage. **Biometrika**, v. 81, p. 425–455, 1994b.

DONOHO, D. L.; JOHNSTONE, I. M. Adapting to unknown smoothness via wavelet shrinkage. **J Am Stat Assoc**, v. 90, p. 1200–1224, 1995.

EINEVOLL, G. T.; FRANKE, F.; HAGEN, E. *et al.* Towards reliable spike train recordings from thousands of neurons with multielectrodes. **Curr Opin Neurobiol**, v. 22, p. 7–11, 2012.

GRIFFIN, J.; BROWN, P. Hierarchical shrinkage priors for regression models. **Bayesian Anal**, v. 12, n. 1, p. 135–159, 2017.

JOHNSTONE, L. M. SILVERMAN, B. W. Empirical Bayes selection of wavelet thresholds. **Ann Stat**, v. 33, p. 1700–1752, 2005.

KARAGIANNIS, G.; KONOMI, B.; LIN, G. A Bayesian mixed shrinkage prior procedure for spatial-stochastic basis selection and evaluation of gPC expansion: applications to elliptic SPDEs. **J Comput Phys**, v. 284, p. 528–546, 2015.

LEWICKI, M. S. A review of methods for spike sorting: the detection and classification of neural action potentials. **Comput Neur Syst**, v. 9, p. R53–R78, 1998.

LIAN, H. On posterior distribution of Bayesian wavelet thresholding. **J Stat Plan Inference**, v. 141, p. 318–324, 2011.

MALLAT, S. G. **A Wavelet Tour of Signal Processing**. San Diego: Academic Press, 1998.

MORETTIN, P. A. **Waves and Wavelets: From Fourier to Wavelet Analysis of Time Series**. São Paulo: University of São Paulo Press, 2014.

NASON, G. P. Wavelet shrinkage using cross-validation. **J R Stat Soc B**, v. 58, p. 463–479, 1996.

POUZAT, C.; MAZORO, Laurentg. Using noise signature to optimize spike-sorting and to assess neuronal classification quality. **J Neurosci Meth**, v. 122, p. 43–57, 2002.

QUIROGA, R. Q.; NADASDY, Z.; BEN-SHAUL, Y. Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering. **J Neural Comput**, v. 16, p. 1661–87, 2004.

REMÉNYI, N.; VIDA KOVIC, B. Wavelet shrinkage with double weibull prior. **Commun Stat Simul Comput**, v. 44, n. 1, p. 88–104, 2015.



RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR

ENCOLHIMENTO BAYESIANO DE COEFICIENTES DE ONDALETAS SOB
PRIORI DE CHAMPERNOWNE COM APLICAÇÕES
Alex Rodrigo dos Santos Sousa

ROBERT, C. P. **The Bayesian choice**: from decision-theoretic foundations to computational implementation. New York: Springer, 2001.

ROBERT, C. P.; CASELLA, G. MONTE. **Carlo Statistical Methods**. New York: Springer-Verlag, 2005.

SHALCHYAN, V.; JENSEN, W.; FARINA, D. Spike detection and clustering with unsupervised wavelet optimization in extracellular neural recordings. **IEEE Trans Biomed Eng**, v. 59, p. 2576–85, 2012.

SHOHAM, S.; FELLOWS, M. R.; NORMANN, R. A. Robust, automatic spike sorting using mixtures of multivariate t-distributions. **J Neurosci Methods**, v. 127, p., 111–22, 2003.

TORKAMANI, R.; SADEGHZADEH, R. Bayesian compressive sensing using wavelet based Markov random fields. **Signal Process Image Commun**, v. 58, p. 65–72, 2017.

VIDAKOVIC, B. **Statistical modeling by wavelets**. New York: Wiley, 1999.

VIDAKOVIC, B.; RUGGERI, F. BAMSmethod: theory and simulations. **Sankhya Indian J Stat B**, v.63, p. 234–249, 2001.