Estimação por ondaletas em análise de dados funcionais agregados

Ana Luiza Ferreira Viana, SP - UNICAMP
Prof. Dr. Alex Rodrigo dos Santos Sousa (orientador), SP - UNICAMP

Palavras-Chave: análise de dados funcionais, estimação de curvas, wavelet shrinkage.

Introdução

Em diversas áreas da ciência é possível lidar com dados funcionais agregados por curvas componentes desconhecidas, como, por exemplo, na composição de energia elétrica de uma determinada região, em que a curva média de consumo de energia em um dado período de tempo é composta pelas curvas de consumo de cada consumidor. Vários métodos estatísticos foram propostos para estimar curvar componentes a partir de curvas agregadas, muitos considerando a estrutura dos dados como multivariada, devido ao fato de que as observações são obtidas em tempos discretos. Embora aplicado com sucesso em muitos casos, estes métodos não consideram a estrutura funcional dos dados. Assim, foi proposta a utilização de bases de ondaletas para estimar curvas médias individuais das funções componentes a partir de curvas agregadas por tais funções em modelos com erros aleatórios normais.

Metodologia

Modelo estatístico

O problema consiste em estimar um conjunto de L funções componentes desconhecidas, $\alpha_l(t)$, a partir de observações de uma função agregada, $f(t) \in L^2(\mathbb{R})$, descrita como:

$$f(t) = \sum_{l=1}^{L} y_l \alpha_l(t) + e(t) \tag{1}$$

onde y_l são pesos conhecidos e $e(t)$ é um processo gaussiano com média zero e variância σ^2.

A estratégia é representar cada função componente $\alpha_l(t)$ em uma base de ondaletas

ortonormal $\{\psi_{jk}(t)\}$:

$$\alpha_l(t) = \sum_{j,k\in\mathbb{Z}} \gamma_{jk}^{(l)} \psi_{jk}(t) \tag{2}$$

Desta forma, o problema em se estimar as funções α_l torna-se um problema de estimar um número finito de coeficientes de ondaletas $\gamma_{jk}^{(l)}$'s da representação (2). Na prática, a partir de I amostras da curva agregada, observadas em $M = 2^J$ localizações, obtém-se a forma matricial $f = \alpha y + e$.

Tem-se por objetivo estimar os coeficientes de ondaleta e, para isso, é aplicada a transformada discreta de ondaletas.

Dados funcionais

Deseja-se estimar curvas componentes a partir de dados funcionais agregados, que são compostos por uma combinação de curvas desconhecidas. Então tem-se o objetivo de estimar estas curvas componentes. Uma função $\psi \in L_2(\mathbb{R}) = \left\{f : \mathbb{R} \to \mathbb{R} \mid \int f^2 < \infty\right\}$ é uma ondaleta se satisfizer a propriedade $\int \psi(x)dx = 0$.

Pode-se escrever que $f(x) = \sum_{j\in\mathbb{Z}} \sum_{k\in\mathbb{Z}} \theta_{jk}\psi_{jk}(x)$, ou seja, f pode ser representada em bases de ondaletas.

Transformada Discreta de Ondaletas (DWT)

Com o objetivo de estimar os coeficientes aplica-se a transformada discreta de ondaletas que pode ser representada por uma matriz de transformação de ondaletas $\mathbf{W}$. Aplicando $\mathbf{W}$ em ambos os lados de (1) entra-se no domínio das ondaletas e obtém-se:

$$d^* = d + \varepsilon \quad \text{com} \quad d^* = \mathbf{W}\alpha\mathbf{y}, \quad d = \mathbf{W}\mathbf{f}, \quad \varepsilon = \mathbf{W}\mathbf{e}.$$

A transformada discreta de ondaletas mantém as propriedades do erro e_i, logo, ε_i são variáveis aleatórias com distribuição normal de média 0 e variância σ^2 assim como os erros da equação (1). Chama-se $\mathbf{W}\alpha = \mathbf{\Gamma}$ a matriz dos coeficientes desconhecidos.

Regras de limiar e encolhimento

Após a DWT, o problema de estimar f foi substituído por estimar os coeficientes de ondaletas d utilizando os coeficientes de ondaletas empíricos d^*. Como os coeficientes de ondaleta têm um comportamento esparso (com muitos zeros), aplicam-se regras de encolhimento (shrinkage) para zerar coeficientes dominados por ruído, como forma de reduzir o erro.

Existem vários métodos para definir qual o limite para o qual as observações são encolhidas e a partir de que ponto é considerado que não é mais apenas o ruído agindo e o coeficiente de ondaleta não é nulo. Este limite é chamado de limiar (threshold). Também existem vários métodos para definir qual o valor do ruído para os coeficientes observados que não são nulos, para mais detalhes ver Nason (2008).

Para se definir qual o valor do ruído para o qual os coeficientes observados são não nulos é necessário determinar uma regra para o limiar. As principais regras são o limiar duro (hard), o suave (soft) e a Bayesiana.

Para o limiar duro, valores menores que λ são levados para zero e os maiores que λ são mantidos sem alterações. Já para o limiar suave os valores menores que λ são levados para zero e, para os maiores, é calculado o valor do coeficiente subtraindo-se λ.

Regra de encolhimento Bayesiana

Neste método, assume-se que a função $\mathbf{f}$ segue uma distribuição a priori tal que $\pi(\theta; \alpha, \tau) = \alpha\delta_0(\theta) + (1-\alpha)g(\theta; \tau)$, em que α é um valor no intervalo $[0, 1]$, δ é a função Delta de Dirac, g é uma função com distribuição simétrica e unimodal em zero de forma que $g(\theta; \tau) = \frac{\exp(-\theta/\tau)}{\tau(1+\exp(-\theta/\tau))^2}\mathbb{I}_{\mathbb{R}}(\theta)$, com $\alpha \in (0, 1)$, $\tau > 0$ e $\mathbb{I}_{\mathbb{R}}(\cdot)$ a função indicadora em $\mathbb{R}$.

A regra de encolhimento para estimação bayesiana do coeficiente de ondaletas θ pode ser obtida por meio da definição de uma função perda e sua respectiva regra de Bayes. Sabe-se que, sob a seguinte função perda quadrática, $L(\delta, \theta) = (\delta - \theta)^2$, a regra de Bayes que minimiza esta função perda é dada por $\delta(d) = E_\pi(\theta \mid d)$. Ou seja, sob função perda quadrática, a regra de Bayes é dada pelo valor esperado a posteriori de θ.

Assim, encontram-se os coeficientes de ondaletas estimados para cada coeficiente ruidoso θ. Após estimar os valores reais para os dados, aplica-se a transformada discreta inversa de ondaletas (IDWT) para se voltar ao domínio do tempo e se encontrar a função $\mathbf{f}$ estimada sem os ruídos, a partir dos valores estimados.

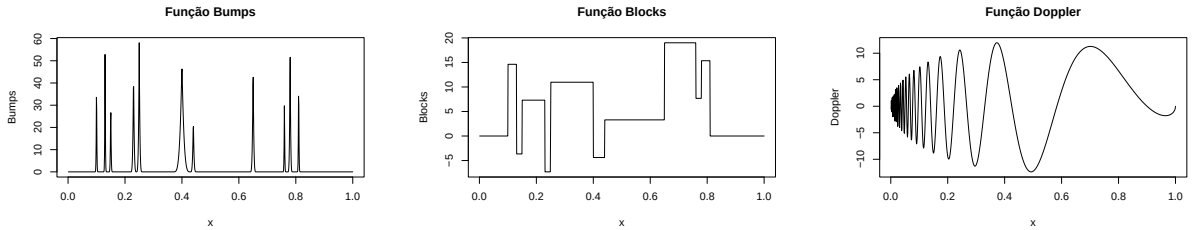
Pela transformada discreta de ondaletas α pode ser estimado utilizando-se $\hat{\alpha} = \mathbf{W^t}\hat{\mathbf{\Gamma}}$. Assim, volta-se para o domínio do tempo e é possível encontrar a função estimada para os dados, de forma que $\hat{\mathbf{f}} = \mathbf{W^t}\hat{\theta}$.

Resultados - aplicação em dados reais e simulação

Estudo de simulação

Para essa parte do projeto, foi conduzido um estudo de simulação em prol de avaliar o desempenho do método de estimação Bayesiano em cenários controlados. O objetivo foi analisar a performance do estimador sob diferentes condições, variando-se a complexidade do sinal, o tamanho da amostra e o nível de ruído presente nos dados.

Para as funções teste foram utilizadas três curvas clássicas do conjunto de Donoho-Johnstone, sendo estas: Bumps, Blocks e Doppler, representadas pelas figuras 1a, 1b e 1c, respectivamente. Todas elas tem picos, descontinuidades e oscilações. Já para o nível de ruído, para cada cenário da simulação foram testados dois níveis de ruído, definidos como SNR = 3 para ruído alto e SNR = 9 para ruído baixo. E, por fim, para o tamanho da amostra, a simulação tinha três cenários: n = 128, n = 1024 e n = 2048 pontos.



(a) Função Bumps (b) Função Blocks (c) Função Doppler

Figura 1: Funções teste de Donoho–Johnstone utilizadas na simulação.

Assim, para cada uma das 18 combinações de cenários (3 curvas x 2 níveis de ruído

x 3 tamanhos de amostra) foram realizadas 200 replicações. Para quantificar a acurácia desta reconstrução, a performance do estimador em cada replicação foi medida pelo Erro Quadrático Médio (MSE). A métrica final de desempenho para cada cenário é a média do Erro Quadrático Médio (AMSE), obtido pela média dos MSEs das 200 replicações.

A Tabela 1 mostra a média dos erros quadráticos médios de cada uma dos 18 cenários. A partir da tabela, pode-se observar a grande relevância do nível de ruído para a qualidade do ajuste dos dados: para ruídos maiores (SNR = 3), em todos os diferentes cenários, o erro quadrático médio aumenta. Além disso, o tamanho da amostra influencia na qualidade do ajuste da curva aos dados, sendo que para amostras maiores têm-se erros menores.

Tabela 1: AMSE Bayesiano para todas as diferentes funções, tamanhos de amostra e níveis de ruído

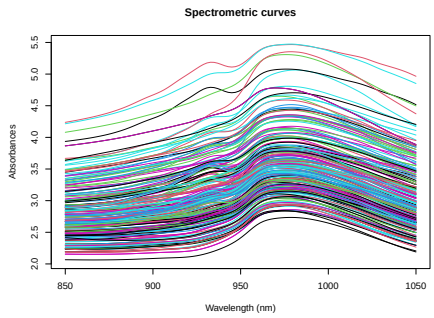
n	**SNR**	**Bumps**	**Blocks**	**Doppler**
128	3	18.9909	17.6669	10.7682
	9	0.6455	0.7087	0.3937
1024	3	6.8195	4.9887	1.7004
	9	0.2708	0.2184	0.0918
2048	3	4.1896	3.0589	1.0088
	9	0.1658	0.1426	0.0521

A partir do estudo de simulação, foi possível perceber o grande poder do estimador que, mesmo para funções com picos, descontinuidades e oscilações, conseguiu estimar bem as curvas, até em cenários desfavoráveis.

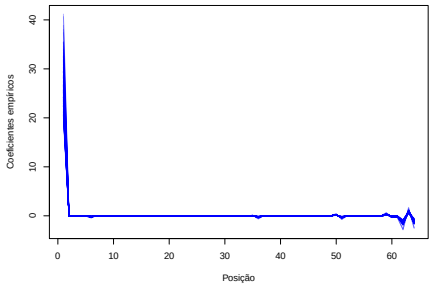
Aplicação em dados reais

Para aplicação da metodologia de ondaletas em um conjunto de dados reais utilizou-se o conjunto disponível no pacote Tecator. Este conjunto de dados é composto por espectros de absorbância de I = 215 amostras de carne, medidas em M = 100 pontos igualmente espaçados, medindo entre 850 a 1050 nms.

Cada amostra de carne pode ser decomposta em componentes de água, proteína e gordura, com diferentes concentrações em cada uma dos três. A Figura 2a ilustra a distribuição dos dados de absorbância utilizados. Os coeficientes empíricos, obtidos através da aplicação da DWT, podem ser observados na imagem 2b.



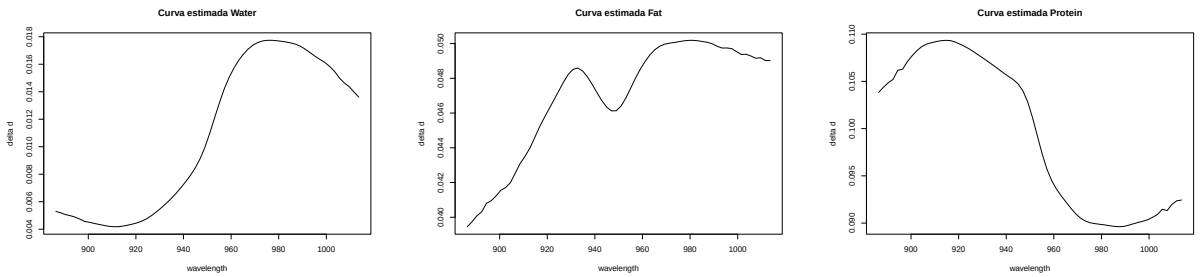
(a) Distribuição dos dados



(b) Distribuição dos coeficientes empíricos

Figura 2: Visualização dos dados do conjunto Tecator e seus coeficientes empíricos.

A curva de absorbância de cada amostra pode ser modelada como uma combinação linear das curvas componentes $\alpha_1(t), \alpha_2(t)$ e $\alpha_3(t)$, associadas à água, gordura e proteína, respectivamente. Assim, foi aplicado a função logística com hiperparamêtros $\tau = 5$, p = 0.8, 50 simulações e o valor de σ conforme descrito acima. As curvas estimadas para água, proteína e gordura, respectivamente, encontram-se nas imagens 3a, 3b, e 3c.



(a) Curva estimada Água (b) Curva estimada Proteína (c) Curva estimada Gordura

Figura 3: Curvas estimadas para os três componentes: Água, Proteína e Gordura.

As estimativas obtidas a partir das imagens acima permitem verificar que as curvas componentes de fato apresentam comportamentos distintos, refletindo as características únicas de cada uma das três substâncias. Isso evidencia que a metodologia de ondaletas, juntamente ao encolhimento bayesiano, foi eficaz em separar e estimar os efeitos individuais dos componentes a partir da curva de absorbância agregada.

Considerações finais

Este trabalho cumpriu com sucesso seus objetivos: primeiramente, aplicar a metodologia de estimação em bases de ondaletas a um conjunto de dados reais para estimar curvas componentes, em segundo lugar, avaliar o desempenho dessa abordagem por meio de um estudo de simulação e, em terceiro, um estudo completo sobre a estimação de dados funcionais agregados por ondaletas. A aplicação prática demonstrou a capacidade do método em recuperar curvas componentes a partir de dados agregados.

Referências bibliográficas

FEBRERO-BANDE, Manuel; OVIEDO DE LA FUENTE, Manuel (2012). Statistical Computing in Functional Data Analysis: The R Package fda.usc. *Journal of Statistical Software*, 51(4), 1-28. URL: `https://www.jstatsoft.org/v51/i04/`.

NASON, G.P. (2008). *Wavelet Methods in Statistics with R*. Springer.

NASON, G. (2022). wavethresh: Wavelets Statistics and Transforms. R package version 4.7.2, `https://CRAN.R-project.org/package=wavethresh`.

SARAIVA, Marley Apolinario (2009). *Análise Não-Paramétrica de Dados Funcionais: Uma Aplicação à Quimioterapia*.

SOUSA, Alex Rodrigo dos Santos. Encolhimento bayesiano sob priori beta de coeficientes de ondaletas em modelos com erros gaussianos e positivos. 2018. 1 recurso online (90 p.) Tese (doutorado) - Universidade Estadual de Campinas, Instituto de Matemática Estatística e Computação Científica, Campinas, SP.

SOUSA, Alex Rodrigo dos Santos (2024). A wavelet-based method in aggregated functional data analysis. *Monte Carlo Methods and Applications*, vol. 30, no. 1, pp. 19-30.