



Aplicações de Análise de Dados Funcionais em Ciência Forense

Eric Pavarim Lima, IMECC - UNICAMP

Prof. Dr. Alex Rodrigo dos Santos Sousa (orientador), IMECC - UNICAMP

Palavras-chave: Dados Funcionais, Splines, Espectroscopia

1 INTRODUÇÃO

O presente trabalho tem como objetivo o estudo e aplicação de métodos estatísticos, em especial de regressão linear não-paramétrica, no contexto de dados funcionais em ciência forense. Para isso, conceitos de quimiometria, dados longitudinais e simulação computacional foram extremamente necessários para alcançar o objetivo proposto.

Os temas abordados foram, cronologicamente, estudo e revisão da teoria essencial, aplicação em um conjunto de dados forense (Tecator), simulação e criação de dashboard para contemplar diferentes cenários e, por fim, estudo e aplicação da razão de verossimilhança em ambiente simulado e em dados reais.

2 METODOLOGIA

A análise de dados funcionais⁵ trata as observações com uma abordagem diferente da tradicional, no sentido de, os dados possuem uma estrutura de dependência intrínseca, como, por exemplo, temporal ou espacial. Vale ressaltar que, por muito tempo, utilizou-se métodos multivariados para lidar com essa repetição de informações em cada observação, entretanto, problemas conceituais e numéricos surgiram com essa abordagem, o que resultou na necessidade de utilizar métodos mais flexíveis e que incorporem a dependência como modelos lineares generalizados aditivos ou, o tópico principal aqui, modelos não paramétricos, como expansão por splines ou wavelets.

Uma das principais métricas funcionais, no âmbito de ciência forense, é a absorvância e pode ser definida como ¹:

$$x = -\log\left(\frac{I_0}{I}\right)$$

Em que $x > 0$ e, I é a intensidade da luz restante em dado comprimento de onda e I_0 é a intensidade de luz incidente em dado comprimento de onda, em outras palavras, é basicamente uma razão da luz antes de entrar e depois de sair, resultando na quantidade absorvida. Trata-se

de uma medida funcional pois diferentes comprimentos de onda, por exemplo, no espectro de luz visível, resultam em diferentes medidas de absorbância que dependem de tal frequência

Uma das principais leis da espectroscopia é a lei de Beer-Lambert, definida por:

$$x_j = \sum_{\ell=1}^m y_{\ell} \alpha_{\ell j} + \varepsilon_j$$

em que x_j é o valor observado do constituinte químico, y_{ℓ} é a concentração do constituinte ℓ ; $\alpha_{\ell j}$ representa a absorbância ou termitância do constituinte; ε_j é o erro aleatório e m é o número de pontos do espectro.

A interpretação da lei é simples: a absorbância total de uma substância, em dado comprimento de onda, é uma combinação linear proporcional das absorbâncias dos constituintes da substância. Partindo para a modelagem estatística, pode-se estimar a curva agregada da absorbância com métodos não-paramétricos e expansão por splines. A expansão por splines é escolhida pois a absorbância tem comportamento estritamente positivo e sem ciclos inerentes a sua natureza, o que favorece o uso de polinômios, e o modelo é definido por:

$$x_i(t) = \sum_{l=1}^L \left[\theta_{0l} + \sum_{j=1}^m \theta_{jl} y_{ij} \right] B_l(t) + \epsilon_i(t)$$

Em que cada θ é um coeficiente das bases de Splines⁴ devido a expansão não paramétrica, com seus devido número de knots. Além disso, uma forma interessante de descrever esse modelo é usando a notação matricial¹, pois a solução e, portanto, a estimação pode ser encontrada utilizando mínimos quadrados ordinários. A notação matricial é dada por:

$$\underbrace{\begin{bmatrix} x_1(t_1) \\ x_2(t_1) \\ \vdots \\ x_n(t_k) \end{bmatrix}}_{\mathbf{X}(t)} = \underbrace{\begin{bmatrix} 1 & y_{11} & \cdots & y_{1m} & B_1(t_1) & y_{11}B_1(t_1) & \cdots & y_{1m}B_L(t_1) \\ 1 & y_{21} & \cdots & y_{2m} & B_1(t_1) & y_{21}B_1(t_1) & \cdots & y_{2m}B_L(t_1) \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 1 & y_{n1} & \cdots & y_{nm} & B_1(t_k) & y_{n1}B_1(t_k) & \cdots & y_{nm}B_L(t_k) \end{bmatrix}}_{\mathbf{D}(t)} \cdot \underbrace{\begin{bmatrix} \theta_{01} \\ \theta_{11} \\ \vdots \\ \theta_{mL} \end{bmatrix}}_{\boldsymbol{\Theta}} + \underbrace{\begin{bmatrix} \varepsilon_1(t_1) \\ \varepsilon_2(t_1) \\ \vdots \\ \varepsilon_n(t_k) \end{bmatrix}}_{\boldsymbol{\varepsilon}(t)}$$

E podemos achar a solução com:

$$\hat{\boldsymbol{\Theta}} = (\mathbf{D}(t)^\top \mathbf{D}(t))^{-1} \mathbf{D}(t)^\top \mathbf{X}(t)$$

e a recuperação das curvas com:

$$\hat{\boldsymbol{\alpha}} = \mathbf{B} \hat{\boldsymbol{\Theta}}$$

Por fim, a razão de verossimilhança para dados funcionais³ foi estudada e tem como objetivo comparar duas hipóteses concorrentes: H_p , de que ambas curvas provêm da mesma origem, e H_d , de que provêm de origens distintas. A razão de verossimilhança (RV), em razão da evidência coletada, ou seja, os dados, é definida como:

$$RV = \frac{\Pr(E \mid H_p)}{\Pr(E \mid H_d)}$$

Após reduzir as curvas com bases spline ($Z = W\boldsymbol{\Theta}$) e padronizar, em escala 0 a 1, a grandeza, como comprimento de onda, obtém-se os coeficientes reduzidos Z , sobre os quais as estatísticas são calculadas. Sejam $\bar{z}_r$ e $\bar{z}_c$ as médias dos grupos recuperado e controle, respectivamente, e $\bar{z} = (\bar{z}_r + \bar{z}_c)/2$ a média global. Além disso, U é matriz de covariância dentro de cada grupo

(estimada), C a matriz de covariância entre grupos e y^* a média ponderada entre os grupos, reduzida. As variâncias utilizadas são:

$$\Sigma_{Hp} = \frac{U}{n_r} + \frac{U}{n_c}, \quad \Sigma_p^* = \frac{U}{n_r + n_c} + C, \quad \Sigma_r = \frac{U}{n_r} + C, \quad \Sigma_c = \frac{U}{n_c} + C$$

Com isso, a razão de verossimilhança funcional assume a forma:

$$RV = \frac{\phi(\bar{z}_r - \bar{z}_c; 0, \Sigma_{Hp}) \cdot \phi(y^* - \bar{z}; 0, \Sigma_p^*)}{\phi(\bar{z}_r - \bar{z}; 0, \Sigma_r) \cdot \phi(\bar{z}_c - \bar{z}; 0, \Sigma_c)}$$

em que $\phi(\cdot; \mu, \Sigma)$ denota a densidade de uma normal multivariada com média μ e matriz de covariância Σ . Valores de RV maiores que 1 indicam maior verossimilhança sob H_p , enquanto valores menores que 1 favorecem H_d .

3 RESULTADOS E DISCUSSÕES:

O primeiro cenário estudado foi a aplicação do método de recuperação das curvas constituintes em um conjunto de dados de amostras de carne chamado Tecator em que o objetivo foi a recuperação das curvas base de Gordura, Proteína e Água utilizando apenas as absorbâncias agregadas das carnes e as respectivas concentrações dos constituintes.

Vale lembrar que, como não há o acesso prévio das curvas originais, não pode-se inferir quão assertiva foi a recuperação, portanto, o segundo cenário foi de simulação com curvas conhecidas. No entanto, o valor da recuperação, mesmo com essa incerteza, dá-se pela aplicação de dados reais.

Abaixo, na Figura 1, segue as curvas de absorbância totais do banco de dados, enquanto na Figura 2, as recuperações das curvas constituintes, ou seja, a calibração do modelo.

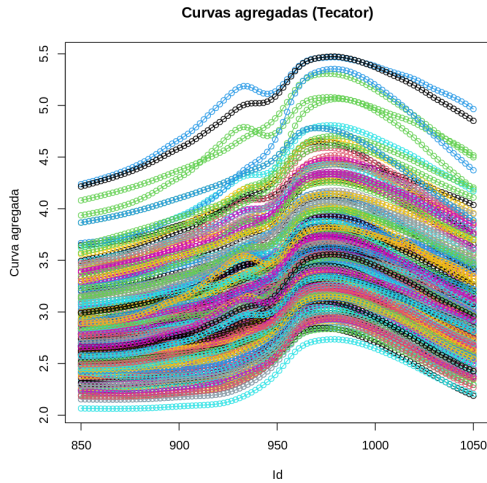


Figura 1: Curvas de absorbância de carnes do banco Tecator

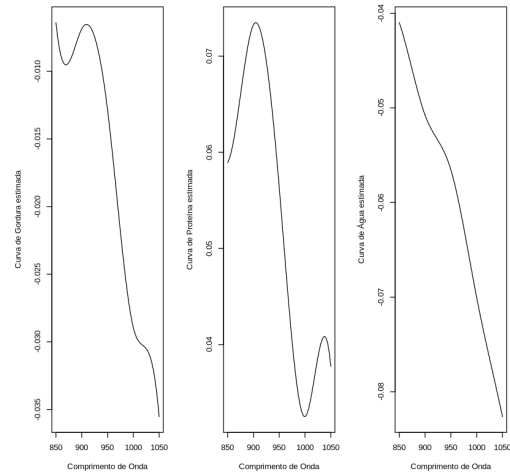


Figura 2: Recuperação das curvas de Gordura, Proteína e Água, respectivamente

Para o segundo cenário, foi desenvolvido um Dashboard interativo em Shiny, no R, para controle geral de hiperparâmetros e métricas como ajuste no ruído dos dados, número de knots, quantidade amostral etc. Esta implementação foi bem sucedida pois possibilitou uma gama maior de cenários interessantes e a identificação de como essas configurações estatísticas se relacionam com o êxito da estimação.

Vale ressaltar que ambos possuem variância, quantidade amostral e grau polinomial controlados e diferem apenas na quantidade de knots. Essa simulação foi importante para mostrar o poder da recuperação de curvas desconhecidas, com o método proposto, e como alguns ajustes alteram totalmente tal recuperação. As funções utilizadas e agregadas, e posteriormente recuperadas, chamam-se Logito e Spahet ⁶. Abaixo nas Figuras 3 e 4, encontram-se dois cenários, de ajuste insuficiente e excedente na recuperação.

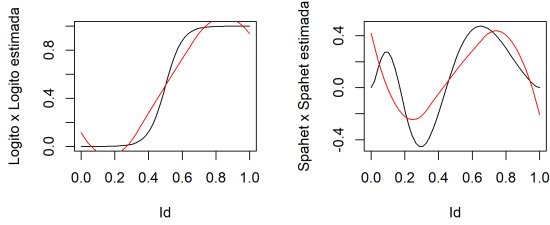


Figura 3: Cenário com 4 knots, demonstrando sub-ajuste e recuperação falha

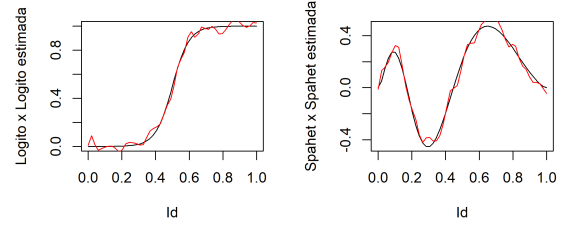


Figura 4: Cenário com 30 knots, demonstrando superajuste e modelagem do ruído

Partindo para o estudo de razão de verossimilhança funcional, foi simulado duas curvas, evidentemente, distintas para testa-las e verificar se a hipótese alternativa, de distinção da origem é satisfeita. O resultado foi bem sucedido com $RV = 3.94477268918904e - 312$ o que indicamente um denominador muito maior que o numerador. Abaixo, nas Figuras 5 e 6, há as curvas utilizadas e sua reduções com splines ao lado.

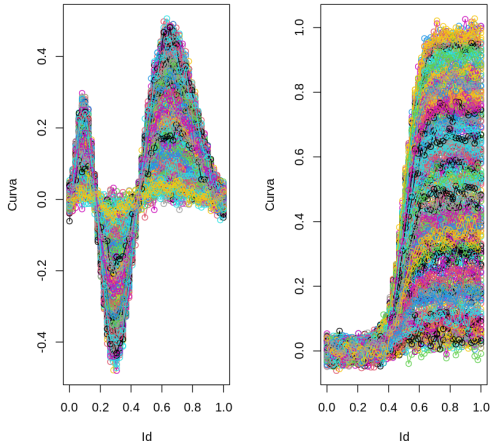


Figura 5: Curvas simuladas para a razão de verossimilhança

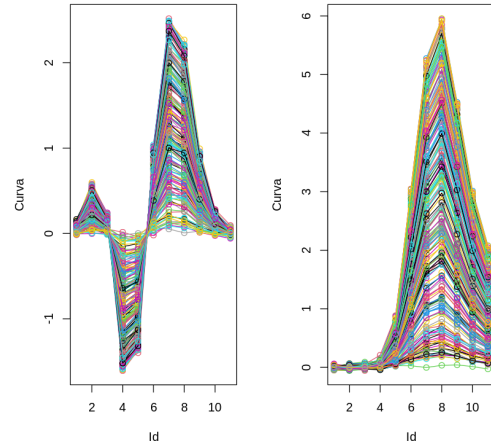


Figura 6: Curvas simuladas reduzidas para a razão de verossimilhança

Por fim, o banco de dados Milkns⁷ foi utilizado. Esse banco de dados contém amostras de leite ao decorrer de várias semanas. A variável utilizada foi a curva total (chamada Tot) na semana 1. Abaixo na Figura 7, encontra-se as curvas gerais da Semana 1 e na Figuras 8, as curvas utilizadas para teste (idênticas) reduzidas. Como esperado, o valor da razão de verossimilhança foi exatamente 1, comprovando a mesma origem das curvas.

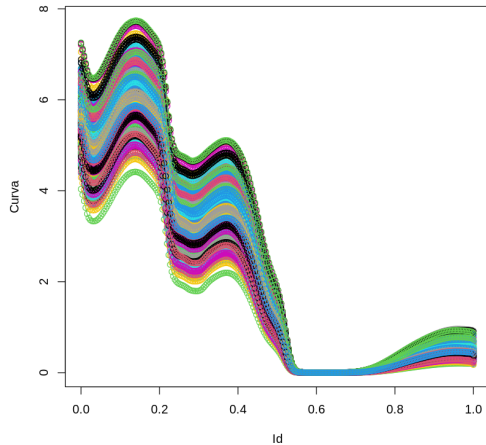


Figura 7: Curvas gerais da Semana 1 da variável Tot do banco de dados Milkness

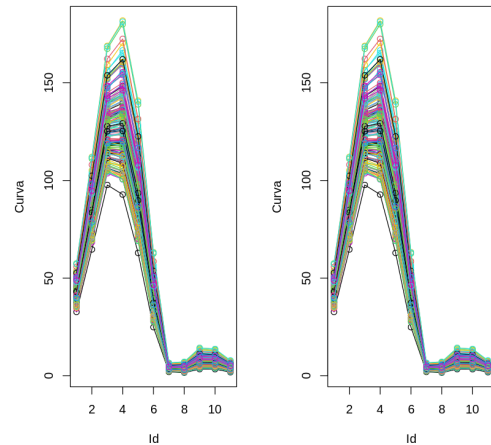


Figura 8: Curvas, por grupo, testadas e reduzidas para razão de verossimilhança

4 CONCLUSÕES

Em suma, a aplicação estatística para calibração de modelos em ciência forense se mostra eficaz, principalmente pela natureza comum de estrutura temporal ou dependente de observações como em problemas de absorbância, termitância etc. No entanto, vale ressaltar que trata-se de uma intersecção recente e, portanto, com um potencial latente enorme e este trabalho serviu para mostrar uma pequena demonstração da força conjunta dessas áreas da ciência.

5 BIBLIOGRAFIA

- [1] Saraiva, Marley Apolinario, and Ronaldo Dias. Análise não-paramétrica de dados funcionais: uma aplicação a quimiometria. Diss. Master's thesis, Universidade Estadual de Campinas, Campinas, 2009.
- [2] Ramsay, James O., and Bernard W. Silverman. Functional data analysis. New York, NY: Springer New York, 2005.
- [3] Martyna, Agnieszka, et al. "The evidential value of microspectrophotometry measurements made for pen inks." *Analytical Methods* 5.23 (2013): 6788-6795.
- [4] James, Gareth, et al. An introduction to statistical learning: with applications in R. Vol. 103. New York: springer, 2013.
- [5] Ferraty, Frédéric, and Philippe Vieu. Nonparametric functional data analysis: theory and practice. New York, NY: Springer New York, 2006.
- [6] Sousa, Alex Rodrigo dos Santos. "Simulation studies to compare bayesian wavelet shrinkage methods in aggregated functional data." *arXiv preprint arXiv:2210.04966* (2022).
- [7] Diaz-Olivares, Jose A., et al. "Near-infrared spectra dataset of milk composition in transmittance mode." *Data in Brief* 51 (2023): 109767.