ISSN: 2317-0840

Estudo de simulação na análise de dados funcionais: Spline×Fourier

Matheus C. Silva[1], Gabriel Edson S. Silva[2], Ernandes G. Moura[3†], Luíz Leonardo D. Garcia[4]

[1] *IFMA-Instituto Federal do Maranhão. E-mail:* matheusifsjp@gmail.com.
[2] *IFMA-Instituto Federal do Maranhão. E-mail:* gabrielssousasjp@gmail.com.
[3] *IFMA-Instituto Federal do Maranhão.*
[4] *IFMA-Instituto Federal do Maranhão. E-mail:* luiz.garcia@ifma.edu.br.

Resumo: *A análise funcional utiliza combinações lineares de funções básicas como o principal método para representar funções. O uso de funções básicas é um dispositivo computacional bem adaptado para armazenar informações sobre funções, uma vez que é muito flexível e tem o poder computacional de encaixar até mesmo centenas de milhares de pontos de dados. Além disso, permite que os cálculos necessários sejam expressos dentro do contexto familiar da álgebra matricial o que facilita a implementação em software estatístico. Além disso, devido à simplicidade e eficácia para lidar com diferentes problemas de suavização semiparamétrica, a regressão funcional (Spline, Fourier, etc.) recentemente se tornou uma ferramenta popular para resolver vários problemas de estimativa nas mais variadas ciências. Neste artigo, usamos um estudo de simulação para comparar um método com nós equidistantes em um modelo de spline de regressão com um modelo de base Fourier. Ambos os métodos o número de nós para Spline e o número de bases para Fourier foram determinado pelo algorítimo busca direta. Em nosso estudo de simulação não identificamos vantagens entre os métodos.*

Palavras-chave: Regressão não paramétrica; Regressão semi-paramétrica; Splines; Regressão Funcional.

Abstract: *Functional analysis uses linear combinations of basic functions as the primary method for representing functions. The use of basic functions is a well-designed computing device for storing information about functions as it is very flexible and has the computing power to fit even hundreds of thousands of data points. In addition, it allows the necessary calculations to be expressed within the familiar context of matrix algebra which facilitates implementation in statistical software. In addition, because of the simplicity and effectiveness of dealing with different problems of semi-parametric smoothing, functional regression (Spline, Fourier, etc.) has recently become a popular tool for solving various estimation problems in the most varied sciences. In this paper, we used a simulation study to compare a method with equidistant nodes in a regression spline model with a Fourier-based model. Both methods the number of nodes for Spline and the number of bases for Fourier were determined by the direct search algorithm.In our simulation study we did not identify advantages between the methods.*

Keywords: Nonparametric regression; Semiparametric regression; splines; Functional Regression.

[†]Autor correspondente: ernandes.moura@ifma.edu.br.

Introdução

A análise de dados funcionais (FDA) é um campo de estudo que lida com a análise e teoria de dados cujas unidades de observação são funções (curvas) definidas em qualquer domínio contínuo (MORRIS, 2014). A maioria dos trabalhos em análise de dados funcional é baseada em uma variante do Modelo Linear Funcional (FLM), introduzida primeiramente por (Ramsay and Dalzell 1991). A análise de regressão funcional (FRA) e a análise funcional utilizam combinações lineares de funções bases como o principal método para representar funções. O uso de funções bases é uma ferramenta computacional bem adaptada para armazenar informações sobre funções, uma vez que é muito flexível e tem o poder computacional de executar centenas de milhares de pontos de dados (MONTESINOS-LÓPEZ *et. al*, 2018).

Existe uma gama de diferentes sistemas de funções de base, tais como funções de base polinomial, funções de base Wavelet, dentre outras. Duas das funções base mais populares serão abordadas nesse estudo, Fourier e Spline. Dessa forma, uma série de Fourier é uma expansão de uma função periódica em termos de uma soma infinita de combinações de senos e cossenos, enquanto que uma Spline nada mais é que um polinômio por partes, com limites em pontos chamados break points. Dessa forma, o objetivo deste artigo é comparar um ajuste de Spline com nós igualmente espaçados com um sistema de bases Fourier.

1 Material e Métodos

1.1 Dados simulados

Usamos R (R Core Team, 2018) para realizar as simulações e para escrever uma função suavizada utilizando Spline e Fourier. Um estudo de simulação com n = 250 observações foi gerado como conjunto de dados. Os x_i foram igualmente espaçados gerados de uma distribuição uniforme no intervalo [0, 1]. A funções de regressão (curva verdadeira) estudada é conhecida como 'função de colisão' (RUPPERT et al. 2003).

1.2 Fourier

Uma série de Fourier é uma expansão de uma função periódica em termos de uma soma infinita de combinações de senos e cossenos da seguinte forma:

$$f(x) = c_0 + \sum_{j=1}^{m} [c_{2j-1} sen(jwx) + c_{2j} \cos(jwx)] \quad (1)$$

Em que $b = 2m + 1$ é o número total de bases fourier. A constante ω está relacionada ao período τ pela relação $\omega = \frac{2\pi}{\tau}$ e, τ pode ser definido como amplitude do espaço da posição, conforme sugere (RAMSAY, HOOKER, and GRAVES 2009). Além disso, note que para garantir a ortogonalidade, o número total de bases b é sempre ímpar (intercepto e sucessivos pares seno/cosseno), de modo que para cada senóide que entra no modelo, um cossenoide também terá que fazer parte. Dessa forma, a estimativa de $f(x)$ pode ser alcançada por meio da estimativa dos coeficientes $\beta = (\beta_1, \cdots, \beta_n)^T$ através da matriz de bases Fourier B. Em que

$$\mathbf{B} = \begin{bmatrix} 1 & sen(\omega x_1) & cos(\omega x_1) & sen(2\omega x_1) & cos(2\omega x_1) & \cdots & sen\left(\frac{b-1}{2} w x_1\right) & cos\left(\frac{b-1}{2} w x_1\right) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & sen(\omega x_n) & cos(\omega x_n) & sen(2\omega x_n) & cos(2\omega x_n) & \cdots & sen\left(\frac{b-1}{2} w x_n\right) & cos\left(\frac{b-1}{2} w x_n\right) \end{bmatrix}$$

Uma estimativa de mínimos quadrados $\hat{\beta}$ para β pode ser obtida por mínimos quadrados. Logo,

$$\hat{\beta} = (\mathbf{B^T B})^{-1} \mathbf{B^T y} \tag{2}$$

Logo, para nosso exemplo fictício temos, $\hat{\mathbf{f}}(\mathbf{x}) = \mathbf{B}\hat{\beta}$.

1.3 Spline

São mais flexíveis que as séries de Fourier e é caracterizada por o número de nós (pontos em que os segmentos se conectam), a ordem e o grau do polinômio. Considere um vetor de observações $y = \{y_1, y_2, \cdots, y_n\}$, satisfazendo o modelo $y_i = f(x_i) + \varepsilon_i$, em que $f(x)$ é uma função de regressão desconhecida e os ε_i são erros independentes com variância constante σ^2. Assumimos que $f(x)$ pode ser modelado por um spline de grau p da seguinte forma:

$$f(x) = \beta_0 + \beta_1 x + \cdots + \beta_p x^p + \sum_{k=1}^{K} \beta_{pk} (x - \xi_k)_+^p \tag{3}$$

Onde $\xi_1, \cdots, \xi_K$ é um conjunto de nós pré-fixados e, geralmente, são igualmente espaçados e, $(x - \xi_k)_+ = (x - \xi_k)$ se $x \geq \xi_k$ ou zero caso contrário. Assim, a estimativa de $f(x)$ pode ser alcançada através da estimativa dos coeficientes $\beta_0, \beta_1, \cdots, \beta_p, \beta_{p1}, \cdots, \beta_{pK}$ da seguinte maneira. Sejam $\mathbf{y} = \{y_1, \cdots, y_n\}^T$,$\varepsilon = \{\varepsilon_1, \cdots, \varepsilon_n\}^T$, e

$$X = \begin{bmatrix} 1 & x_1 & \cdots & x_1^p & (x_1 - \xi_1)_+^p & \cdots & (x_1 - \xi_K)_+^p \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \cdots & x_n^p & (x_n - \xi_1)_+^p & \cdots & (x_n - \xi_K)_+^p \end{bmatrix}$$

Usando esta notação, o modelo matricial fica:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon \tag{4}$$

Uma estimativa de mínimos quadrados $\hat{\beta}$ para β pode ser obtido. Logo,

$$\hat{\beta} = (\mathbf{X^T X})^{-1} \mathbf{X^T y} \tag{5}$$

Então, $\hat{f}(x)$ é obtido por:

$$\hat{\mathbf{f}}(\mathbf{x}) = \mathbf{X}\hat{\beta} \tag{6}$$

2 Resultados e Discussão

2.1 Fourier

O bom desempenho do modelo de regressão funcional depende fortemente da escolha do tipo certo de funções básicas, do número necessário de bases, do grau do polinômio

(para Splines), dos nós (para Splines), do período (em Fourier), entre e outros. Assim sendo, na figura 1 estão representadas a curva verdadeira no painel A, as curvas ajustada dos painéis de B-D, sendo que no painel D estão representadas as curva que maximizou a correlação com a verdadeira, sendo portanto, considerada ótima para esse estudo.

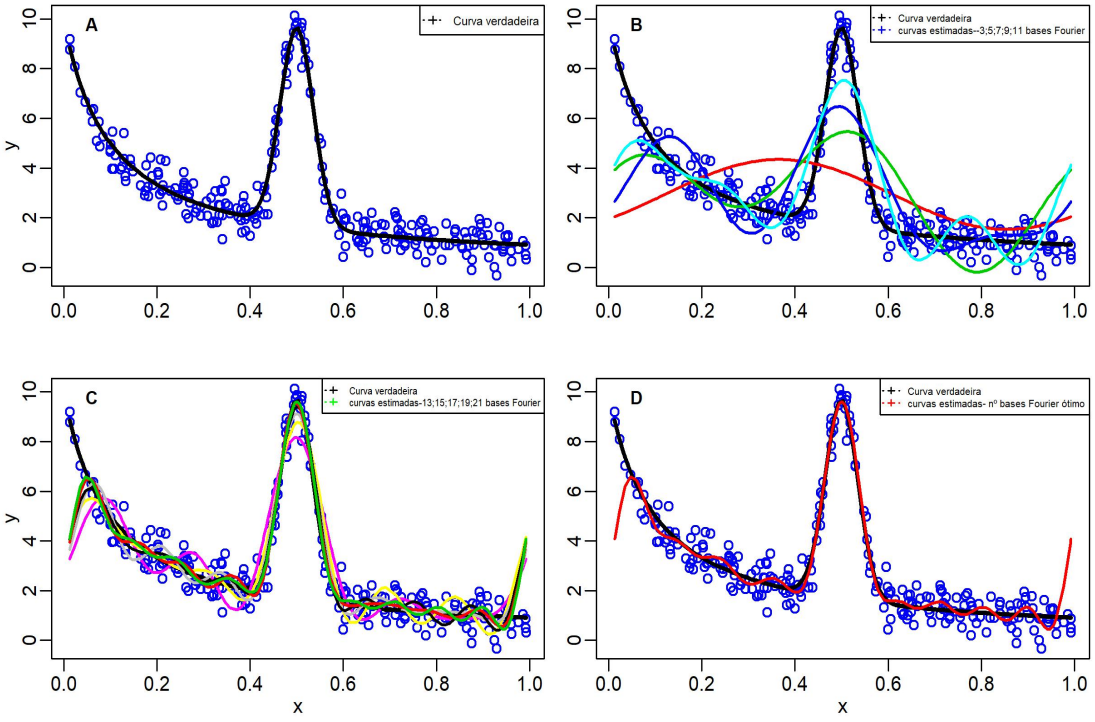

Figura 1: Diagrama de dispersão do fenômeno hipotético. Os pontos coloridos representam os 250 pontos de dados simulados. A curva verdadeira se encontra no painel A (em preto). No painel B, estão representas as curvas suavizadas (3;5;7;9 e 11) bases Fourier. No painel C, estão representas as curvas suavizadas (13;15;17;19 e 21) bases Fourier. Finalmente, no painel D estão representadas as curvas: verdadeira (em preto) e curva que maximizou a correlação em vermelho. Fonte: Do autor (2019).

2.2 Spline

As funções bases de uma spline são definidas de tal maneira que a função é contínua e tem derivações contínuas em todo seu domínio (incluindo nos nós) de ordem q-1; aqui q é a ordem do polinômio, e foi assumido grau $p = 3$. Todavia, a ideia é escolher uma combinação de nós que otimiza o critério de informação escolhido (correlação nesse estudo). No entanto, antes de contemplar tal abordagem, vale a pena considerar o número de modelos possíveis (RUPPERT, 2003). Assim, utilizamos o método de busca completa e ajustamos as curvas splines desde de um nó até 20 nós em busca de uma configuração ideal. Esses resultados estão ilustrados na Figura 2 nos painéis de A-D. O número de nós que proporcionou a melhor suavização foi $K = 16$ e está representado na Figura 2 no painel D.

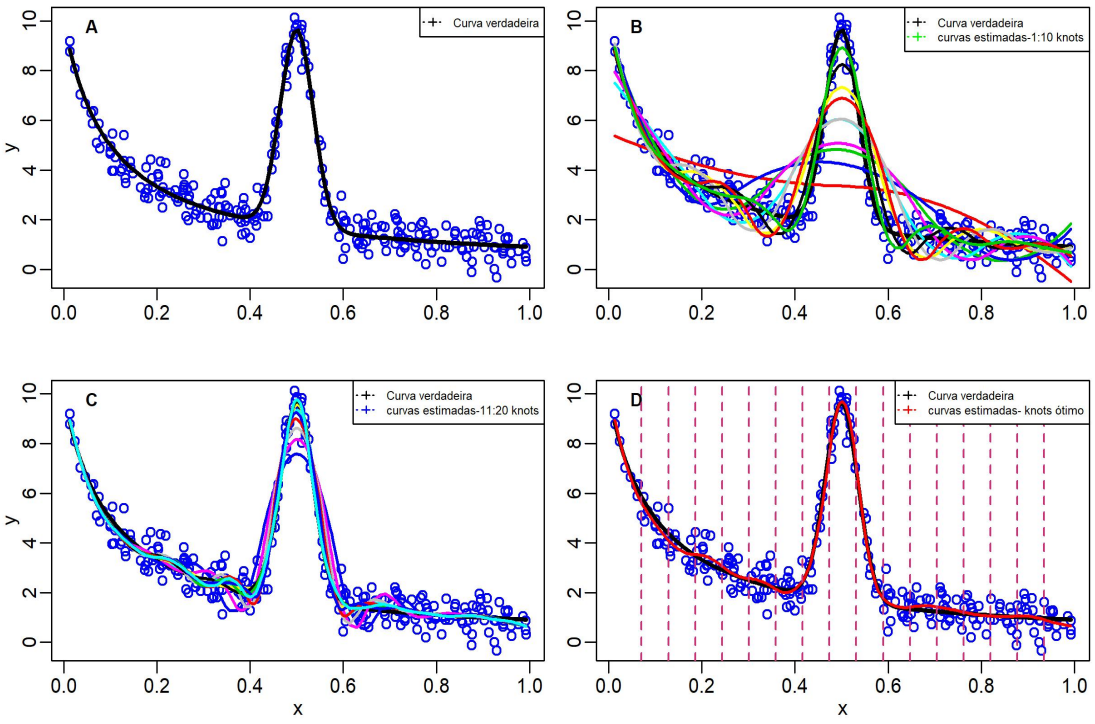

Figura 2: Diagrama de dispersão do fenômeno hipotético. Os pontos coloridos representam os 250 pontos de dados simulados. A curva verdadeira se encontra no painel A (em preto). No painel B, estão representadas as curvas suavizadas (1 a 10) nós. No painel C, estão representadas as curvas suavizadas (11 a 20) nós. Finalmente, no painel D estão representadas as curvas: verdadeira em preto e a que maximizou a correlação ($K = 16$) em vermelho. Linhas tracejadas em vermelho, representam os 16 knots. Fonte: Do autor (2019).

2.3 Fourier x Spline

Dois tipos satisfatórios de base são a base de Fourier para curvas periódicas e o modelo Spline para curvas não periódicas. Comparamos essas duas técnicas em um cenário simples de simulação. Utilizou-se o algorítimo busca completa para encontrar um modelo ótimo em ambos os casos. Esses resultados estão representados na Figura 3 abaixo.

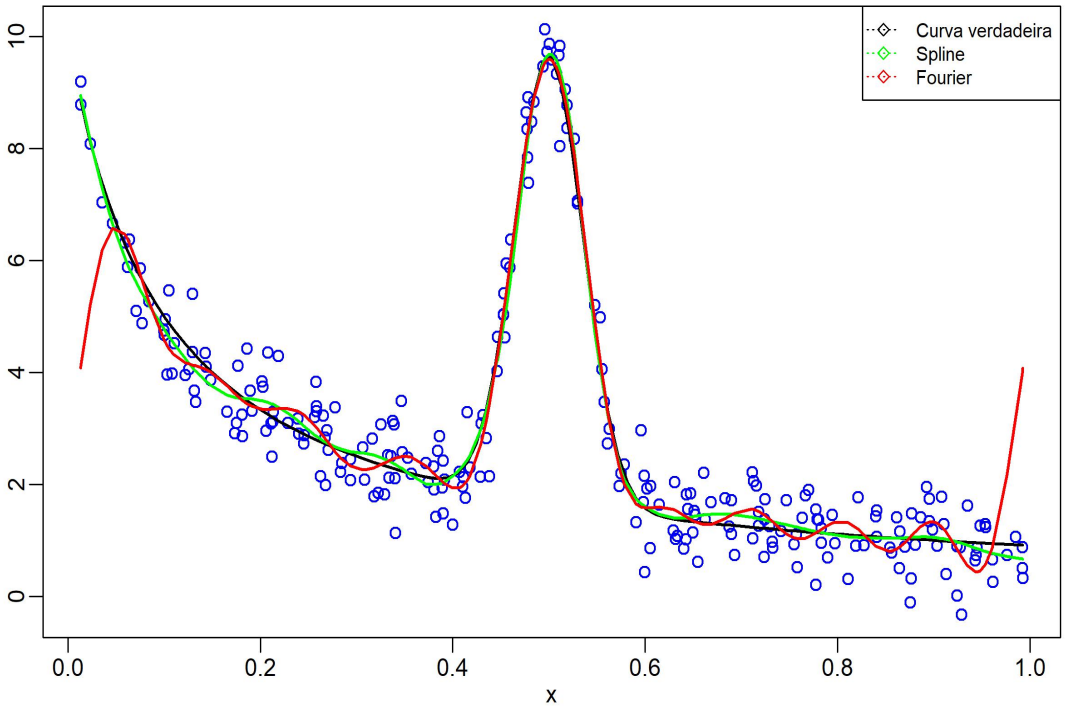

Figura 3: Diagrama de dispersão do fenômeno hipotético. Os pontos representam os 250 pontos de dados simulados. A curva verdadeira em preto, curva suavizada via bases Fourier em vermelho e a curva suavizada via Spline em azul. Fonte: Do autor (2019).

Conclusão

Sistema de Bases de Fourier e Splines são métodos de suavização largamente utilizados. Comparamos o desempenho desses dois métodos de Sistemas de Bases com nós igualmente espaçados para Spline. Em nosso estudo de simulação, não identificamos vantagens entre os métodos. No entanto, num senário de dados reais o pesquisador tem que ter em mente que a escolha do sistema de bases adequada é de suma importância. Assim sendo, o bom desempenho do modelo de regressão funcional depende fortemente da escolha do tipo certo de funções bases, do número necessário de bases, do grau do polinômio (para Splines), dos nós (para Splines) e, do período (para Fourier).

Agradecimentos

Gostaríamos de agradecer ao Instituto Federal do Maranhão - IFMA pelo apoio estrutural e financeiro para concretização desse trabalho.

Referências Bibliográficas

MONTESINOS-LÓPEZ, A.; MONTESINOS-LÓPEZ, O.A.; LOS CAMPOS, G.; CROSSA, J.; BURGUEÑO, J.; LUNA-VAZQUEZ, J. Bayesian functional regression as an alternative statistical analysis of high-throughput phenotyping data of modern agriculture. *Plant methods*, v.14, n.1, p.46, 2018.

MORRIS, J.S. Functional regression. *Annual Review of Statistics and Its Application*, v.2, p.321-359, 2015.

RAMSAY, J.O.; DALZELL, C.J. Some tools for functional data analysis. *Journal of the Royal Statistical Society: Series B (Methodological)*, v.53, n.3, p.539-561, 1991.

RAMSAY, J.O.; HOOKER, G.; GRAVES, S. *Functional Data Analysis with R and MATLAB*. New York: Springer Science e Business Media, 2009.

RUPPERT, D.; WAND, M. P.; CARROLL, Raymond J. *Semiparametric regression*. Cambridge university press, 2003.

R CORE TEAM. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. 2018. ISBN 3-900051-07-0, URL https://www.R-project.org/.