ISSN: 2317-0840

Estratégia para posicionamento de nó para regressão spline penalizado

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

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

Resumo: *Apresentamos um novo método para a seleção de sequências de nós para curvas de regressão P-spline. O método parte do pressuposto que os próprios dados determinam a quantidade e a posição dos nós. Assim, esse novo esquema de colocação de nós assume que os nós são uma variável aleatória através de uma grade fina de possíveis candidatos a nós no intervalo da co-variável. Dessa forma, através de uma busca em grade determina-se o knot que maximiza a correlação em cada iteração. Esse novo esquema de colocação de nós obteve excelentes resultados comparativamente a métodos convencionais de alocação de nós em um estudo de simulação e, além disso, nosso estudo de simulação evidencia que essa estratégia torna o modelo mais parcimonioso. Os resultados fornecem orientação na seleção do número de nós não necessariamente equidistantes em um modelo de de regressão spline penalizada.*

Palavras-chave: Regressão não paramétrica; Regressão semi-paramétrica; Splines penalizados; Colocação do nós.

Abstract: *We present a new method for the selection of node sequences for P-spline regression curves. The method assumes that the data themselves determine the number and position of the nodes. Thus, this new node placement scheme assumes that nodes are a random variable across a thin grid of possible candidate nodes in the covariate range. Thus, through a grid search, we determine the knot that maximizes the correlation in each iteration. This new node placement scheme has obtained excellent results compared to conventional node allocation methods in a simulation study and, furthermore, our simulation study shows that this strategy makes the model more parsimonious.The results provide guidance in selecting the number of nodes not necessarily equidistant in a penalized spline regression model.*

Keywords: Nonparametric regression; Semiparametric regression; Penalized splines; Knot placement.

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

Introdução

Uma spline é uma função contínua formada pela conexão de segmentos de polinômios sobe certas condições específicas. Os pontos em que os segmentos se conectam são chamados de nós da spline. Devido à sua simplicidade e eficácia para lidar com diferentes problemas de suavização semiparamétrica, a regressão por spline penalizada recentemente se tornou uma ferramenta popular para resolver vários problemas de estimativa (Yao e Lee, 2008).

A forma de uma spline pode ser controlada escolhendo-se cuidadosamente o número de nós e suas localizações exatas. Dessa forma, uma das grandes vantagens de se utilizar regressão spline é devido a flexibilidade onde a tendência (padrão) dos dados muda rapidamente. Assim sendo, o principal desafio é escolher o número ideal de nós e sua respectiva localização. Em particular, Ruppert et al. (2003, Seção 5.5.3) propõem uma estratégia para obter o número e a localização dos nós (denominado de knot fixo nesse trabalho).Todavia, embora essa estratégia funcione bem em muitas situações práticas, não faz uso de nenhuma informação dos dados, exceto o tamanho da amostra. Ademais, devido a essa particularidade, não há garantia que os nós sejam colocados em locais críticos, isto é, locais nos quais a função de regressão subjacente possui mudanças nítidas.

Diante do exposto, se faz necessário, um algoritmo mais geral, que use os dados para escolher o número de nós e, suas respectivas posições. Dessa forma, o objetivo deste artigo é propor um novo esquema de colocação de nós em regressão spline, em que, ao invés de determinar os nós e suas localizações a priori, assume-se que cada nó é uma variável aleatória e, assim, os dados é quem determinam a quantidade e a localização dos nós. Este esquema consiste em ajustar uma spline começando com apenas um nó e, estabelece uma grade fina de possíveis candidatos ao primeiro nó, faz-se as estimativas para todos os "candidatos" e escolhe-se o nó ótimo por algum critério (correlação por exemplo), posteriormente, repete-se o processo para escolha do segundo nó ótimo, dado que já foi determinado o primeiro nó ótimo na primeira iteração, repete-se o processo por um número razoável de iterações. Este novo esquema de colocação de nós teve um ótimo desempenho no estudo de simulação relatado abaixo.

Material e Métodos

Dados simulados

Usamos R (R Core Team, 2018) para realizar as simulações e para escrever uma função para encaixar os nós para os dois métodos de seleção de nós e parâmetros de suavização descritos na Seção 1.3. Um estudo de simulação com n = 100 observações por conjunto de dados (três conjuntos de dados, resultados não mostrados). Os x_i foram igualmente espaçados gerados de uma distribuição uniforme no intervalo [0, 1]. Uma função (curva verdadeira) de regressão foi simulada e, enquanto os erros ε_i são iid $N(0, 0.15^2)$. O parâmetro de suavização λ foi escolhido pela validação cruzada generalizada (GCV) e o grau da spline é $p = 3$.

Spline penalizado

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 penalizado de grau p da seguinte forma:

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

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

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

Considerando as definições supracitadas, o modelo matricial fica:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon. \tag{2}$$

Então, para um determinado parâmetro de suavização λ , uma estimativa de mínimos quadrados penalizados $\hat{\beta}_\lambda$ para β pode ser obtido. Logo,

$$\hat{\beta}_\lambda = (\mathbf{X^T X} + \lambda \mathbf{D})^{-1} \mathbf{X^T y}, \tag{3}$$

em que $D = diag\{0_{p+1\,\mathrm{x}\,p+1}, 1_{K\,\mathrm{x}\,K}\}$ ou, matricialmente,

$$D = \left[\begin{array}{cccc|cccc} 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 1 \end{array}\right] = \begin{bmatrix} 0_{p+1\,\mathrm{x}\,p+1} & 0_{p+1\,\mathrm{x}\,K} \\ 0_{K\,\mathrm{x}\,p+1} & I_{K\,\mathrm{x}\,K} \end{bmatrix}.$$

Então, $\hat{f}(x)$ é obtido por:

$$\hat{\mathbf{f}}_\lambda(\mathbf{x}) = \mathbf{X}\hat{\beta}_\lambda. \tag{4}$$

A qualidade das estimativas, depende das escolhas de $(\xi_1, \cdots, \xi_K)$ e de λ. A próxima seção apresentamos uma nova estratégia para colocação de nós, não necessariamente igualmente espaçados.

Esquema proposto de colocação de nós

Motivação

A maioria dos métodos existentes de seleção de nós (knots) são "igualmente espaçados" (Yao e Lee, 2008; Ruppert ,2002) . Todavia, não há garantia que esses nós igualmente espaçados fiquem em posições em que realmente seja necessário, isto é, não há garantia que mesmo saturando o espaço de x com uma grande quantidade de nós igualmente espaçado, esses se localizem em regiões que favoreçam a captar máximos e mínimos locais e globais em $f(x)$. Ruppert et al. (2003, Seção 5.5.3) propôs o seguinte método de escolha de nós:

$$\mathrm{K} = \min(\frac{1}{4} \text{ número de unique } x_i, 35) \tag{5}$$

$$\xi_k = (\frac{k+1}{K+2}) \text{ quantil da amostra unique } x_i \tag{6}$$

em que $k = 1, \cdots, K$. Assim, K é o número total dos nós. Esse método de colocação de nós é simples e fácil de implementar. No entanto, conforme ilustrado na próxima subseção, esse método não garante que os nós sejam colocados em todos os locais críticos. Diante dessa limitação, se faz necessário um novo esquema de colocação de nós, que realmente detecte locais críticos.

O procedimento proposto

Conduzido pela discussão supracitada, o seguinte procedimento de colocação de nós foi proposto para concorrer com os métodos de colocação de nós igualmente espaçados.

[1] Estabelece-se uma grade fina com pontos igualmente espaçados no intervalo [min(x),max(x)].

[2] Assume-se que cada ponto de [1] seja um candidato ao primeiro nó ideal ${\xi_1}'$ e, dessa forma, o nó ideal é escolhido por algum critério de informação (máxima correlação nesse estudo).

[3] Dado que escolheu um nó ideal em [2], repete o procedimento para encontrar um segundo nó ideal ${\xi_2}'$ para fazer par com o primeiro nó ideal ${\xi_1}'$ e, assim, ajustando um modelo com dois nós.

[4] Repete-se o processo de incremento de nós até obter o ${\xi_K}'$ e, por último, obter um novo conjunto de nós para análise final da seguinte forma $(\xi_1, \cdots, \xi_K)$ = unique(${\xi_1}', \cdots, {\xi_K}'$), esse processo retira os inúmeros nós repetidos.

Para uma melhor compreensão, considera a ilustração abaixo das duas primeiras iterações.

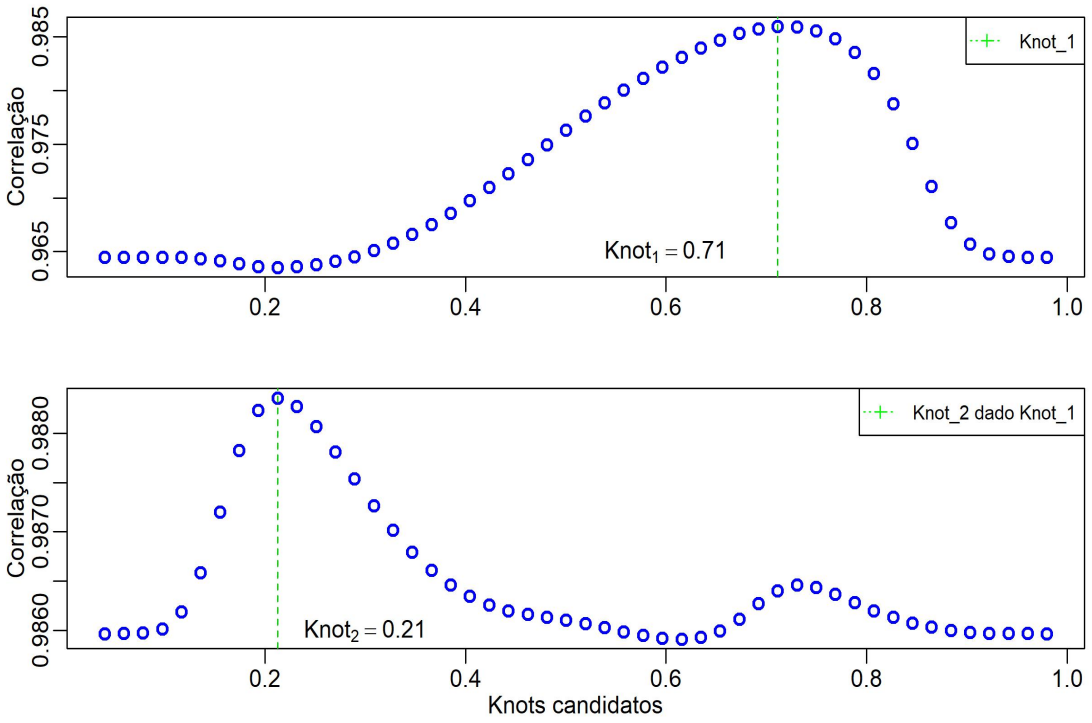

Figura 1: Escolhas dos Knots ótimos. Fonte: Do autor (2019).

Pela Figura 1, pode-se observar que entre todos os nós candidatos, o 0.71 foi o que maximizou a correlação na primeira iteração. Logo, esse foi adotado como primeiro nó ótimo e foi fixado para próxima iteração. Note ainda, que dado o primeiro nó ótimo, o segundo nó que maximizou a correlação foi 0.21. Assim para a terceira iteração já havia dois nós fixados e, assim por diante até completar o número de iteração desejada (100 iterações nesse estudo).

Escolhendo o parâmetro de suavização

O papel do parâmetro de suavização na spline penalizada é controlar a suavidade da curva ajustada. Para calcular o valor ideal do parâmetro de suavização, o critérios de seleção foi considerado o parâmetro de validação cruzada generalizada (GCV). O método GCV é computacionalmente simples e muito bem usado na literatura em regressão splines (CRAVEN e WAHBA, 1978). O método GCV consiste em selecionar λ de modo que minimize

$$GCV(\lambda) = \frac{SSE}{(1 - df_\lambda/n)^2} \tag{7}$$

em que df_λ são os "graus de liberdade"correspondente ao parâmetro de suavização λ e é definido como traço da matriz chapéu H_λ ou matriz suavizadora no contexto de regressão não paramétrica, em que $H_\lambda = X(X^TX + \lambda D)^{-1}X^T$. Assim, $df_\lambda = tr[H_\lambda]$ e o SSE é a soma dos quadrados dos resíduos dado por $SSE = \sum_{i=1}^{n}(y - \widehat{f_\lambda(x)})^2$. Note que a matriz chapéu H_λ é quadrada, simétrica e de ordem n e, é uma função de λ. Ela tem muitos usos, entre os quais, obter a medida dos graus de liberdade efetivos do ajuste definido acima.

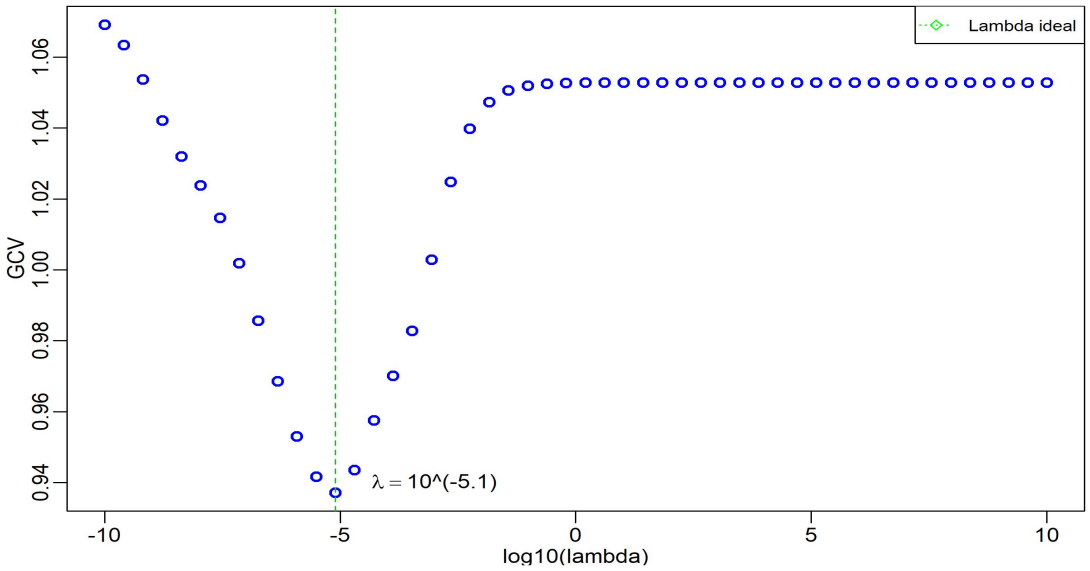

Figura 2: Valores de validação cruzada generalizada(GCV) para obtenção do lambda ótimo para ajuste da curva de suavização spline. Fonte: Do autor (2019).

A Figura 2 ilustra a obtenção do lambda ótimo para uma iteração. Neste caso, o lambda que minimizou validação cruzada generalizada(GCV) foi de $\lambda = 10^{-5.1}$, sendo esse considerado como ótimo para a respectiva iteração.

Resultados e Discussão

Uma das principais razões por trás do sucesso do método proposto é que, muitas vezes, os locais críticos para a colocação dos nós estão nos extremos locais da função-alvo, enquanto, métodos de alocações de nós igualmente espaçados, não garante nós em locais de extremos locais. Para avaliar as estimativas de regressão dos dois métodos, utilizou-se o critério de correlação de Pearson, bem como uma análise visual. Dessa forma, pode-se observar pela Figura 3**B** um melhor ajuste em relação à Figura 3**A**.

É importante ressaltar que nosso método difere do proposto por Montoya *et.al* (2014), em que cada ponto de observação foi considerado como um nó. Entretanto, como os nós para o presente trabalho são variáveis aleatórias, pode ocorrer de coincidir que determinado nó seja a própria observação. Ademais, o estudo de simulação sugere que o modelo proposto é muito mais parcimonioso que o método proposto por Ruppert et al. (2003). Na figura 3 abaixo estão representadas as estimativas de regressão e a função verdadeira para o método proposto e o método proposto por Ruppert et al. (2003).

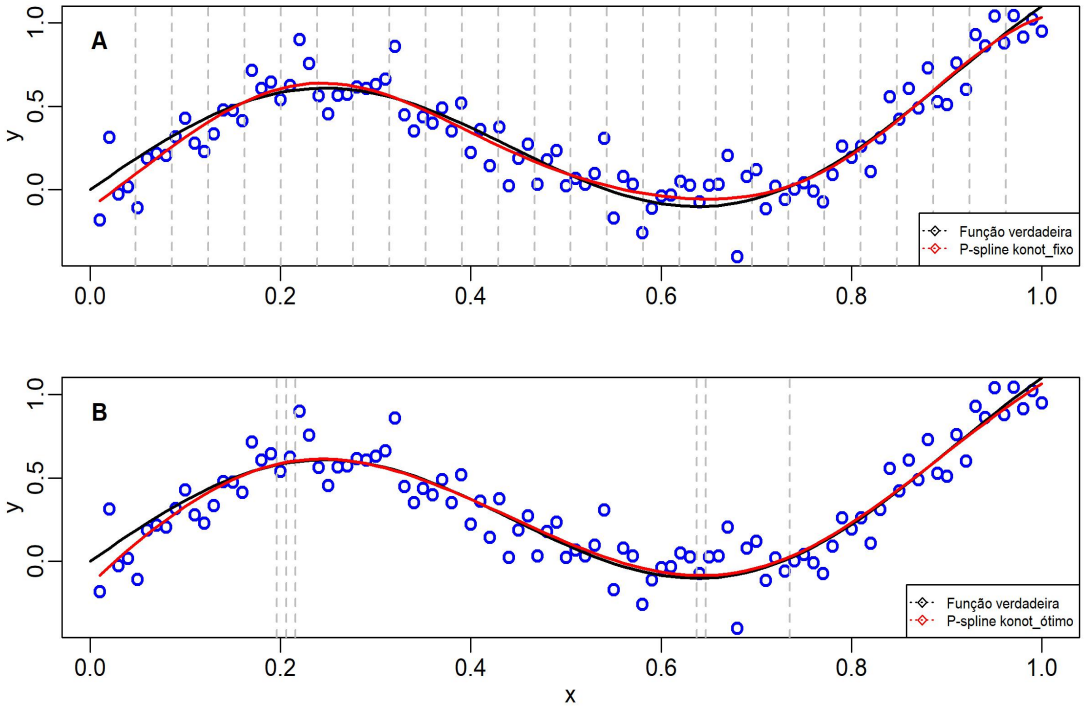

Figura 3: Diagrama de dispersão do fenômeno hipotético. Os pontos em azul representam os 100 pontos de dados medidos. Regressão P-spline com knots fixo igualmente espaçados e knots ótimos. Fonte: Do autor (2019).

Conclusão

Nós comparamos o desempenho do método proposto com um método de seleção de nós fixos e igualmente espaçados em um modelo de spline de regressão penalizado. Deve-se ter em mente que nossos resultados surgem de um estudo de simulação. No entanto, sob nossas configurações de simulação, assumir os nós como variáveis aleatórias em que os dados determinam a quantidade e a posição exata desses nós, faz bem tanto em ajuste como em parcimônia de modelo. Dessa forma, os resultados de simulação sugerem que o método proposto é mais eficiente que os de alocação de nós igualmente espaçados.

Agradecimentos

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

Referências Bibliográficas

CRAVEN, P.; WAHBA, G. Smoothing Noisy Data with Spline Functions - Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation. *Numerische mathematik*, v.31, n.4, p.377-403, 1978.

MONTOYA, E. L.; ULLOA, N.; MILLER, V. A simulation study comparing knot selection methods with equally spaced knots in a penalized regression spline. *International Journal of Statistics and Probability*, v.3, n.3, p.96-110, 2014.

RUPPERT, D. Selecting the number of knots for penalized splines. *Journal of computational and graphical statistics*, v.11, n.4, p.735-757, 2002.

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

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

YAO, F.; LEE, T. CM. On knot placement for penalized spline regression. *Journal of the Korean Statistical Society*, v.37, n.3, p.259-267, 2008.