SPLINES DE NÓS LIVRES PARA AJUSTE DE CURVAS EM PROBLEMAS NA ÁREA DA SAÚDE

Tiago Pereira Marques

Dissertação apresentada à Universidade Estadual Paulista "Júlio de Mesquita Filho" para a obtenção do título de Mestre em Biometria.

BOTUCATU
São Paulo - Brasil
Fevereiro– 2021

SPLINES DE NÓS LIVRES PARA AJUSTE DE CURVAS EM PROBLEMAS NA ÁREA DA SAÚDE

Tiago Pereira Marques

Orientadora: Profa. Dra. **Miriam Harumi Tsunemi**

Dissertação apresentada à Universidade Estadual Paulista "Júlio de Mesquita Filho" para a obtenção do título de Mestre em Biometria.

BOTUCATU
São Paulo - Brasil
Fevereiro– 2021

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉC. AQUIS. TRATAMENTO DA INFORM.
DIVISÃO TÉCNICA DE BIBLIOTECA E DOCUMENTAÇÃO - CÂMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: ROSEMEIRE APARECIDA VICENTE-CRB 8/5651

Marques, Tiago Pereira.
Splines de nós livres para ajuste de curvas em problemas na área da saúde / Tiago Pereira Marques. - Botucatu, 2021

Dissertação (mestrado) - Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências de Botucatu
Orientador: Miriam Harumi Tsunemi
Capes: 90194000

1. Biometria. 2. Modelos lineares. 3. Dinâmica não linear. 4. Teorema de Bayes. 5. Inferência Bayesiana.

Palavras-chave: BASS; Inferência Bayesiana; RJMCMC; Regressão linear; Splines.

ATA DA DEFESA PÚBLICA DA DISSERTAÇÃO DE MESTRADO DE TIAGO PEREIRA MARQUES, DISCENTE DO PROGRAMA DE PÓS-GRADUAÇÃO EM BIOMETRIA, DO INSTITUTO DE BIOCIÊNCIAS - CÂMPUS DE BOTUCATU.

Aos 25 dias do mês de fevereiro do ano de 2021, às 08:30 horas, por meio de Videoconferência, realizou-se a defesa de DISSERTAÇÃO DE MESTRADO de TIAGO PEREIRA MARQUES, intitulada **SPLINES DE NÓS LIVRES PARA AJUSTE DE CURVAS EM PROBLEMAS NA ÁREA DA SAÚDE**. A Comissão Examinadora foi constituida pelos seguintes membros: Profa. Dra. MIRIAM HARUMI TSUNEMI (Orientador(a) - Participação Virtual) do(a) Departamento de Bioestatística, Biologia Vegetal, Parasitologia e Zoologia / Instituto de Biociências de Botucatu - UNESP , Profa. Dra. LUZIA APARECIDA TRINCA (Participação Virtual) do(a) Departamento de Bioestatística / Instituto de Biociências de Botucatu - UNESP , Prof. Dr. CRISTIAN MARCELO VILLEGAS LOBOS (Participação Virtual) do(a) Depto. de Ciências Exatas / Esalq / Universidade de São Paulo. Após a exposição pelo mestrando e arguição pelos membros da Comissão Examinadora que participaram do ato, de forma presencial e/ou virtual, o discente recebeu o conceito final:_ _ Aprovado _ _ _ . Nada mais havendo, foi lavrada a presente ata, que após lida e aprovada, foi assinada pelo(a) Presidente(a) da Comissão Examinadora.

Miriam Harumi Tsunemi

Profa. Dra. MIRIAM HARUMI TSUNEMI

Dedicatória

Dedico a quem me faz sonhar!

Agradecimentos

Agradeço a CAPES por oferecer a bolsa de mestrado com a qual pude seguir até aqui no programa de pós-graduação em Biometria (Código de Financiamento 001).

Agradeço ao grupo PET-Engenharia Agronômica da UNESP de Botucatu-SP, grupo responsável por criar minha vontade em ser docente, a qual, sem dúvida, foi uma das razões que me fez optar em seguir para o mestrado. Os inúmeros cursos que organizamos e ministramos fizeram-me acreditar que talvez tivesse prazer em ensinar, o que, para mim, é o mais importante para ser professor, até mesmo em nível universitário.

Agradeço a meus colegas da pós-graduação por tantos bons momentos vividos, pelas discussões envolvendo os conteúdos das disciplinas, tópicos diversos em matemática e estatística e, até mesmo, as discussões banais que faziam o dia ser mais interessante.

Aos professores do programa de pós-gradução em Biometria pela dedicação e prontidão a tudo que foi a eles solicitados.

Um agradecimento especial à minha orientadora Miriam Harumi Tsunemi, por ser um poço de paciência comigo e pelas contribuições ao trabalho.

Agradeço aos professores Prof. Dr. Vladimir Eliodoro Costa e Prof. Dr. Ricardo Miguel Costa de Freitas, e também à Beatriz de Oliveira Garcia por fornecerem os bancos de dados utilizados no trabalho.

Agradeço à banca por ter aceito prontamente o convite e dedicar parte de seu tempo a fazer correções e possivelmente contribuições fundamentais ao trabalho.

Agradeço à minha família, com a qual ainda resido, por me dar muito carinho e incentivo todos os dias para seguir em frente, sempre confiando em mim e no meu potencial.

Agradeço à minha namorada, Ariadne Magalhães Carneiro, a qual conheci durante a disciplina de Princípios de Inferência Estatística, por ter tornado um dos momentos mais desafiadores da minha vida em um dos mais felizes e belos, sendo uma das pessoas mais carinhosas que conheci. Uma pessoa que sempre quero fazer sorrir e com quem quero estar acompanhado para viver tantos outros momentos. E também a seus pais por terem me tratado tão bem ao me receberem em sua casa e por apoiarem o nosso relacionamento.

Agradeço a Deus, principalmente, por ter me dado o direito a vida, tão surpreendente e intensa, e por dar razão a coisas que não veria sentido sem crer em sua existência.

Sumário

Lista de Figuras

Página

Lista de Tabelas

Página

SPLINES DE NÓS LIVRES PARA AJUSTE DE CURVAS EM PROBLEMAS NA ÁREA DA SAÚDE

Autor: TIAGO PEREIRA MARQUES
Orientadora: Profa. Dra. MIRIAM HARUMI TSUNEMI

RESUMO

Encontrar modelos que expliquem a relação entre variáveis é uma das principais aplicações da estatística. Os modelos mais simples são os chamados modelos de regressão linear que são, muitas vezes, incapazes de aproximar apropriadamente a relação entre as variáveis, a qual nem sempre é linear. Para essas situações, podem ser usados os modelos de regressão não-linear, os quais demandam conhecimentos específicos do processo envolvido e ferramentas mais complexas para ajuste que os modelos lineares. Outra opção é aproximar funções não-lineares por expansões de base. Expansões de base por *splines* aproximam funções não-lineares por polinômios segmentados de ordem q contínuos até a sua derivada de ordem $q-2$. O maior desafio nesses métodos é encontrar o posicionamento correto dos nós (onde dividir os intervalos para melhor aproximar a função desejada). Métodos chamados de *splines* de nós livres consideram a posição dos nós, e muitas vezes o

número, como parâmetros livres que são, então, definidos pelos dados. No contexto Bayesiano, existem metodologias fundamentadas no algoritmo MARS (*Multivariate Adaptive Regression Splines*), que selecionam os nós adaptativamente com o método de Monte Carlo via cadeias de Markov com saltos reversíveis (RJMCMC). O presente trabalho é focado no estudo de uma dessas metodologias, *Bayesian Adaptive Splines Surfaces* (BASS), no contexto de ajuste de curvas. A metodologia foi aplicada em dois problemas da área da saúde. Um com objetivo de validar um protocolo de crioablação vertebral analisando as curvas de temperatura na sonda de crioablação e em termopares a distâncias crescentes das sondas, e o outro com o objetivo de determinar o período ótimo, após ingestão da uréia ^{13}C, para a realização do teste respiratório para detecção da presença de *Helicobacter pylori* analisando as curvas de DOB (*Delta Over Baseline*). Os resultados do método BASS foram comparados aos resultados de método de *splines* de nós livres que utiliza algoritmo genético para encontrar a posição dos nós. Apesar do método levar a pistas interessantes a respeito dos problemas da área da saúde, na avaliação dos modelos foram observados graves problemas de heterocedasticidade.

Palavras-Chave: RJMCMC, BASS, Splines, Regressão Linear e Inferência Bayesiana.

FREE-KNOT SPLINES FOR CURVE FITTING IN HEALTH PROBLEMS

Author: TIAGO PEREIRA MARQUES
Adviser: Profa. Dra. MIRIAM HARUMI TSUNEMI

SUMMARY

Finding models that explain the relationship among variables is one of the main applications of statistics. The simplest models are the so-called linear regression models which, many times, are unable to approximate properly the relationship among variables that is not always linear. For these situations, non-linear regression models can be used, which demand specific knowledge of the process involved and more complex tools for fitting than linear models. Another option is to approximate these non-linear functions through basis expansions. Spline basis expansion approximate non-linear functions by segmented polynomials of q order that are continuous up to its $q-2$ order derivative. The greatest challenge in these methods is to find the right placement of knots (where to divide the intervals to better approximate the desired function). Free-knot splines methodologies set the positions of the knots, and sometimes the number of knots, as free parameters that are, then,

defined by the data. In the Bayesian context, there exist methodologies inspired by the MARS algorithm, that use reversible jump Markov chain Monte Carlo methods (RJMCMC) to adaptively select the knots' locations and number. The following work focus on the study of one of those methodologies, Bayesian Adaptive Splines Surfaces (BASS), in the context of curve fitting. The methodology was applied in two health problems. In one the goal is to validate a vertebral cryoablation protocol analyzing temperature curves in the cryoablation catheter and at the thermocouples placed at growing distances from the cryoablation catheter and in the other problem the goal is finding the optimal time, after ingestion of ^{13}C urea, to take the breath test to *Helicobacter pylori* detection by analyzing DOB (Delta Over Baseline) curves. The BASS results were compared to a free-knot splines approach that uses a genetic algorithm to find the optimal positions of knots. Despite presenting interesting clues about the health problems, several heteroscedasticity problems were observed in model evaluation.

Keywords: RJMCMC, BASS, Splines, Linear Regression and Bayesian Inference.

1 INTRODUÇÃO

Ao se trabalhar com modelagem estatística, são frequentes as situações em que um modelo linear não é suficiente para explicar a relação entre as variáveis. Os chamados modelos não-lineares se fundamentam em processos físicos que precisam ser conhecidos a fundo para que se consiga descrever a relação entre as variáveis a serem modeladas, além de exigir ferramentas mais complexas para o ajuste.

Uma alternativa para explorar relações de não-linearidade é realizar transformações nas variáveis de forma a obter um novo conjunto de variáveis entre as quais a relação que se estabelece é linear. Uma das formas mais comuns de transformação é a expansão polinomial, na qual a relação não-linear entre variáveis é aproximada por meio de um polinômio. A regressão *spline* é uma estratégia de regressão polinomial segmentada, em que o intervalo de variação da variável independente é dividido em vários segmentos ajustando, em cada, um polinômio distinto, todos com uma mesma ordem q, cuja função resultante é contínua até a derivada de ordem $q-2$, utilizada quando a relação entre as variáveis for muito complexa, tal que a aproximação por vários polinômios de baixo grau é mais conveniente que por um único de grau elevado.

Através de estudo sobre o tema, notou-se que a proposição de novas metodologias para encontrar a posição ótima dos nós, no contexto das *splines* de nós livres, é um campo de pesquisa promissor. Este tema também mostra-se interessante para estudos mais aprofundados sob a ótica Bayesiana onde se destaca a proposta de Denison et al. (1998), que adaptou a metodologia MARS (Friedman & Silverman, 1989), extensão das árvores de regressão utilizando produtos tensores de bases de potências truncadas para aproximação de funções contínuas, para o contexto Bayesi-

ano utilizando o método de Monte Carlo via cadeias de Markov com saltos reversíveis (RJMCMC) (Green, 1995). Esta proposta passou por diferentes adaptações como em DiMatteo et al. (2001) e Lindstrom (2002), em que o método BASS (Francom et al., 2018, 2019) é uma implementação recente prontamente disponível em R.

O presente trabalho busca aplicar o método BASS e uma metodologia de otimização em *splines* de nós livres utilizando algoritmo genético a dois problemas da área de saúde relacionados ao ajuste de curvas. A escolha pelo método BASS, em relação a outros no contexto Bayesiano, consiste da sua facilidade de implementação pelo pacote **BASS** (Francom & Sansó, 2019) no R.

2 OBJETIVO

A partir da pesquisa sobre o tema e estudos práticos de problemas da área da saúde, são propostos os seguintes objetivos:

- discutir os potenciais e dificuldades da metodologia BASS em problemas de ajuste de curva, apresentando as principais limitações e as funções do pacote convRJMCMC criado com a finalidade de apresentar um diagnóstico de convergência mais confiável para o método BASS e algoritmos RJMCMC em geral;

- fornecer novas informações relacionadas aos problemas de saúde estudados, como definir as regiões de morte celular, com e sem dano celular no protocolo de crioablação apresentado e o período ótimo para realização do exame UBT.

3 REVISÃO DE LITERATURA

Um dos objetivos mais elementares da modelagem estatística, apresentado de forma genérica por Friedman (1991), é descrever a relação entre uma variável dependente Y e um vetor de p variáveis independentes $\mathbf{X}$, para a qual se supõem:

$$Y = f^*(\mathbf{X}) + \boldsymbol{\epsilon}^*, \tag{1}$$

por uma função $f(.)$ que se aproxima à função $f^*(.)$ desconhecida, em que $\boldsymbol{\epsilon}^*$ é um componente estocástico aditivo que representa, geralmente, a dependência de Y a outras variáveis além de $\mathbf{X}$ que não são observadas ou controladas (Friedman, 1991).

Por razões de simplicidade, o modelo será restrito ao seguinte contexto:

$$Y = f(\mathbf{X}) + \boldsymbol{\epsilon}, \tag{2}$$

em que $\boldsymbol{\epsilon}$ é um componente estocástico aditivo que também considera o erro representado pela distância entre $f(.)$ e $f^*(.)$, ou seja, a diferença entre a função aproximada e a real (Friedman, 1991).

Os métodos aqui detalhados são de interesse para situações em que se deseja aproximar relações não-lineares por processos de expansões de base *splines*. As duas metodologias apresentadas no trabalho são propostas distintas de *splines* de nós livres, uma utilizando algoritmo genético e outra o método de Monte Carlo via cadeias de Markov com saltos reversíveis (RJMCMC), denominada metodologia BASS, uma metodologia Bayesiana que é o foco principal do trabalho.

Assim, essa revisão, será dividida em três partes. Uma dedicada especialmente à discussão da aproximação por funções *splines* dentro do contexto de

modelos lineares, destacando os conceitos fundamentais e algumas bases de funções *splines*, buscando a apresentação geral do tema e fornecendo os subsídios teóricos necessários para o encerramento da primeira parte com o tema das *splines* de nós livres. Na segunda, serão apresentados os métodos RJMCMC e o BASS, detalhando algumas das diferenças principais entre o BASS que utiliza o algoritmo RJMCMC com as metodologias de otimização utilizadas para *splines* de nós livres.

Visto o método BASS depender do RJMCMC, será necessário apresentar critérios apropriados para a análise de convergência, discutidos na terceira parte da revisão de literatura junto a alguns critérios de métodos MCMC tradicionais utilizados no presente trabalho.

3.1 Do Modelo de Regressão Linear às Splines de Nós Livres

O modelo linear para a relação entre um vetor de observações $\mathbf{Y}_{n\times 1}$ de uma variável dependente Y e a respectiva matriz de dados das p variáveis independentes $\mathbf{X}$, com uma coluna de 1's representando o intercepto, $\mathbf{X}_{n\times(p+1)}$, é dado por:

$$\mathbf{Y}_{n\times 1} = \mathbf{X}_{n\times(p+1)}\boldsymbol{\beta}_{p+1\times 1} + \boldsymbol{\epsilon}_{n\times 1} \tag{3}$$

em que $\boldsymbol{\beta}_{(p+1)\times 1}$ é um vetor coluna com $p+1$ coeficientes. Temo-se as seguintes suposições para o modelo linear múltiplo (Graybill, 1961; Montgomery et al., 2012; Faraway, 2014; Hastie et al., 2017):

1. Erros tem distribuição normal com média 0 e variância constante (homocedástica) e não são correlacionados.

2. A relação entre a variável dependente Y e o vetor das variáveis independentes $\mathbf{X}$ é linear.

   O modelo de regressão linear simples é um caso particular do modelo linear múltiplo dado pela relação:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \tag{4}$$

que representa uma reta, em que β_0 é o intercepto e β_1 sua inclinação, y_i é a i-ésima observação da variável dependente Y e x_i é a i-ésima observação da única variável independente X, para $i = 1, \ldots, n$.

Quando a suposição de linearidade é violada, uma alternativa é transformar o vetor de variáveis independentes originais $\mathbf{X}$, adicionando ou trocando por novas variáveis que são produto de transformações das originais. Este procedimento é denominado de expansão de base, forma usualmente utilizada para implementar os métodos de regressão baseados em funções *spline*. Para a devida conceituação, o processo de expansão de base linear será definido apropriadamente nas seções seguintes, junto aos principais métodos de regressão por funções *spline*.

3.1.1 Expansões de Base

As diferentes formas de funções *spline* são convenientemente expressas como expansões de base, recurso extensamente utilizado por ferramentas computacionais. Portanto, para fins de maior clareza, serão apresentadas as definições formais das funções *spline* (quando aplicável) e por suas expansões de base lineares.

Definição 1. Expansão de Base Linear (Hastie et al., 2017, págs. 139 e 140) Seja $\mathbf{X}$ um vetor com p variáveis e $h_l(\mathbf{X}) : \mathbb{R}^p \to \mathbb{R}$ a l-ésima transformação de $\mathbf{X}$, para $l = 1, \ldots, L$. Se

$$f(\mathbf{X}) = \sum_{l=1}^{L} \beta_l h_l(\mathbf{X})$$

então $f(\mathbf{X})$ é uma expansão de base linear em $\mathbf{X}$ e β_l é o coeficiente associado a transformação l.

Alguns exemplos de expansões de base são:

- $h_l(\mathbf{X}) = 1$, que representa o intercepto do modelo;

- $h_l(\mathbf{X}) = \log(X_j), \sqrt{X_j}, \ldots,$ para $j = 1, \ldots, p$, transformações não lineares aplicadas a um conjunto das variáveis de $\mathbf{X}$ ou a todas variáveis originais, como em $h_l(\mathbf{X}) = ||\mathbf{X}||$, tal que $||\mathbf{X}|| = \sqrt{X_1^2 + X_2^2 + \cdots + X_p^2}$ é a norma de $\mathbf{X}$;

- $h_l(\mathbf{X}) = I_{(a_k,b_k)}(X_k)$, para $k = 1, \ldots, p$, funções indicadoras de que valores de determinada variável X_k de $\mathbf{X}$ pertencem a um determinado intervalo (a_k, b_k).

A definição 1 apresenta uma forma geral, para o caso multivariado, de expansão de base. A partir daqui, nessa parte da revisão de literatura, será considerado o caso univariado, de forma que a l-ésima transformação da variável X será denotada por $h_l(X) : \mathbb{R} \to \mathbb{R}$, formando expansões de bases criando um conjunto de novas variáveis independentes a partir de uma única variável independente original. Um exemplo é o polinômio cúbico que pode ser expresso pelas bases:

$$h_1(X) = 1,\ h_2(X) = X,\ h_3(X) = X^2 \ e\ h_4(X) = X^3,$$

tal que nosso modelo é escrito na forma

$$f(X) = \sum_{l=1}^{4} \beta_l h_l(X) = \beta_1 + \beta_2 X + \beta_3 X^2 + \beta_4 X^3. \tag{5}$$

3.1.2 Dos Polinômios às *Splines*

Para fins de devidamente uniformizar a notação a respeito das funções *spline*, essenciais para elaboração dos modelos de regressão *spline*, serão apresentadas as definições de espaço de polinômios, polinômios segmentados e funções *splines* polinomiais de ordem q.

Definição 2. Espaço dos polinômios de ordem q (Schumaker, 2007, pág. 3)
O conjunto $\mathcal{P}_q$ definido pela expressão

$$\mathcal{P}_q = \left\{ p(X) : p(X) = \sum_{i=1}^{q} c_i X^{i-1} \right\} \tag{6}$$

é denominado espaço dos polinômios $p(X)$ de ordem q, em que $c_1, \ldots, c_q$ e $X \in \mathbb{R}$. De forma que, se $p(X) \in \mathcal{P}_q$, então $p(X)$ é um polinômio de ordem q (grau $q-1$).

Observe que a expressão utilizada na equação (6) para representar o polinômio $p(X)$ é uma série de potências. Por exemplo, o polinômio cúbico, de grau 3 ou ordem 4, pode ser expresso como série de potências por:

$$p(X) = \sum_{i=1}^{4} c_i X^{i-1} = c_1 + c_2 X + c_3 X^2 + c_4 X^3, \tag{7}$$

expressão de mesma forma que a obtida na equação (5).

A definição 2 estabelece uma expressão geral para o conjunto de polinômios de ordem q. Visto a estrutura simples, polinômios são extensamente usados para aproximação de funções. Uma grande dificuldade, nesta abordagem, é quando tal aproximação exige polinômios de grau muito elevado. Para redução do grau de um polinômio $p(X)$, uma alternativa é segmentar o intervalo de variação de X de modo a definir, em cada subintervalo, um polinômio distinto de grau reduzido, resultando em um polinômio segmentado (Bojanov et al., 1993).

Definição 3. Espaço dos polinômios segmentados de ordem q (Schumaker, 2007, pág. 4)
Sejam $[a, b] \subset \mathbb{R}$ um intervalo fechado e $\Delta = \{\xi_i\}_0^{K+1}$ um conjunto, tal que $a = \xi_0 < \xi_1 < \ldots < \xi_K < \xi_{K+1} = b$, de forma que Δ particiona o intervalo $[a, b]$ em $K+1$ subintervalos $I_i = [\xi_i, \xi_{i+1})$, para $i = 0, 1, \ldots, K$. Dessa forma, ξ_i é o i-ésimo nó do polinômio segmentado, sendo ξ_0 e ξ_{K+1} denominados nós exteriores e ξ_i, para $i = 1, \ldots, K$, nós interiores.
Dado um inteiro positivo q, existem os polinômios $p_i \in \mathcal{P}_q$ (definição 2), tal que

$$\mathcal{PP}_q(\Delta) = \{f : f(X) = p_i \,\text{para}\, X \in I_i,\, i = 0, 1, \ldots, K.\}$$

é denominado de espaço dos polinômios segmentados de ordem q com nós em $\xi_1, \ldots, \xi_K$. Como visto, o conjunto Δ é uma partição do intervalo $[a, b]$ e f é uma

função expressa por um conjunto de $K + 1$ polinômios definidos para cada um dos subintervalos da partição Δ.

Os polinômios segmentados, na forma da definição 3, embora consigam aproximar adequadamente inúmeras funções, não apresentam certas propriedades desejáveis como a continuidade do polinômio e de suas derivadas até determinada ordem. Funções *splines* adicionam restrições aos polinômios segmentados, forçando sua continuidade, para um polinômio segmentado de ordem q, até a derivada de ordem $q - 2$ (Hastie et al., 2017).

Definição 4. Funções de classe $\mathcal{C}^m$ (Lima, 2009, pág. 103)
Seja I um intervalo. Dizemos que, dado $X \in I$, $f(X) : I \to \mathbb{R}$ é uma função de classe $\mathcal{C}^m$, ou $f(X) \in \mathcal{C}^m$, se $f(X)$ é derivável m vezes e a função $f(X)^{(m)} : I \to \mathbb{R}$, m-ésima derivada de $f(X)$, é contínua no intervalo I.

Se $f(X) \in \mathcal{C}^m$, a derivada até ordem m de $f(X)$ será contínua, visto a necessidade da continuidade das funções expressas pelas derivadas de ordem inferiores a m para existência da m-ésima derivada. Temos, assim, a seguinte definição formal de espaço das funções *spline* polinomiais de ordem q:

Definição 5. Espaço das funções *spline* polinomiais de ordem q (Bojanov et al., 1993; Shikin & Plis, 1995; Schumaker, 2007; Micula & Micula, 2012, págs. 19 e 20; págs. 11 e 12; pág. 5; págs. 3 e 4)
Seja Δ uma partição do intervalo $[a, b] \subset \mathbb{R}$ (definição 3) e q um inteiro positivo. O conjunto

$$\mathcal{S}_q(\Delta) = \mathcal{PP}_q(\Delta) \cap \mathcal{C}^{q-2}[a, b]$$

é denominado espaço das funções *spline* polinomiais de ordem q com nós em $\xi_1, \ldots, \xi_K$. Dessa forma, se $f(X) \in S_q(\Delta)$, $f(X)$ é uma função *spline* polinomial de ordem q com nós interiores em $\xi_1, \ldots, \xi_K$, ou seja, uma função polinomial segmentada de ordem q com as derivadas até ordem $q - 2$ contínuas.

Com o objetivo de ilustrar as definições acima, a função $f(x)$ na equação (8) é um exemplo de *spline* de ordem 4 (cúbica):

$$f(x) = \begin{cases} p_1 &= 1 - x + x^2 - 0{,}1x^3, \quad 0 \leq x < 1; \\ p_2 &= 1{,}8 - 3{,}4x + 3{,}4x^2 - 0{,}9x^3, \quad 1 \leq x < 2; \\ p_3 &= -15{,}8 + 23x - 9{,}8x^2 + 1{,}3x^3, \quad 2 \leq x < 3; \\ p_4 &= 22 - 14{,}8x + 2{,}8x^2 - 0{,}1x^3, \quad 3 \leq x \leq 4. \end{cases} \tag{8}$$

em que $\xi_0 = a = 0$, $\xi_1 = 1$, $\xi_2 = 2$, $\xi_3 = 3$ e $\xi_4 = b = 4$. As restrições de continuidade das derivadas de primeira e segunda ordem, $f^{(1)}(x)$ e $f^{(2)}(x)$, respectivamente, podem ser verificadas pelas expressões:

$$f^{(1)}(x) = \begin{cases} -1 + 2x - 0{,}3x^2, \quad 0 \leq x < 1; \\ -3{,}4 + 6{,}8x - 2{,}7x^2, \quad 1 \leq x < 2; \\ 23 - 19{,}6x + 3{,}9x^2, \quad 2 \leq x < 3; \\ -14{,}8 + 5{,}6x - 0{,}3x^2, \quad 3 \leq x \leq 4. \end{cases} \tag{9}$$

e

$$f^{(2)}(x) = \begin{cases} 2 - 0{,}6x, \quad 0 \leq x < 1; \\ 6{,}8 - 5{,}4x, \quad 1 \leq x < 2; \\ -19{,}6 + 7{,}8x, \quad 2 \leq x < 3; \\ 5{,}6 - 0{,}6x, \quad 3 \leq x \leq 4. \end{cases} \tag{10}$$

A Tabela 1 ilustra que a função na equação (8), polinômio segmentado de ordem 4, é contínua até a derivada de segunda ordem e é, portanto, uma função *spline* de ordem 4.

O Teorema 1 permite expressar as funções *spline* como expansões de base lineares:

Teorema 1. A função $f(X) \in \mathcal{S}_q(\Delta)$ se e somente se $f(X)$ puder ser escrita na forma

Tabela 1. Avaliação de continuidade da função *spline* cúbica da equação (8) até derivada de segunda ordem.

	Limites		
$a =$	1	2	3
$\lim_{x\to a-} f(x)$	0,9	1,4	0,1
$\lim_{x\to a+} f(x)$	0,9	1,4	0,1
$\lim_{x\to a-} f^{(1)}(x)$	0,7	−0,6	−0,7
$\lim_{x\to a+} f^{(1)}(x)$	0,7	−0,6	−0,7
$\lim_{x\to a-} f^{(2)}(x)$	1,4	−4,0	3,8
$\lim_{x\to a+} f^{(2)}(x)$	1,4	−4,0	3,8

$$f(X) = \sum_{j=1}^{q} a_j X^{j-1} + \sum_{i=1}^{K} c_i [X - \xi_i]_+^{q-1}$$

em que a_j e $c_i \in \mathbb{R}$; $j = 1, \ldots, q$ e $i = 1, \ldots, K$; e $_+$ denota parte positiva, dada por $[\eta]_+ = max(\eta, 0)$ (Bojanov et al., 1993, pág. 20).

A forma apresentada no Teorema 1 é conhecida como de potências truncadas, ou seja, toda função *spline* pode ser escrita na forma de potências truncadas (Bojanov et al., 1993), retomando a Definição 1, tem-se as seguintes bases para as funções *spline*, denominadas bases de potências truncadas (Hastie et al., 2017):

$$\begin{aligned} h_j(X) &= X^{j-1},\ j = 1, \ldots, q, \\ h_{q+i}(X) &= [X - \xi_i]_+^{q-1},\ i = 1, \ldots, K. \end{aligned} \tag{11}$$

É comum o uso de *splines* cúbicas (cúbica é referente ao grau do polinômio), ou seja, polinômios segmentados de ordem $q = 4$ com derivadas contínuas até ordem 2, cujas bases são dadas por Hastie et al. (2017):

$$\begin{aligned} h_j(X) &= X^{j-1},\ j = 1, \ldots, 4, \\ h_{4+i}(X) &= [X - \xi_i]_+^3,\ i = 1, \ldots, K. \end{aligned} \tag{12}$$

O ajuste de funções *spline*, visto suas expansões de base lineares, pode ser realizado pelo método de mínimos quadrados (Hastie et al., 2017).

A função na equação (8), por exemplo, é escrita com bases de potências truncadas na forma:

$$f(x) = 1 - x + x^2 - 0{,}1x^3 - 0{,}8\,[x-1]_+^3 + 2{,}2\,[x-2]_+^3 - 1{,}4\,[x-3]_+^3\,, \tag{13}$$

com $x \in [0, 4]$.

Temos, também, as chamadas *B-Splines*, que possuem propriedades que favorecem seu ajuste computacional de forma mais eficiente (Braibant & Fleury, 1984; Johs & Hale, 2008).

Definição 6. ***B-Splines*** (Piegl & Tiller, 1996, pág. 47)
Como proposto na definição 3, sejam $[a, b] \subset \mathbb{R}$ um intervalo fechado e $\Delta = \{\xi_i\}_0^{K+1}$ uma partição do intervalo $[a, b]$, tal que $a = \xi_0 < \xi_1 < \ldots < \xi_K < \xi_{K+1} = b$ e $I_i = [\xi_i, \xi_{i+1})$ é o i-ésimo subintervalo de $[a, b]$, para $i = 0, 1, \ldots, K-1$. Considere os nós exteriores ξ_0 e ξ_{K+1}, tal que se obtém a sequência aumentada de $K + 2Q$ nós:

- $\gamma_1 \leq \ldots \leq \gamma_Q \leq \xi_0$;
- $\gamma_{j+Q} = \xi_j$, para $j = 1, \ldots, K$;
- $\xi_{K+1} \leq \gamma_{K+Q+1} \leq \ldots \leq \gamma_{K+2Q}$.

com $Q \geq q$, em que q é a ordem da *spline*.
Denota-se a i-ésima base de uma *B-Spline* de ordem q por $\mathfrak{B}_{i,q}(X)$, $i = 1, \ldots, K + 2Q - q$, determinadas recursivamente, tal que, para $q > 1$:

$$\mathfrak{B}_{i,q}(X) = \frac{X - \gamma_i}{\gamma_{i+q-1} - \gamma_i}\mathfrak{B}_{i,q-1}(X) + \frac{\gamma_{i+q} - X}{\gamma_{i+q} - \gamma_{i+1}}\mathfrak{B}_{i+1,q-1}(X),\ i = 1, \ldots, K + 2Q - q$$

e para $q = 1$:

$$\mathfrak{B}_{i,1}(X) = I_{[\gamma_i, \gamma_{i+1})}(X)$$

é uma função indicadora para $X \in [\gamma_i, \gamma_{i+1})$, $i = 1, \ldots, K + 2Q - 1$.
Se $f(X)$ pode ser escrita por

$$f(X) = \sum_{i=1}^{K+2Q-q} c_i \mathfrak{B}_{i,q}(X), \tag{14}$$

temos que $f(X)$ é uma *B-Spline* de ordem q com nós em γ_s, para $s = 1, \ldots, K + 2Q$.

As *B-Splines* são tidas como generalizações das chamadas curvas de Bézier e tem a vantagem, em relação às bases de potências trucadas, de serem definidas por um conjunto de funções com suporte compacto, além da maior eficiência computacional (Braibant & Fleury, 1984; Johs & Hale, 2008). A propriedade do suporte compacto é notável na definição recursiva de *B-Splines*, visto todas as bases serem definidas considerando funções indicadoras relacionadas a um subintervalo determinado pela sequência aumentada de nós, permitindo alterações locais na função (Johs & Hale, 2008).

É válido notar que qualquer função em base de potências truncadas pode ser escrita na forma de uma *B-Spline* e, no caso, a mudança de coordenada entre as bases pode ser feita utilizando matrizes de mudança de base. No R, é possível gerar as matrizes de mudança de base com a função `chgbasismat` do pacote `freeknotsplines` (Spiriti et al., 2018) que deve ser compilada diretamente do código fonte (o pacote faz uso interno dela, não a liberando para uso direto pelo usuário). De forma genérica uma matriz de mudança de base é usada basicamente para, partindo de um ponto com coordenadas de dimensão p definidas em uma base B, encontrar as coordenadas desse mesmo ponto na base C a partir de um simples produto entre a matriz de mudança de base e o vetor com as coordenadas do ponto na base B. Além

disso, temos que a inversa da matriz de mudança da base B para C é a matriz de mudança da base C para B. Por exemplo, no caso da equação (8), utilizando a função `chgbasismat` obtemos a seguinte matriz de mudança da base *B-Spline* (denotada por B) para a base de potências truncadas (denotada por C) de $\mathbb{R}^7$:

$$[M]_C^B = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1/3 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 2/3 & 0 & 0 & 0 & 0 \\ 1 & 2 & 11/3 & 6 & 0 & 0 & 0 \\ 1 & 3 & 26/3 & 24 & 6 & 0 & 0 \\ 1 & 11/3 & 40/3 & 48 & 18 & 4 & 0 \\ 1 & 4 & 16 & 64 & 27 & 8 & 1 \end{bmatrix} \tag{15}$$

Sabendo que as coordenadas em relação a base de potências truncadas C é dada por:

$$[v]_C = \begin{bmatrix} 1 \\ -1 \\ 1 \\ -0,1 \\ -0,8 \\ 2,2 \\ -1,4 \end{bmatrix} \tag{16}$$

Logo, a matriz de coordenada em relação a base *B-Spline* B é dada por:

$$[v]_B = [M]_C^B \times [v]_C = \begin{bmatrix} 1 \\ 2/3 \\ 2/3 \\ 31/15 \\ -8/15 \\ 4/15 \\ 1,2 \end{bmatrix}, \tag{17}$$

de forma que a expressão da equação (13), pode ser assim escrita como uma B-Spline:

$$\mathfrak{B}_{1,4}(X)+\frac{2}{3}\mathfrak{B}_{2,4}(X)+\frac{2}{3}\mathfrak{B}_{3,4}(X)+\frac{31}{15}\mathfrak{B}_{4,4}(X)-\frac{8}{15}\mathfrak{B}_{5,4}(X)+\frac{4}{15}\mathfrak{B}_{6,4}(X)+1.2\mathfrak{B}_{7,4}(X) \tag{18}$$

para a sequência aumentada de nós $\gamma = \{0,0,0,0,1,2,3,4,4,4,4\}$, sendo 1, 2 e 3 os nós interiores e as repetições de 0 e 4 os nós de controle que se igualam ao mínimo $a = 0$ e ao máximo $b = 4$ da função. O código da função `chgbasismat` para gerar a matriz de mudança de base e a implementação computacional do exemplo apresentado estão no arquivo codigo_matriz_de_mudanca_de_base.R da pasta com os arquivos suplementares cujo acesso está disponibilizado pelo primeiro link do apêndice.

3.1.3 Splines de Nós Livres

Quando os nós são considerados parâmetros livres, estimados utilizando os dados, trata-se do contexto das *splines* de nós livres, o qual engloba um conjunto de metodologias que busca utilizar um pequeno grupo de nós "bem localizados" para o ajuste das *splines*. Neste caso, nota-se a melhora da qualidade da aproximação das *splines* (Jupp, 1978; Lindstrom, 1999; Beliakov, 2004; Molinari et al., 2004; Gálvez & Iglesias, 2011; Gálvez et al., 2015). A seguir é apresentada a definição formal de espaço das funções *splines* polinomiais de ordem q e K nós:

Definição 7. Espaço das *Splines* Polinomiais de Ordem q com K nós (adaptado de Schumaker (2007), página 219)

Seja $\mathcal{S}_q(\Delta)$ uma função *spline* de ordem q com K nós definida a partir de uma dada partição Δ do intervalo $[a, b]$. Temos que

$$\mathcal{S}_{q,K} = \bigcup_{\Delta} \mathcal{S}_q(\Delta) \tag{19}$$

é o espaço das funções *splines* polinomiais de ordem q com K nós, para todas as partições Δ possíveis, em que $\bigcup$ representa a operação de união de conjuntos.

Temos, portanto, que, se $f(X) \in \mathcal{S}_{q,K}$, $f(X)$ é uma *spline* polinomial de ordem q com K nós. Dado que $f(X)$ é uma *spline*, a mesma pode ser escrita por:

$$f(X) = \sum_{l=1}^{L} c_l \delta_l(X) \tag{20}$$

em que $c_l \in \mathbb{R}$ é uma constante, $\delta_l(X)$ é a l-ésima função de base para a *spline* com K nós e ordem q, tal que $X \in [a, b]$ (Meinardus et al., 1989). Os coeficientes c_l podem ser encontrados por mínimos quadrados quando definidas as bases $\delta_l(X)$, para $l = 1, \ldots, L$. Temos que, com os nós sendo parâmetros livres, deve-se encontrar as posições ótimas para os nós e, consequentemente, as bases da *spline* e, por fim, os coeficientes c_l. Os meios para se encontrar essas posições podem ser divididos, na sua maioria, quanto a três formas possíveis (Molinari et al., 2004):

1. Defini-se, inicialmente, um conjunto com vários nós (geralmente equi-espaçados). A cada passo do algoritmo, um único nó é retirado. Encerra-se o algoritmo quando for obtido o número K desejado de nós.

2. Partindo de um conjunto sem nenhum nó, a cada passo do algoritmo, é inserido um nó em uma posição determinada conforme critério previamente especificado. Encerra-se o algoritmo quando for obtido o modelo com o número K desejado de nós.

3. Dado um número K inicial de nós, modificam-se as localizações dos nós até encontrar as posições ótimas.

Os procedimentos estabelecidos acima podem ser estendidos aos modelos de mínimos quadrados penalizados (Lindstrom, 1999) e repetidos, ou analisados, para diferentes números K de nós de modo a se definir, também, o número ótimo de nós. Caso tenha o interesse em se definir o número k ótimo de nós deve ser usado critério de desempenho, que vai definir quais nós serão retirados, adicionados ou modificados, que penalizem a dimensão do modelo, como AIC e BIC. Caso o número de nós k seja fixo, não há necessidade de utilizar um critério que penalize a dimensão do modelo.

Uma forma de ajuste implementado no programa R está disponível no pacote **freeknotsplines** (Spiriti et al., 2018), que utiliza algoritmo genético para encontrar a posição dos nós. Neste algoritmo, para um número de nós fixos, são definidos, aleatoriamente, diversos conjuntos de nós, pertencentes ao intervalo $[a, b]$ de variação de X, e esses são usados para definir bases de funções *B-Spline*. O desempenho do modelo com essas bases é avaliado, por padrão, pelo critério da validação cruzada generalizada (GCV). As bases com os menores valores de GCV são mantidas (bases pai) e passam pelos processos de *crossover* e mutação. No *crossover*, são escolhidas duas bases pai aleatoriamente e é tomado um valor de d do conjunto $\{1, \ldots, K\}$ de índices dos nós, de forma que os $d-1$ primeiros nós são os $d-1$ primeiros nós do primeiro pai e o restante dos nós vem do segundo pai nas respectivas posições, sendo o processo repetido até o número de filhos originados por *crossover* igualar-se ao número de pais anteriormente retidos pelo algoritmo. Após isso, é feito o processo de mutação, em que são selecionadas bases ao acaso, dentre as bases pai e as geradas por *crossover*, nas quais um nó ξ_d qualquer, de cada uma das bases selecionadas, é trocado por outro amostrado uniformemente no intervalo (ξ_{d-1}, ξ_{d+1}). Dessas amostras são selecionadas algumas que tem o menor GCV. Sendo todos esses processos repetidos até se atingir um número limite de gerações ou atingir determinada condição de convergência. Visto tal método não ser foco no trabalho, detalhes do algoritmo e demais considerações podem ser encontradas no artigo de Spiriti et al. (2013).

Na literatura, as abordagens Bayesianas são baseadas no método de Monte Carlo via cadeias de Markov (MCMC) com saltos reversíveis (Denison et al., 1998; DiMatteo et al., 2001; Lindstrom, 2002; Francom et al., 2018, 2019), havendo também uma abordagem fiducial (Sonderegger & Hannig, 2014). Dentre as metodologias Bayesianas a BASS (Francom et al., 2018, 2019), que foi utilizada no trabalho, está disponível no R com o pacote **BASS** (Francom & Sansó, 2019) (mesmo acrônimo da metodologia, para referência a pacotes do R serão utilizados nomes em negrito), e será apresentada a seguir.

3.2 *Bayesian Adaptive Splines Surfaces* (BASS)

3.2.1 Método de Monte Carlo via Cadeias de Markov com Saltos Reversíveis

O método de Monte Carlo via cadeias de Markov com saltos reversíveis representa uma generalização do método de Metropolis-Hastings para situações em que as dimensões do objeto de inferência são desconhecidas, proposto originalmente por Green (1995). Dentre os problemas que Green (1995) buscou resolver com a nova metodologia se destacam:

- seleção de variáveis em modelos de regressão;
- seleção Bayesiana de modelos com diferentes números de parâmetros;
- problemas com múltiplos pontos de mudança;
- segmentação de imagens.

O método assume um conjunto de modelos candidatos contável expresso por $\{\mathcal{M}_k,\, k \in \mathcal{K}\}$, cada $\mathcal{M}_k$ tem um conjunto $\boldsymbol{\theta}_{(k)}$ de parâmetros desconhecidos, de dimensão $\mathbb{R}^{n_k}$, de forma que n_k possa variar entre diferentes modelos, em que $\mathcal{K}$ é o conjunto de todos os índices dos modelos. Dado que os dados y foram observados, a estrutura hierárquica para modelagem da distribuição conjunta do modelo $\left(k, \boldsymbol{\theta}_{(k)}, y\right)$ é dada por:

$$p\left(k, \boldsymbol{\theta}_{(k)}, y\right) = p(k)p\left(\boldsymbol{\theta}_{(k)}|k\right)p\left(y|\boldsymbol{\theta}_{(k)}, k\right) \tag{21}$$

ou seja, o produto da probabilidade do modelo, priori e verossimilhança. A notação acima é abreviada pelo par $\left(\boldsymbol{\theta}_{(k)}, k\right)$ em x, para um dado k, x pertence a $\mathcal{C}_k = \{k\} \times \mathbb{R}^{n_k}$; com x geralmente variando, portanto, em $\mathcal{C} = \bigcup_{k\in\mathcal{K}} \mathcal{C}_k$.

Como no algoritmo tradicional de Metropolis-Hastings, o movimento entre estados $x = \left(\boldsymbol{\theta}_{(k)}, k\right) \in \mathcal{A}$ e $x' = \left(\boldsymbol{\theta}'_{(k')}, k'\right) \in \mathcal{B}$ de dois modelos deve satisfazer, para uma cadeia de Markov com espaço de estados Θ e distribuição estacionária π, a condição de equilíbrio detalhado

$$\int_{(x,x')\in\mathcal{A}\times\mathcal{B}} \pi(dx)P(x, dx') = \int_{(x,x')\in\mathcal{A}\times\mathcal{B}} \pi(dx')P(x', dx) \tag{22}$$

para todos os conjuntos de Borel $\mathcal{A} \times \mathcal{B} \subset \Theta$, em que P é um núcleo (kernel) de transição de Markov (Fan & Sisson, 2011).

Dado que o movimento de x para x' é proposto por uma densidade com forma $q(x, x')$ a condição de equilíbrio detalhado na equação (22) pode ser garantida propondo uma probabilidade de aceitação $\alpha(x, x')$, tal que

$$\int_{(x,x')\in\mathcal{A}\times\mathcal{B}} \pi(x|k)q(x, x')\alpha(x, x')dxdx' = \int_{(x,x')\in\mathcal{A}\times\mathcal{B}} \pi(x'|k')q(x', x)\alpha(x', x)dxdx' \tag{23}$$

em que $\pi(x|k)$ e $\pi(x'|k')$ são as distribuições a posteriori em relação ao modelo $\mathcal{M}_k$ e $\mathcal{M}_{k'}$, respectivamente, e

$$\alpha(x, x') = min\left\{1, \frac{\pi(x|k)q(x, x')}{\pi(x'|k')q(x', x)}\right\}, \tag{24}$$

com $\alpha(x', x)$ sendo definido de forma semelhante (Fan & Sisson, 2011). Passos de algoritmos RJMCMC que não dependem de movimentos entre modelos diferentes, como a atualização de parâmetros, podem seguir outros procedimentos de amostragem Bayesiana. Por outro lado, movimentos entre modelos diferentes devem seguir as condições aqui estabelecidas (Fan & Sisson, 2011). Maior detalhamento da forma

geral dos algoritmos RJMCMC, bem como seu uso em diferentes aplicações, pode ser encontrado em Green (1995).

3.2.2 *Bayesian Adaptive Splines Surfaces* (BASS)

A metodologia *Bayesian Adaptive Splines Surfaces* (BASS) consiste em uma estratégia Bayesiana para aproximação de curvas e superfícies multidimensionais utilizando bases *spline* e está implementada no R através do pacote **BASS** (Francom & Sansó, 2019). Esta estratégia é baseada no trabalho de Denison et al. (1998) que é a primeira proposta Bayesiana para ajuste de curvas usando o método RJMCMC inspirada no algoritmo MARS.

No método BASS modela-se y_i, para $i = 1, \ldots, n$, como:

$$y_i = f(\boldsymbol{x}_i^*) + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2); \tag{25}$$

em que

$$f(\boldsymbol{X}^*) = a_0 + \sum_{l=1}^{L} a_l B_l(\boldsymbol{X}^*). \tag{26}$$

Tem-se que y_i e $\boldsymbol{x}_i^*$ são os valores observados de Y e $\mathbf{X}^*$ na i-ésima observação, respectivamente, em que $\boldsymbol{X}^*$ é a padronização de $\mathbf{X}$ no intervalo $[0, 1]$; ϵ_i é o erro do modelo na i-ésima observação e σ^2 é a variância de ϵ_i (Francom & Sansó, 2019). Em que as funções de base B_l são definidas por:

$$B_l(\boldsymbol{X}^*) = \prod_{z=1}^{Z_l} g_{zl}[\psi_{zl}(X^*_{v_{zl}} - \xi_{zl})]_+^{q-1} \tag{27}$$

em que X_v^* é a v-ésima variável do vetor $\boldsymbol{X}^*$, para $v = 1, \ldots, p$, $\psi_{zl} \in \{-1, 1\}$ é denominado como sinal, $\xi_{zl} \in [0, 1]$ é um nó, v_{zl} seleciona uma variável, tal que seja exclusiva em determinada função de base, não se repita, Z_l é o grau de interação (quantas funções de potências truncadas formam a base), $g_{zl} = [(\psi_{zl} + 1)/2 - \psi_{zl}\xi_{zl}]^{-(q-1)}$ é uma constante normalizadora, tal que as funções de base tenham valor máximo igual a 1, $[.]_+$ denota parte positiva e q define a ordem das *splines*.

O objetivo do método BASS é estimar $\boldsymbol{\theta} = \{\sigma^2, L, \mathbf{a}, \mathbf{Z}, \boldsymbol{\psi}, \boldsymbol{\xi}, \mathbf{v}\}$, em que $\mathbf{Z}$ é o vetor de graus de interação de dimensão L, $\boldsymbol{\psi}$ é o vetor de sinais $\left\{\{\psi_{zl}\}_{z=1}^{Z_l}\right\}_{l=1}^{L}$, $\boldsymbol{\xi}$ é o vetor de nós, $\mathbf{v}$ é o vetor que indica quais variáveis são usadas ($\boldsymbol{\xi}$ e $\mathbf{v}$ definidos de maneira similar a $\boldsymbol{\psi}$) e $\mathbf{a} = \{a_0, a_1, \ldots, a_L\}$ é o vetor de coeficientes do modelo. Seja $\mathbf{B}$ a matriz $n \times (L+1)$ de funções de base (incluindo o intercepto), em que T indica matriz transposta, tem-se as prioris:

$$\begin{aligned} \mathbf{a}|\sigma^2, \tau, \mathbf{B} &\sim N\left(\mathbf{0}, \sigma^2 \left(\mathbf{B}^T\mathbf{B}\right)^{-1}/\tau\right) && (28) \\ \sigma^2 &\sim InvGamma(g_1, g_2) && (29) \\ \tau &\sim Gamma(a_\tau, b_\tau) && (30) \end{aligned}$$

em que $a_\tau = 1$ e $b_\tau = \frac{1}{n}$ (parâmetros de forma e taxa), em que n é o tamanho amostral, para centralizar a distribuição sobre a priori informativa unitária e $g_1 = g_2 = 0$ resultando em uma priori não informativa $p(\sigma^2) \propto 1/\sigma^2$. É usada uma priori Poisson com parâmetro λ truncada entre 0 e L_{max} para o número de funções de base L. Por sua vez, a taxa λ da Poisson tem uma hiper-priori Gamma, em que c é a massa de uma Poisson truncada, $c = 1 - \sum_{l=0}^{L_{max}} e^{-\lambda}\lambda^l/l!$, ou seja,

$$\begin{aligned} p(L|\lambda) &= \frac{e^{-\lambda}\lambda^L}{cL!} && (31) \\ \lambda &\sim Gamma(h_1, h_2) && (32) \end{aligned}$$

em que $h_1 = h_2 = 10$ (forma e taxa) geralmente induzindo um pequeno número de funções de base (Francom & Sansó, 2019).

As distribuições de $\mathbf{Z}$, $\boldsymbol{\psi}$, $\boldsymbol{\xi}$ e $\mathbf{v}$ são distribuições uniformes sobre um espaço restrito ao número mínimo de pontos não nulos b nos vetores de uma base selecionada (Francom et al., 2018), que por padrão é estabelecido como $min(20, 0.1 \times n)$ no pacote **BASS** (Francom & Sansó, 2019), em que n é o número de observações no conjunto de dados. No caso, essa restrição, faz com que marginalmente em relação às dimensões dos nossos preditores não seja possível ter menos de b pontos em cada

uma das partições criadas pelas funções de base. As prioris para tais quantidades dependem da função de base em questão, o subscrito l indica a função de base. De modo que a priori para Z_l, para cada base l, é dada por uma distribuição uniforme discreta:

$$P(Z_l|L) \sim Unif\{1, \ldots, Z_{max}\},\ l = 1, \ldots, L. \tag{33}$$

em que Z_{max} é o grau máximo de interação possível na l-ésima função de base.

A priori para os sinais, variáveis e nós corresponde a uma distribuição uniforme discreta que satisfaz a restrição proposta, denotada por

$$P(\psi_l, \mathbf{v}_l, \xi_l|Z_l, L, \mathbf{X}^*) = \begin{cases} c_{Z_l} & \text{se} \quad b_l \geq b, \\ 0 & \text{caso contrário,} \end{cases} \tag{34}$$

em que b_l é o número de valores não nulos presente nos vetores da base. O valor c_{Z_l} é o recíproco do número de funções de base com grau de interação Z_l possíveis, valor que depende de $\mathbf{X}^*$, dado pela expressão (Francom et al., 2019, arquivo suplementar):

$$c_{Z_l} = \left(\frac{1}{2}\right)^{Z_l} \left(\prod_{z=1}^{Z_l} \frac{1}{n_{v_{zl}}}\right) \binom{p}{Z_l}^{-1} \left(\frac{1}{Z_{max}}\right) \tag{35}$$

em que p é o número de variáveis do vetor $\mathbf{X}^*$ (dimensão de $\mathbf{v}$) e $n_{v_{zl}}$ o número de observações únicas de X_v^* candidatas a serem definidas como um nó.

Basicamente, isso indica que os sinais, variáveis e nós que definem as bases serão amostradas de uma uniforme discreta dentre todas as bases possíveis, ou seja, definindo-se o valor de Z_l as probabilidades de selecionar uma base dentre todas as possíveis é equiprovável, salvo a restrição do número de vetores nulos na base, tal que pode ser obtida a seguinte distribuição conjunta (Francom et al., 2018):

$$\begin{aligned} P(\boldsymbol{\theta}, \tau) = P(L, \sigma^2, \mathbf{a}, \mathbf{Z}, \boldsymbol{\psi}, \mathbf{v}, \boldsymbol{\xi}, \lambda, \tau|\mathbf{X}^*) \quad &= \quad P(\lambda)P(\tau|\mathbf{X}^*)P(\sigma^2)P(L|\lambda)P(\mathbf{a}|L, \sigma^2, \mathbf{B}, \tau) \times \\ &\prod_{l=1}^{L} P(Z_l|L)P(\psi_l, v_l, \xi_l|Z_l, L, \mathbf{X}^*). \end{aligned} \tag{36}$$

E, também, tem-se o seguinte núcleo da distribuição a posteriori (Francom et al., 2019, arquivo suplementar):

$$\begin{aligned}\pi(L,\lambda,\mathbf{a},\sigma^2,\tau,\mathbf{Z},\boldsymbol{\psi},\boldsymbol{\xi},\mathbf{v}|\mathbf{y}) \quad \propto \quad & N(\mathbf{y}|\mathbf{Ba},\sigma^2\mathbf{I})N\left(\mathbf{a}|\mathbf{0},\frac{\sigma^2}{\tau}(\mathbf{B}^T\mathbf{B})^{-1}\right)\times \\ & Pois(L|\lambda)Gamma(\lambda|h_1,h_2)L!Gamma(\tau|b_1,b_2)\times \\ & \prod_{l=1}^{L}\left(\frac{1}{2}\right)^{Z_l}\left(\prod_{z=1}^{Z_l}\frac{1}{n_{v_{zl}}}\right)\binom{p}{Z_l}^{-1}\left(\frac{1}{Z_{max}}\right)I(b_l>b),\end{aligned} \tag{37}$$

em que $L!$ representa todas as formas que você pode obter as L funções de base (ordenação).

No algoritmo BASS são propostos 3 movimentos no RJMCMC (Denison et al., 1998; Nott et al., 2005; Francom et al., 2018):

(a) Nascimento: adição de uma nova função de base com probabilidade $P_{nascimento}$;

(b) Morte: remoção de uma função de base com probabilidade P_{morte};

(c) Mudança: Mudança de uma função de base já existente com probabilidade $P_{mudança}$.

A ideia básica do algoritmo empregado consiste em:

1. Inicializar $\boldsymbol{\theta}_{t=0}$.

2. Propor um passo de nascimento, morte ou mudança com probabilidades $P_{nascimento}$, P_{morte} e $P_{mudança}$, respectivamente. Gerar os valores propostos para $\mathbf{Z}_{t+1}$, $\boldsymbol{\psi}_{t+1}$, $\mathbf{v}_{t+1}$ e $\mathbf{t}_{t+1}$.

3. Após gerar os valores propostos em (2), calcular a probabilidade de aceitação de Metropolis-Hastings e determinar se os valores propostos serão aceitos ou os antigos mantidos.

4. Amostrar τ_{t+1}, $\boldsymbol{\sigma}^2_{t+1}$ e a_{t+1} das distribuições condicionais completas (amostrador de Gibbs).

5. Retornar ao passo (2) para $t = t + 1$ até $t \geq t_{max}$, em que t_{max} é o número de iterações da cadeia.

Informações específicas do algoritmo implementado constam no arquivo suplementar de Francom et al. (2019). As principais informações estão expostas abaixo.

Para os passos de nascimento temos que $b(L^A \to L^P)$ é a probabilidade proposta de sair do modelo atual para o modelo proposto (com adição de uma função de base). Dado que O é uma função de densidade de probabilidade discreta que relaciona as probabilidades de incluir um grau de interação de acordo com a frequência que esses graus de interação já ocorrem no modelo e φ é uma função similar, mas para as variáveis atualmente usadas no modelo:

$$b(L^A \to L^P) = P_{nascimento} O(Z_{L+1}|L^A)\varphi(\mathbf{v}_{L+1}|L^A)\left(\frac{1}{2}\right)^{Z_{L+1}} \prod_{z=1}^{Z_{L+1}} \frac{1}{n_{v_{zL+1}}} \tag{38}$$

a probabilidade de aceitar o passo é o máximo entre 1 e

$$\alpha_{nascimento} = \frac{[L+1, \mathbf{Z}^*, \boldsymbol{\psi}^*, \boldsymbol{\xi}^*, \mathbf{v}^*|\mathbf{y}, \sigma^2, \tau, \lambda]\, b(L^P \to L^A)}{[L, \mathbf{Z}, \boldsymbol{\psi}, \boldsymbol{\xi}, \mathbf{v}|\mathbf{y}, \sigma^2, \tau, \lambda]\, b(L^A \to L^P)} \tag{39}$$

tem-se que os asteriscos $*$ em $\mathbf{Z}$, $\boldsymbol{\psi}$, $\boldsymbol{\xi}$, $\mathbf{v}$ e em $\mathbf{a}$ e $\mathbf{B}$ indicam os parâmetros, ou a matriz de vetores da base no caso de $\mathbf{B}$, do novo modelo proposto; e que $b(L^P \to L^A) = P_{morte}\frac{1}{L+1}$, dado que é escolhida uma base ao acaso para ser removida, e que a distribuição condicional tem forma:

$$
\begin{aligned}
\left[L, \mathbf{Z}, \boldsymbol{\psi}, \boldsymbol{\xi}, \mathbf{v} | \mathbf{y}, \sigma^2, \tau, \lambda\right] \quad \propto \quad & \left(\frac{\tau}{1+\tau}\right)^{(L+1)/2} \exp\left\{\frac{-1}{2\sigma^2}\left[\mathbf{y}^T\mathbf{y} - \frac{1}{1+\tau}\hat{\mathbf{a}}^T\mathbf{B}^T\mathbf{y}\right]\right\} \times \\
& Pois(L|\lambda) L! \prod_{l=1}^{L}\left(\frac{1}{2}\right)^{Z_l}\left(\prod_{z=1}^{Z_l}\frac{1}{n_{v_{zl}}}\right)\binom{p}{Z_l}^{-1}\left(\frac{1}{Z_{max}}\right) \times \\
& I(b_l > b) \\
\propto \quad & \left(\frac{\tau}{\tau+1}\right)^{(L+1)/2} \exp\left\{\frac{-1}{2\sigma^2}\left[\frac{1}{1+\tau}\hat{\mathbf{a}}^T\mathbf{B}^T\mathbf{y}\right]\right\} \lambda^L \times \\
& \prod_{l=1}^{L}\left(\frac{1}{2}\right)^{Z_l}\left(\prod_{z=1}^{Z_l}\frac{1}{n_{v_{zl}}}\right)\binom{p}{Z_l}^{-1}\left(\frac{1}{Z_{Max}}\right) I(b_l > b). \quad (40)
\end{aligned}
$$

Com $\mathbf{B}^*$ sendo o candidato a conjunto de funções de base e $\hat{\mathbf{a}}^*$ os coeficientes (pesos) obtidos por mínimos quadrados, tem-se a razão,

$$
\begin{aligned}
\frac{\left[L+1, \mathbf{Z}^*, \boldsymbol{\psi}^*, \boldsymbol{\xi}^*, \mathbf{v}^* | \mathbf{y}, \sigma^2, \tau, \lambda\right]}{\left[L, \mathbf{Z}, \boldsymbol{\psi}, \boldsymbol{\xi}, \mathbf{v} | \mathbf{y}, \sigma^2, \tau, \lambda\right]} \quad = \quad & \left(\frac{\tau}{1+\tau}\right)^{1/2} \exp\left\{\frac{\left[\hat{\mathbf{a}}^*\mathbf{B}^{*T}\mathbf{y} - \hat{\mathbf{a}}^T\mathbf{B}^T\mathbf{y}\right]}{2\sigma^2(1+\tau)}\right\} \lambda \times \\
& \left(\frac{1}{2}\right)^{Z_{L+1}}\left(\prod_{z=1}^{Z_{L+1}}\frac{1}{n_{v_{jL+1}}}\right)\binom{p}{Z_{L+1}}^{-1}\left(\frac{1}{Z_{max}}\right) \times \\
& I(b_{L+1} > b). \quad (41)
\end{aligned}
$$

Dessa forma, tem-se que:

$$
\begin{aligned}
\alpha_{nascimento} \quad = \quad & \left(\frac{\tau}{1+\tau}\right)^{1/2} \exp\left\{\frac{1}{2\sigma^2(1+\tau)}[\hat{\mathbf{a}}^*\mathbf{B}^{*T}\mathbf{y} - \hat{\mathbf{a}}^T\mathbf{B}^T\mathbf{y}]\right\} \lambda \times \\
& \binom{p}{Z_{L+1}}^{-1}\left(\frac{1}{Z_{max}}\right) I(b_{L+1} > b) \frac{P_{morte}/(L+1)}{P_{nascimento} O(Z_{L+1}|L^C)\varphi(\mathbf{v}_{L+1}|L^A)} \quad (42)
\end{aligned}
$$

O passo de morte é recíproco ao passo de nascimento:

$$
\begin{aligned}
\alpha_{morte} \quad = \quad & \frac{\left[L-1, \mathbf{Z}^*, \boldsymbol{\psi}^*, \boldsymbol{\xi}^*, \mathbf{v}^* | \mathbf{y}, \sigma^2, \tau, \lambda\right] b(L^A \to L^P)}{\left[L, \mathbf{Z}, \boldsymbol{\psi}, \boldsymbol{\xi}, \mathbf{v} | \mathbf{y}, \sigma^2, \tau, \lambda\right] b(L^P \to L^A)} \\
= \quad & \left(\frac{\tau}{1+\tau}\right)^{-1/2} \exp\left\{\frac{1}{2\sigma^2(1+\tau)}[\hat{\mathbf{a}}^*\mathbf{B}^{*T}\mathbf{y} - \hat{\mathbf{a}}^T\mathbf{B}^T\mathbf{y}]\right\} (1/\lambda) \times \\
& \binom{p}{Z_L} Z_{max} \frac{L\, P_{nascimento} O(Z_L|L^P)\varphi(\mathbf{v}_L|L^P)}{P_{morte}}. \quad (43)
\end{aligned}
$$

Já para o passo de mudança a probabilidade de aceitação é dada por:

$$\begin{aligned}\alpha_{mudança} &= \frac{[L, \mathbf{Z}, \boldsymbol{\psi}^*, \boldsymbol{\xi}^*, \mathbf{v}|\mathbf{y}, \sigma^2, \tau, \lambda]}{[L, \mathbf{Z}, \boldsymbol{\psi}, \boldsymbol{\xi}, \mathbf{v}|\mathbf{y}, \sigma^2, \tau, \lambda]} \\ &= \exp\left\{\frac{1}{2\sigma^2(1+\tau)}[\hat{\mathbf{a}}^*\mathbf{B}^{*T}\mathbf{y} - \hat{\mathbf{a}}^T\mathbf{B}^T\mathbf{y}]\right\}. \end{aligned} \tag{44}$$

As distribuições condicionais, considerando as prioris utilizadas por padrão, tem forma:

$$[\sigma^2|\mathbf{y}, L, \mathbf{Z}, \boldsymbol{\psi}, \boldsymbol{\xi}, \mathbf{v}, \tau, \lambda] \sim InvGamma\left(N/2 + g_1, g_2 + \frac{1}{2}\left[\mathbf{y}^T\mathbf{y} - \frac{1}{1+\tau}\hat{\mathbf{a}}^T\mathbf{B}^T\mathbf{y}\right]\right); \tag{45}$$

$$[\mathbf{a}|.] \sim N\left(\frac{\hat{\mathbf{a}}}{1+\tau}, \frac{\sigma^2}{1+\tau}(\mathbf{B}^T\mathbf{B})^{-1}\right); \tag{46}$$

$$[\tau|.] \sim Gamma\left((L+1)/2 + b_1, b_2 + \frac{1}{2\sigma^2}\mathbf{a}^T\mathbf{B}^T\mathbf{B}\mathbf{a}\right). \tag{47}$$

Apresentado o ajuste de curvas pelo método BASS, a seguir será apresentado o diagnóstico de convergência apropriado para metodologias que envolvem amostradores RJMCMC.

3.3 Diagnóstico de Convergência em Métodos RJMCMC

Como visto na seção anterior, soluções analíticas para a distribuição a posteriori são, para muitos problemas de cunho prático, difíceis ou impossíveis de serem obtidas (Paulino et al., 2018). Com núcleos de transição adequados e quando a distribuição estacionária é alcançada, é possível obter uma amostra da distribuição a posteriori utilizando cadeias de Markov. Na prática, após um número suficientemente grande de iterações, é possível obter uma amostra próxima à posteriori de interesse. A grande dificuldade tratada nos métodos de convergência está em identificar qual é

esse número de iterações, o qual é desconhecido antes de se iniciarem as simulações (BROOKS & ROBERTS, 1998; Gamerman & Lopes, 2006).

Os métodos de avaliação de convergência se dividem em dois grandes grupos (Gamerman & Lopes, 2006):

1. Empíricos, que usam somente os valores de saída amostrados pelas cadeias, com aplicação ampla, embora, na teoria, não garantam a convergência para a verdadeira distribuição.

2. Teóricos, que também consideram propriedades específicas da cadeia de Markov para o problema de inferência em análise, sendo específicos para cada aplicação e geralmente inviáveis de serem encontrados para a maioria dos problemas práticos.

Algoritmos RJMCMC são largamente utilizados para problemas de inferência Bayesiana trans-dimensionais, que tem por característica alternar entre diferentes espaços paramétricos, gerando duas dificuldades na análise de convergência:

1. Identificar parâmetros cuja interpretação se mantenha a mesma em qualquer modelo e espaço paramétrico em consideração (Brooks & Giudici, 2000; Gamerman & Lopes, 2006), os quais nem sempre são encontrados na amostra obtida pela cadeia de Markov. Nesses casos, uma das estatísticas possíveis de serem usadas é o *deviance* (Brooks & Giudici, 2000).

2. São necessários métodos que considerem essa estrutura mais complexa, em que diferentes modelos são considerados, para uma análise de convergência confiável.

Uma alternativa é generalizar métodos empregados no contexto em que as dimensões do modelo são fixas (cis-dimensional) para o contexto em que as dimensões do modelo constantemente se alteram ao longo da cadeia de Markov (trans-dimensional). Como exemplo, tem-se as generalizações de Brooks & Giudici

(2000) e Castelloe & Zimmerman (2002) para os potenciais de redução de escala propostos em Gelman & Rubin (1992).

Na presente revisão, serão apresentados o critério de Castelloe & Zimmerman (2002) e o gráfico de traços para diagnóstico de convergência.

3.3.1 Critério de Castelloe & Zimmerman (2002)

O critério de Castelloe & Zimmerman (2002) é uma modificação do critério de Brooks & Giudici (2000) para análise de convergência de algoritmos RJMCMC. A proposta de Brooks & Giudici (2000) utiliza a análise de variância (ANOVA) de dois fatores (*Two-Way* ANOVA) para computar os potenciais de redução de escala de maneira análoga a de Gelman & Rubin (1992), agora decompostos de maneira a detectar diferenças não só entre cadeias, mas também dentro de um mesmo modelo e entre modelos nas diferentes cadeias de Markov (Brooks & Giudici, 2000; Castelloe & Zimmerman, 2002). Entretanto, a proposição de Brooks & Giudici (2000), considerou um modelo ANOVA balanceado, ignorando que modelos podem não ser visitados com a mesma frequência. A abordagem de Castelloe & Zimmerman (2002) buscou contornar esse problema com uma ANOVA levando em conta o desbalanceamento.

Para apresentar as estatísticas usadas pelos métodos, que dependem das quantidades do Quadro 1 será utilizada a notação proposta por Castelloe & Zimmerman (2002).

As quantidades presentes nas equações (48) a (52) são equivalentes para ambos os métodos:

Quadro 1. Principais quantidades utilizadas para o cálculo dos potenciais de redução de escala e suas descrições.

Quantidade	Descrição
$\boldsymbol{\theta}$	Vetor de parâmetros.
θ	Parâmetro escalar pertencente a $\boldsymbol{\theta}$.
C	Número de cadeias.
T	Tamanho do lote.
M	Número de modelos distintos visitados (por qualquer cadeia).
$\boldsymbol{\theta}^r_{cm}$	Valor de $\boldsymbol{\theta}$ para r-ésima ocorrência do modelo m na cadeia c.
R_{cm}	Número de vezes que o modelo m ocorreu na cadeia c.

Fonte: Quadro elaborado a partir das informações de Castelloe & Zimmerman (2002).

$$R_{.m} = \sum_{c=1}^{C} R_{cm}, \text{ número de ocorrências do modelo } m; \tag{48}$$

$$\overline{\boldsymbol{\theta}}^{\cdot}_{cm} = \frac{1}{R_{cm}} \sum_{r=1}^{R_{cm}} \boldsymbol{\theta}^r_{cm}, \text{ média dos valores de } \boldsymbol{\theta} \text{ para o modelo } m \text{ e cadeia } c; \tag{49}$$

$$\overline{\boldsymbol{\theta}}^{\cdot}_{c.} = \frac{1}{T} \sum_{m=1}^{M} \sum_{r=1}^{R_{cm}} \boldsymbol{\theta}^r_{cm}, \text{ média dos valores de } \boldsymbol{\theta} \text{ para a cadeia } c; \tag{50}$$

$$\overline{\boldsymbol{\theta}}^{\cdot}_{.m} = \frac{1}{R_{.m}} \sum_{c=1}^{C} \sum_{r=1}^{R_{cm}} \boldsymbol{\theta}^r_{cm}, \text{ média dos valores de } \boldsymbol{\theta} \text{ para o modelo } m; \tag{51}$$

$$\overline{\boldsymbol{\theta}}^{\cdot}_{..} = \frac{1}{CT} \sum_{c=1}^{C} \sum_{m=1}^{M} \sum_{r=1}^{R_{cm}} \boldsymbol{\theta}^r_{cm}, \text{ média dos valores de } \boldsymbol{\theta}. \tag{52}$$

O método de Brooks & Giudici (2000), que se restringe a analisar a convergência de um parâmetro por vez, difere primeiramente do método de Castelloe & Zimmerman (2002) nos cálculos das estimativas de variância, que em Brooks & Giudici (2000) são dadas por:

$$\hat{V}(\theta) = \frac{1}{CT-1}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}(\theta_{cm}^{r}-\overline{\theta}_{..}^{\cdot})^2 \quad (53)$$

$$W_c(\theta) = \frac{1}{C}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}\frac{(\theta_{cm}^{r}-\overline{\theta}_{c.}^{\cdot})^2}{T-1} \quad (54)$$

$$W_m(\theta) = \frac{1}{M}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}\frac{(\theta_{cm}^{r}-\overline{\theta}_{.m}^{\cdot})^2}{R_{.m}-1} \quad (55)$$

$$W_mW_c(\theta) = \frac{1}{CM}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}\frac{(\theta_{cm}^{r}-\overline{\theta}_{cm}^{\cdot})^2}{R_{cm}-1} \quad (56)$$

$$B_m(\theta) = \sum_{m=1}^{M}\frac{(\overline{\theta}_{.m}^{\cdot}-\overline{\theta}_{..}^{\cdot})^2}{M-1} \quad (57)$$

$$B_mW_c(\theta) = \sum_{c=1}^{C}\sum_{m=1}^{M}\frac{(\overline{\theta}_{cm}^{\cdot}-\overline{\theta}_{c.}^{\cdot})}{C(M-1)} \quad (58)$$

No caso do diagnóstico de Brooks & Giudici (2000) tem-se as seguintes considerações quanto às equações (53) a (58):

1. $\hat{V}(\theta)$ e $W_c(\theta)$ devem ser boas aproximações para a variância total nos dados.

2. $W_m(\theta)$ e $W_mW_c(\theta)$ devem ser boas aproximações da variância média dentro de cada modelo.

3. $B_m(\theta)$ e $B_mW_c(\theta)$ devem ser boas aproximações para variância entre modelos.

Assim, especificando um valor T para tamanho do lote, deverão ser encontrados valores para as equações (53) a (58) para cada lote. Finalmente, o diagnóstico proposto por Brooks & Giudici (2000) consiste do uso de 3 análises gráficas:

1. Plotar $\hat{V}(\theta)$ e $W_c(\theta)$ calculados para cada lote.

2. Plotar $W_m(\theta)$ e $W_mW_c(\theta)$ calculados para cada lote.

3. Plotar $B_m(\theta)$ e $B_mW_c(\theta)$ calculados para cada lote.

A convergência será observada se os valores assumidos para essas estatísticas forem semelhantes nos diferentes lotes, ou seja, se os valores de $\hat{V}(\theta)$ forem próximos aos de $W_c(\theta)$, de $W_m(\theta)$ forem próximos aos de $W_m W_c(\theta)$, e de $B_m(\theta)$ forem próximos aos de $B_m W_c(\theta)$. Os valores devem ser plotados para cada lote na ordem de ocorrência. Por exemplo, supondo 5 lotes em 5.000 iterações, os valores para o primeiro lote corresponderão às 1.000 primeiras interações, seguido pelas interações de 1.000 a 2.000 até chegar ao lote com as iterações de 4.000 a 5.000, conforme ilustra a figura 1.

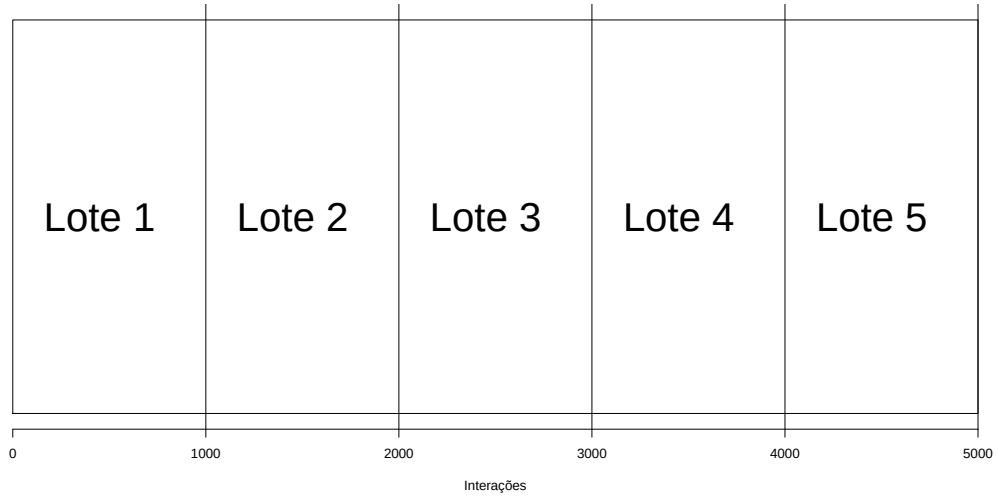

Figura 1: Ilustração da formação de lotes considerando 5.000 interações e número de lotes igual a 5.

Já no critério de Castelloe & Zimmerman (2002), as estimativas de variância, quando avaliada a convergência de um parâmetro por vez, são calculadas por

$$\hat{V}(\theta) = \frac{1}{CT-1}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}(\theta_{cm}^{r}-\overline{\theta}_{..}^{\cdot})^2 \quad (59)$$

$$W_c(\theta) = \frac{1}{C(T-1)}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}(\theta_{cm}^{r}-\overline{\theta}_{c.}^{\cdot})^2 \quad (60)$$

$$W_m(\theta) = \frac{1}{CT-M}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}(\theta_{cm}^{r}-\overline{\theta}_{.m}^{\cdot})^2 \quad (61)$$

$$W_mW_c(\theta) = \frac{1}{C(T-M)}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}(\theta_{cm}^{r}-\overline{\theta}_{cm}^{\cdot})^2 \quad (62)$$

que por sua vez, quando avaliada a convergência de múltiplos parâmetros simultaneamente, são dadas por

$$\hat{V}(\boldsymbol{\theta}) = \frac{1}{CT-1}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}(\boldsymbol{\theta}_{cm}^{r}-\overline{\boldsymbol{\theta}}_{..}^{\cdot})(\boldsymbol{\theta}_{cm}^{r}-\overline{\boldsymbol{\theta}}_{..}^{\cdot})^T \quad (63)$$

$$W_c(\boldsymbol{\theta}) = \frac{1}{C(T-1)}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}(\boldsymbol{\theta}_{cm}^{r}-\overline{\boldsymbol{\theta}}_{c.}^{\cdot})(\boldsymbol{\theta}_{cm}^{r}-\overline{\boldsymbol{\theta}}_{c.}^{\cdot})^T \quad (64)$$

$$W_m(\boldsymbol{\theta}) = \frac{1}{CT-M}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}(\boldsymbol{\theta}_{cm}^{r}-\overline{\boldsymbol{\theta}}_{.m}^{\cdot})(\boldsymbol{\theta}_{cm}^{r}-\overline{\boldsymbol{\theta}}_{.m}^{\cdot})^T \quad (65)$$

$$W_mW_c(\boldsymbol{\theta}) = \frac{1}{C(T-M)}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{r=1}^{R_{cm}}(\boldsymbol{\theta}_{cm}^{r}-\overline{\boldsymbol{\theta}}_{cm}^{\cdot})(\boldsymbol{\theta}_{cm}^{r}-\overline{\boldsymbol{\theta}}_{cm}^{\cdot})^T \quad (66)$$

Os valores dos potenciais de redução de escala são calculados, no caso com um único parâmetro, para cada parâmetro θ_i de $\boldsymbol{\theta}$, por:

$$PSRF_1(\theta_i) = \frac{\hat{V}(\theta_i)}{W_c(\theta_i)}, \quad (67)$$

$$PSRF_2(\theta_i) = \frac{W_m(\theta_i)}{W_mW_c(\theta_i)}, \quad (68)$$

e para o caso com múltiplos parâmetros por

$$MPSRF_1(\boldsymbol{\theta}) = \text{máximo autovalor de } [W_c(\boldsymbol{\theta})]^{-1}\hat{V}(\boldsymbol{\theta}), \quad (69)$$

$$MPSRF_2(\boldsymbol{\theta}) = \text{máximo autovalor de } [W_mW_c(\boldsymbol{\theta})]^{-1}W_m(\boldsymbol{\theta}). \quad (70)$$

A estratégia de diagnóstico sugerida por Castelloe & Zimmerman (2002) é plotar, quando $\boldsymbol{\theta}$ for composto por mais de um parâmetro, os valores de $MPSRF_1(\boldsymbol{\theta})$ e $MPSRF_2(\boldsymbol{\theta})$ (juntos ou separados) vs o número do índice de lote (*index plot*), sendo a convergência alcançada quando ambos os valores estiverem próximos de 1 ao longo dos diferentes lotes, plotar os maiores autovalores de $\hat{V}(\boldsymbol{\theta})$ e $W_c(\boldsymbol{\theta})$ (juntos) vs o número do índice de lote; e os maiores autovalores de $W_m(\boldsymbol{\theta})$ e $W_mW_c(\boldsymbol{\theta})$ (juntos) vs o número do índice de lote.

Nos gráficos dos maiores autovalores, quando os valores observados nos gráficos estiverem próximos nos diferentes lotes é observada a convergência das cadeias. Também podem ser plotados os gráficos individuais para cada parâmetro θ_i de $\boldsymbol{\theta}$ de maneira análoga ao caso com múltiplos parâmetros, plotando os gráficos $PSRF_1(\theta_i)$ e $PSRF_2(\theta_i)$ (juntos ou separados) vs o número do índice de lote com valores próximos a 1 indicando convergência; $\hat{V}(\theta_i)$ e $W_c(\theta_i)$ (juntos) vs o número do índice de lote e $\hat{W_m}(\theta_i)$ e $W_mW_c(\theta_i)$ (juntos) vs o número do índice de lote com valores próximos entre si indicando convergência, para $i = 1, \ldots, p$, em que p é o número de parâmetros de $\boldsymbol{\theta}$. Esses procedimentos mostrados para cada θ_i são, obviamente, os únicos possíveis de serem utilizados quando $\boldsymbol{\theta}$ for composto apenas por um único parâmetro.

A ideia por trás do critério multivariado é que os potenciais de escala encontrados funcionam como um limite superior para os valores possíveis de cada potencial de redução de escala individual, garantido pelo teorema da majoração que é dado por Castelloe & Zimmerman (2002):

$$MPSRF_1(\boldsymbol{\theta}) \geq \max_i PSRF_1(\theta_i) \text{ e } MPSRF_2(\boldsymbol{\theta}) \geq \max_i PSRF_2(\theta_i) \qquad (71)$$

Os estimadores encontrados por Castelloe & Zimmerman (2002) são baseados em três modelos ANOVA para o caso univariado e MANOVA para o caso multivariado. Em todos os modelos ANOVA temos o parâmetro θ_i como resposta, e nos modelos MANOVA o vetor paramétrico $\boldsymbol{\theta}$. Na primeira ANOVA (ANOVA1) considera-se apenas os efeitos da cadeia; na segunda (ANOVA2) os efeitos simples do

modelo; e na terceira (ANOVA3) os efeitos simples mais o efeito da interação entre modelo e cadeia.

O estimador $W_c(\theta_i)$ é o erro quadrático médio dos resíduos e $V(\theta_i)$ o erro quadrático médio total na ANOVA1; $W_m(\theta_i)$ erro quadrático médio dos resíduos na ANOVA2; e $W_cW_m(\theta_i)$ erro quadrático médio dos resíduos da ANOVA3 (Castelloe & Zimmerman, 2002). Para o modelo MANOVA as considerações são semelhantes, todavia, nesse caso, os estimadores encontrados são equivalentes aos maiores autovalores das respectivas razões entre as matrizes de variância e covariância estimadas e os seus respectivos graus de liberdade.

Nos modelos ANOVA e MANOVA tem-se que o efeito da cadeia é fixo e do modelo é aleatório (Castelloe & Zimmerman, 2002). Todavia, os estimadores propostos em Castelloe & Zimmerman (2002) coincidem com os valores da ANOVA e MANOVA considerando ambos os efeitos fixos. De tal modo, os valores foram estimados, no presente trabalho, utilizando a construção dos modelos ANOVA considerando ambos os efeitos como fixos.

Os estimadores propostos em Castelloe & Zimmerman (2002) não consideram a possibilidade de alguns dos modelos não serem visitados por todas as cadeias, não existindo, portanto, efeitos de interações entre modelo e cadeia. Dessa forma é necessário fazer as correções dos graus de liberdade para o cálculo correto do erro quadrático médio dos resíduos na ANOVA3 e da medida equivalente nos modelos MANOVA. Tal correção é padrão nas funções ANOVA e MANOVA do R (R Core Team, 2019).

Nas Figuras 2 a 4 tem-se exemplos dos gráficos gerados utilizando o critério de Castelloe & Zimmerman (2002) para o caso univariado. Foram geradas 3 cadeias distintas com tamanho 100.000 cada de uma distribuição normal com média $\mu = 20$ e variância $\sigma^2 = 2,25$, sendo o algoritmo executado em 20 lotes. Para simular a variável indicadora do modelo usou-se a função `sample` do R que gerou as variáveis indicadoras de modelo 1,2 e 3 na proporção aproximada $1 : 2 : 1$.

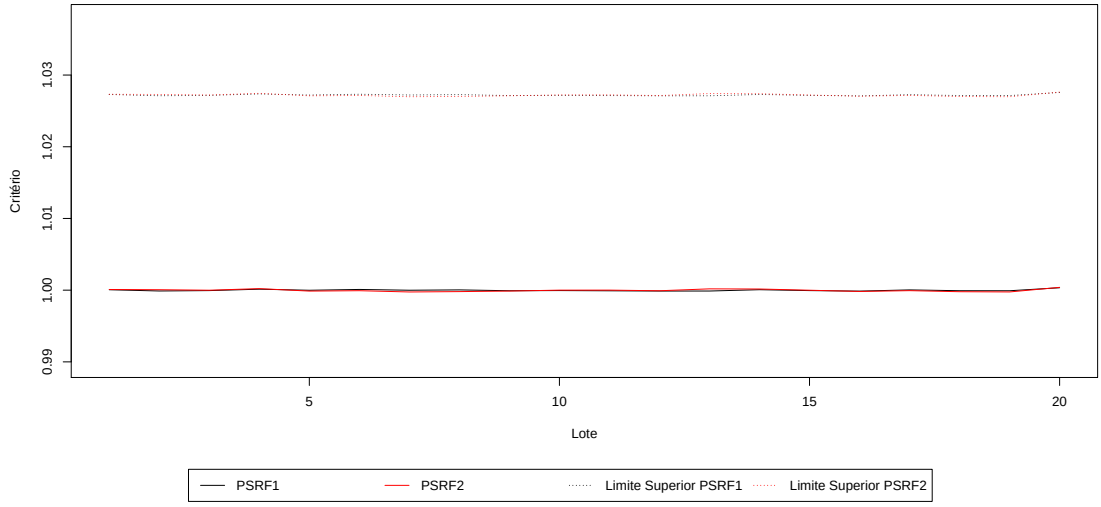

Figura 2: $PSRF1$ e $PSRF2$ vs Lotes.

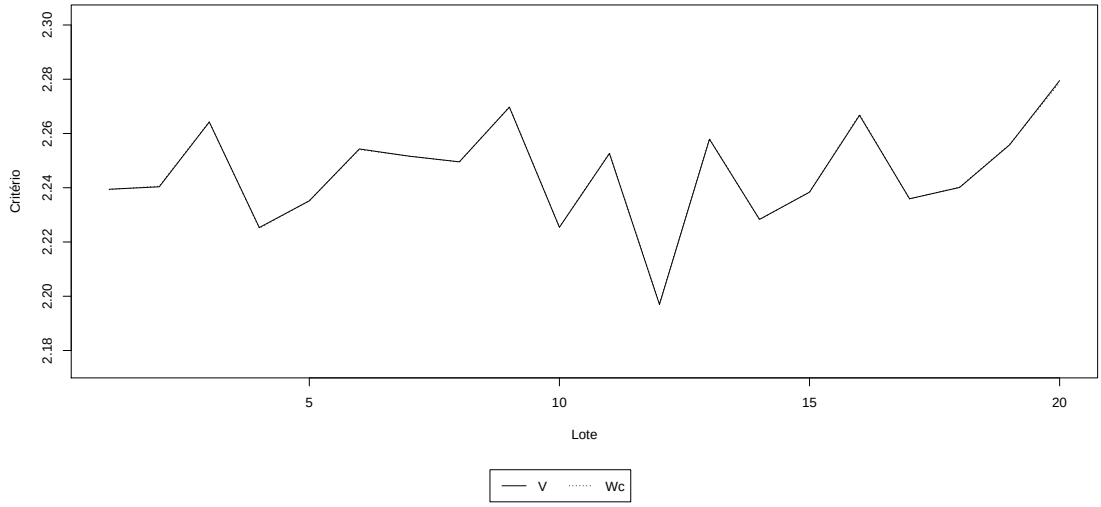

Figura 3: V e Wc vs Lotes.

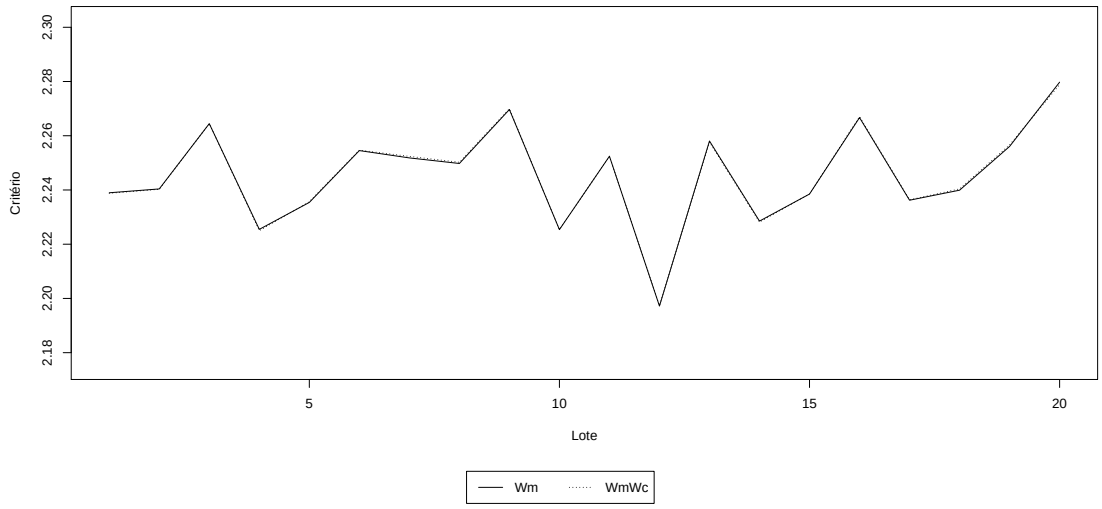

Figura 4: Wm e $WmWc$ vs Lotes.

Nas Figuras 5 a 7 tem-se exemplos dos gráficos gerados utilizando o critério de Castelloe & Zimmerman (2002) para o caso multivariado. Foram feitas 3 amostras distintas para cada parâmetro diferente a fim de simular as cadeias, de forma a ilustrar vetores paramétricos iguais entre diferentes cadeias, utilizando 3 parâmetros diferentes, todos com distribuição normal com médias μ iguais a 0, 1 e 1 e variâncias σ^2 iguais a 1, 4 e 9, respectivamente. A simulação das variáveis indicadoras do modelo foi feita da mesma forma que para o caso univariado.

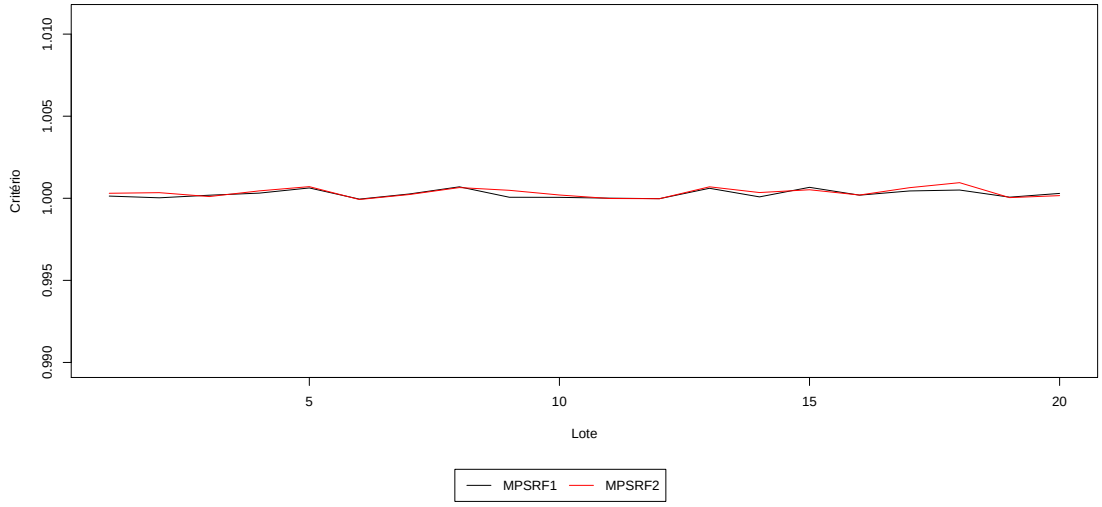

Figura 5: $MPSRF1$ e $MPSRF2$ vs Lotes.

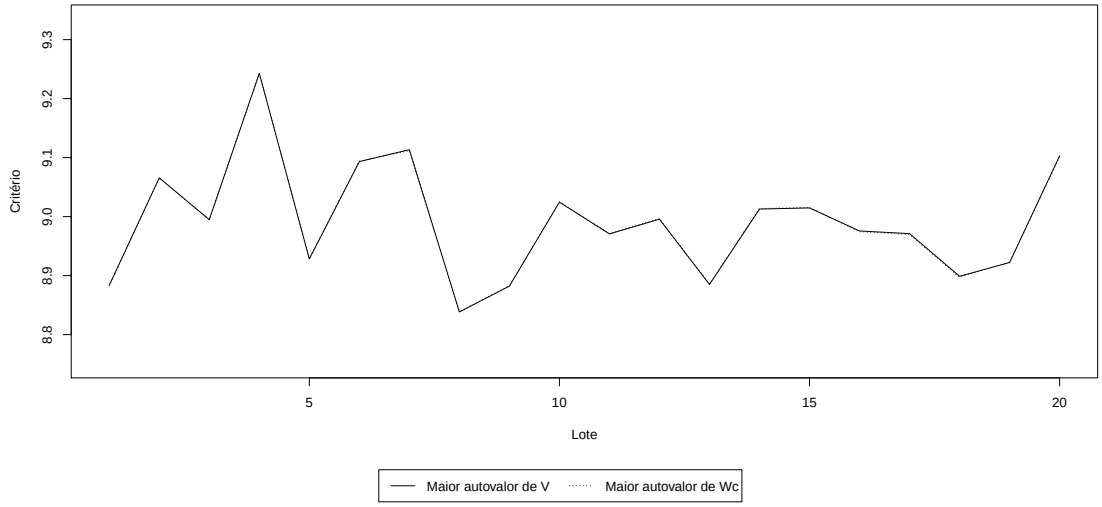

Figura 6: Maior Autovalor de V e Wc vs Lotes.

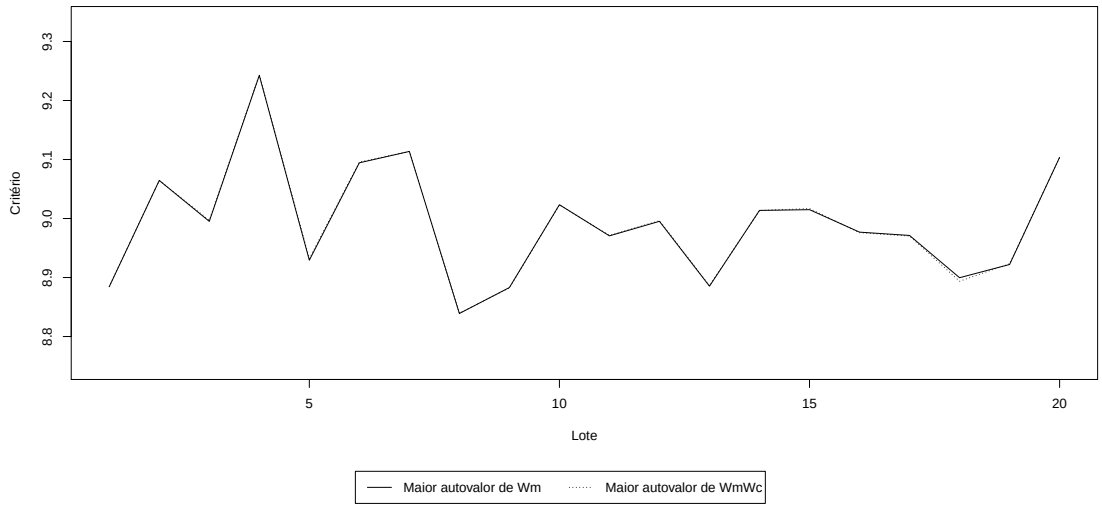

Figura 7: Maior Autovalor de Wm e $WmWc$ vs Lotes.

O código com os exemplos apresentados utilizando o pacote convRJMCMC está no arquivo exemplos_RJMCMC.R dos arquivos suplementares. O repositório do pacote convRJMCMC no GitHub pode ser acessado pelo segundo link do apêndice.

3.3.2 Gráfico de Traço

O gráfico de traço, ou *traceplot*, é uma das ferramentas mais utilizadas para diagnóstico de convergência para diferentes algoritmos baseados em simulações aleatórias. Ele consiste de gráficos de linhas que representam as iterações da simulação na abscissa e o valor de interesse, obtido em cada iteração da simulação, na ordenada. Usualmente, para diagnósticos de métodos de Monte Carlo via cadeias de Markov, são analisadas as medidas para um mesmo parâmetro tomado em cadeias diferentes. A convergência é observada quando os valores para os parâmetros nas diferentes cadeias parecem oscilar todos sobre um único valor comum.

A Figura 8 é um exemplo de gráfico de traço gerado com a ajuda do pacote **coda** (Plummer et al., 2006) mostrando o caso de convergência entre as

distribuições, em que as estimativas das 3 cadeias estão oscilando em torno do mesmo valor.

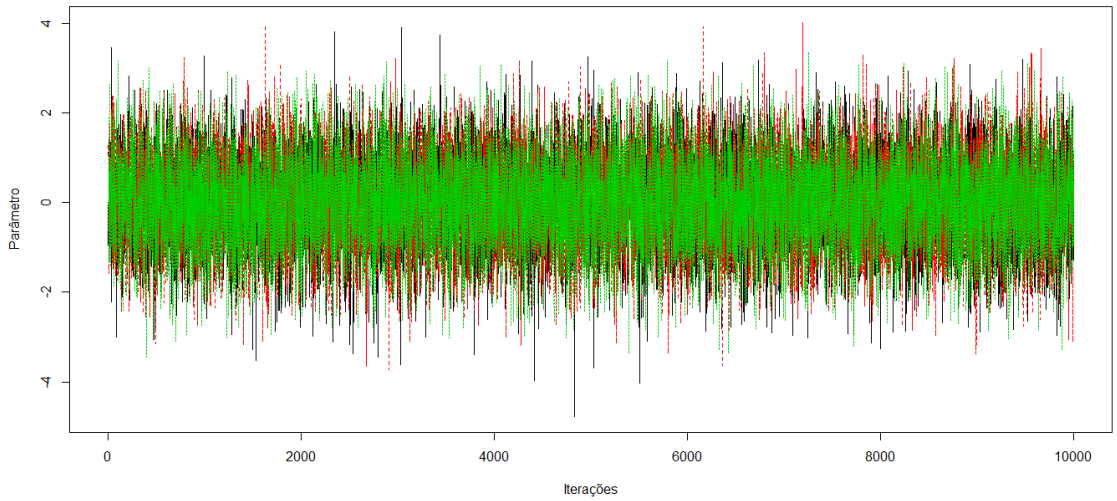

Figura 8: Exemplo de gráfico de traços com 3 cadeias quando as distribuições convergem. As 3 cadeias são amostras de distribuições normal padrão com 10000 observações cada.

3.3.3 Gráfico de Autocorrelação

Uma série temporal é um exemplo de processo estocástico. Uma série pode ser representada pelo conjunto $X_1, \ldots, X_t, \ldots, X_n$, em que X_t é uma variável aleatória que representa o estado do processo no momento t. Ao adicionarmos o atraso a, temos os seguintes pares $(X_t, X_{t+a}) : (X_1, X_{1+a}), \ldots, (X_{n-a}, X_n)$ (Venables, 2002). Assim, a autocorrelação é, para cada atraso a, o coeficiente de correlação calculado entre as sequências para os valores observados da série temporal.

O resultado obtido em uma cadeia de Markov pode ser visto como uma série temporal, em que valores estimados para um mesmo parâmetro são retidos em diferentes iterações, que podem ser vistas como o tempo de uma série temporal, da cadeia. Uma forma de se avaliar a independência em cadeias de Markov é com o uso da autocorrelação, sendo esperado, quando independentes, que oscile próximo

ao valor 0.

Na Figura 9, temos um exemplo de gráfico de autocorrelação utilizando a função `acf` no R. Ela foi aplicada a uma amostra simulada de uma distribuição normal padrão ($\mu = 0, \sigma^2 = 1$) com 10000 observações. O resultado obtido é o que se espera em uma amostra de observações independentes.

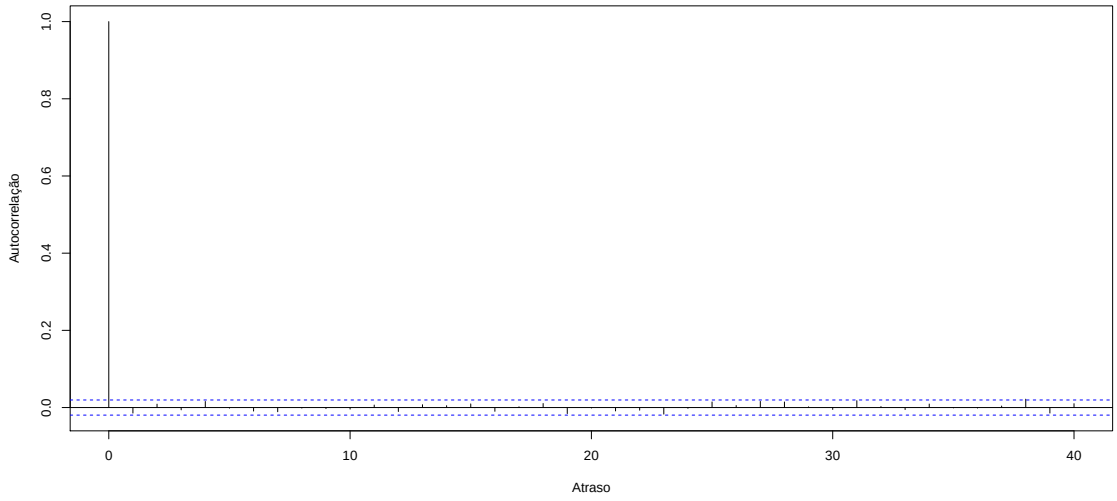

Figura 9: Exemplo de gráfico de autocorrelação para amostra independente da distribuição normal padrão com 10000 observações.

Gráficos de autocorrelação não são usados para verificar a convergência para a distribuição estacionária, mas sim, para verificar a independência da amostra gerada, feito em uma cadeia por vez.

4 MATERIAL E MÉTODOS

Para uma melhor organização, os dois problemas da área da saúde analisados serão apresentados em seções distintas. Porém, antes serão descritos detalhes relacionados às simulações e algumas funções criadas para gerar os resultados finais.

4.1 Análise de Convergência e Dependência

Visto a ausência de pacotes já desenvolvidos para plotar os gráficos de diagnóstico de convergência propostos por Castelloe & Zimmerman (2002), foram criadas as funções `CZ_ANOVA` e `plot.CZ_ANOVA` e inseridas no pacote convRJMCMC (Marques & Tsunemi, 2021), junto as funções `CZ_MANOVA` e `plot.CZ_MANOVA` que são usadas para avaliação de convergência de múltiplos parâmetros. A função `CZ_ANOVA` cria um objeto de classe "CZ_ANOVA" que contém um conjunto de 11 listas:

- `sm_aov1`: Lista com os resumos do modelo ANOVA1 para cada lote;
- `sm_aov2`: Lista com os resumos do modelo ANOVA2 para cada lote;
- `sm_aov3`: Lista com os resumos do modelo ANOVA3 para cada lote;
- `PSRF1`: Lista de valores calculados para o critério PSRF1 para cada lote;
- `PSRF2`: Lista de valores calculados para o critério PSRF2 para cada lote;
- `ub_PSRF1`: Lista de valores calculados para o limite superior do critério PSRF1 para cada lote;
- `ub_PSRF2`: Lista de valores calculados para o limite superior do critério PSRF2 para cada lote;

- `V`: Lista de valores calculados de $V(\theta)$ para cada lote;
- `Wc`: Lista de valores calculados de $Wc(\theta)$ para cada lote;
- `Wm`: Lista de valores calculados de $Wm(\theta)$ para cada lote;
- `Wcm`: Lista de valores calculados de $WmWc(\theta)$ para cada lote.

Os argumentos são:

- `theta`: é o vetor de observações do parâmetro escolhido do modelo para análise de convergência;
- `chains`: vetor de variáveis indicadoras para qual cadeia pertence cada observação;
- `models`: a variável indicadora do modelo a qual pertence a observação;
- `mcmciterations`: vetor com as iterações (após *burn-in* e *thin*) a que pertencem as observações;
- `nbatches`: número de lotes;
- `batchsize`: tamanho do lote;
- `confidence`: nível de confiança do intervalo de confiança unilateral para os limites superiores dos critérios PSRF1 e PSRF2 e;
- `division`: critério utilizado para divisão dos lotes, em que `Batch` divide o banco de dados formando partições com as iterações da cadeia e `Sequential` que cria lotes consecutivos de tamanho igual a metade do número de iterações e em número igual a `nbatches` com sobreposição das iterações entre lotes. Somente o primeiro critério de divisão foi utilizado no trabalho.

A segunda função, `plot.CZ_ANOVA`, utiliza o objeto de classe "CZ_ANOVA" para plotar os 3 gráficos de diagnóstico de convergência: PSRF1

e PSRF2 vs Lote; $V(\theta)$ e $Wc(\theta)$ vs lote; e $Wm(\theta)$ e $WmWc(\theta)$ vs lote. Os gráficos de autocorrelação foram obtidos com a função `acf` do R (R Core Team, 2019) e os gráficos de traços com a função `traceplot` do pacote **coda** (Plummer et al., 2006).

4.2 Crioablação Vertebral em Suínos

Crioablação diz respeito a qualquer método utilizado para a destruição de tecidos por resfriamento, a destruição do tecido pode se dar por lesões causadas pelo frio nos tecidos ou por mecanismos indiretos que alteram o microambiente celular, impedindo a viabilidade do tecido (Erinjeri & Clark, 2010). É um método utilizado para tratamento localizado de câncer, em que é introduzida uma sonda de crioablação no tecido afetado, com o objetivo de destruir as células cancerígenas ocasionando o menor dano possível a tecidos saudáveis do organismo. Assim, é necessário que os protocolos de tratamento de câncer por crioablação sejam validados para que possam ser usados com segurança no tratamento de pacientes, grande motivador para estudos como o de Freitas et al. (2015) e Freitas (2015), do qual obteve-se acesso a parte dos dados.

O experimento de Freitas et al. (2015) e Freitas (2015) foi realizado em 6 vértebras de um mesmo suíno em 2 ciclos de congelamento (por 2 minutos) e descongelamento (por 8 minutos). Em cada uma das vértebras foi inserida uma sonda de crioablação (CIOPROBE) e 4 termopares (PROBE 3, PROBE 5, PROBE 7 e PROBE 9) localizados a distâncias radiais crescentes da sonda (Tabela 2). As temperaturas foram registradas a partir do início do congelamento em intervalos de 30 segundos e finalizando em 600 segundos (10 minutos), havendo para cada tempo de registro uma medição de temperatura por sonda ou termopar por vértebra nos dois ciclos de congelamento. A sonda e os termopares foram removidos para limpeza após o término do primeiro ciclo e reinseridos nas mesmas vértebras para realização do segundo ciclo (Freitas et al., 2015; Freitas, 2015).

O ajuste das curvas de temperatura foi feito sem considerar a dependência temporal dos dados com *splines cúbicas* (ordem 4 ou grau 3) utilizando

Tabela 2. Distância dos termopares a sonda de crioablação em milímetros.

Ciclo	1						2					
Vertebra	1	2	3	4	5	6	1	2	3	4	5	6
PROBE 3	3,8	3,3	3,6	3,7	3,7	3,3	3,8	3,3	3,6	3,6	3,6	3,4
PROBE 5	5,1	4,3	4,4	4,2	4,6	4,8	4,6	4,3	4,4	4,5	4,4	4,8
PROBE 7	7,9	7,5	6,6	6,4	6,6	6,8	7,9	7,5	6,6	6,4	6,6	6,9
PROBE 9	8,1	8,4	9,6	8,9	8,5	8,4	8,8	8,4	9,6	8,7	8,5	8,6

a função `bass` do pacote **BASS** (Francom & Sansó, 2019) e `fit.search.numknots` do pacote **freeknotsplines** (Spiriti et al., 2018) do R, que foi removido do CRAN no dia 03 de março de 2020, todavia o código fonte do mesmo pode ser encontrado no seguinte endereço https://github.com/cran/freeknotsplines. Neste trabalho serão apresentados os resultados obtidos no ajuste da curva de temperatura no PROBE 5 durante o segundo ciclo de congelamento. Foi escolhido o PROBE 5 no segundo ciclo por estar localizado a uma distância intermediária entre todos os termopares, no qual foi possível observar grande parte dos problemas encontrados nos demais termopares.

O diagnóstico de convergência do método BASS foi feito usando gráfico de traços e os gráficos de diagnóstico propostos por Castelloe & Zimmerman (2002) implementados pelo pacote convRJMCMC (Marques & Tsunemi, 2021) do R. Além disso, os gráficos de autocorrelação foram utilizados para avaliar a dependência da posteriori do parâmetro σ^2 nas cadeias de Markov geradas, a fim de obter um salto (*thin*) adequado. Foram simuladas 3 cadeias para cada curva, em que uma curva é feita para ajustar os dados de cada conjunto entre sonda ou termopar e ciclo. A semente escolhida para o PROBE 5 durante o segundo ciclo de congelamento foi 219 (sorteada a partir de uma distribuição uniforme discreta de 0 a 1000). Tal procedimento de sorteio foi repetido ao longo do trabalho. Nas simulações foram utilizados salto de 100, *burn-in* de 10000 e 1010000 iterações anteriormente ao *burn-in* e salto em cada cadeia (é necessário especificar o número de iterações totais, antes

do descarte de parte da amostra), totalizando uma amostra com 10000 observações da posteriori em cada cadeia ou 30000 observações no total, fixando o grau da *spline* como 3 (`degree`).

No algoritmo do pacote `freeknotsplines` foi definido o número mínimo de nós igual a 1 (`minknot`), máximo igual a 10 (`maxknot`) e grau 3 (`degree`). O número máximo de 10 foi utilizado por considerar que o mesmo, para o contexto, já seria um modelo bastante complexo, sendo improvável que o modelo ótimo utilizasse mais de 10 nós. Caso, todavia, o método selecionasse o modelo ótimo com 10 nós seriam avaliados números maiores de nós para ver se algum dos modelos com mais nós se adequariam melhor aos dados o que, todavia, não foi necessário para os dados avaliados. As outras configurações foram mantidas como padrão, o que significa o uso de algoritmo genético para encontrar a curva ótima de acordo com o critério da validação cruzada generalizado (Craven & Wahba, 1979). A semente utilizada para o PROBE 5 durante o segundo ciclo de congelamento foi 949 (escolhida por sorteio).

Para plotar as curvas médias e intervalos de credibilidade a partir das densidades preditivas criou-se a função `pred_line`. A mesma utiliza as cadeias geradas e armazenadas nos objetos "chain1" , "chain2" e "chain3" e plota a curva média entre todas as densidades preditivas e o intervalo de credibilidade 95% considerando, também, a distribuição empírica da variância dos resíduos. O argumento "yourcolor" indica a cor desejada para as curvas plotadas, "lw" indica a espessura da linha ("lwd" função `plot`), "lt1" o tipo de linha usado para a curva média ("lty" da função `plot`), "lt2" o tipo de linha usado para as curvas que indicam os limites inferior e superior do intervalo de credibilidade 95% e "mcmcit" as iterações a serem utilizadas das 3 cadeias (após *burn-in*) para plotar as curvas.

Para plotar as curvas e intervalos de confiança do modelo ajustado pela função `fit.search.numknots` criou-se a função `fpred_line`. A mesma utiliza o objeto "freeknotobj" , gerado pela função `fit.search.numknots`, e que representa o modelo ajustado pela função, e plota a curva do modelo e o intervalo de confiança 95% considerando, também, a variância estimada para os resíduos. Os argumentos

de formatação das curvas são similares aos da função `pred_line`.

Para análise de resíduos e avaliação do modelo foram plotados uma sequência de gráficos que contemplam: resíduos vs tempo (variável independente), resíduos vs temperatura estimada, temperatura estimada vs temperatura observada e gráfico quantil-quantil da normal. Os mesmos foram obtidos com o auxílio de duas funções. Uma, `BASSresiduals`, que cria um objeto de classe "BASSresiduals" que contempla uma tabela de dados (objeto "data.frame") com o valor da variável independente (tempo), da variável dependente (temperatura observada), estimativa do modelo para cada observação da variável independente (temperatura estimada) e os resíduos do modelo para cada observação. Ela utiliza os objetos "chain1" , "chain2" e "chain3" gerados pela função `bass` para cada cadeia, "x_data" que é o vetor dos valores observados da variável independente, "y_data" que é o vetor dos valores observados da variável dependente e "mcmcit" indica as iterações utilizadas de cada uma das cadeias (após *burn-in*). E outra função `plot.BASSresiduals` que plota os gráficos de interesse utilizando o objeto de classe "BASSresiduals" gerado.

Analogamente, temos a função `fnsresiduals` que a partir do objeto "fns_model", gerado pela função `fit.search.numknots`, junto a um conjunto de vetores "x_data" e "y_data", que representam os dados observados de tempo e temperatura, respectivamente, gera um objeto da classe `fnsresiduals` de mesma estrutura que o objeto de classe `BASSresiduals`. Por sua vez, os gráficos são gerados pela função `plot.fnsresiduals`. As funções criadas e a implementação para gerar os resultados estão no arquivo codigo_crioablacao_P5_c2.R da pasta com os arquivos suplementares.

A avaliação das curvas será feita conforme as seguintes faixas de temperatura (Freitas et al., 2015; Freitas, 2015):

- temperaturas letais, abaixo de $-20°C$, provocam morte celular;
- temperaturas entre $-20°C$ a $8°C$ podem causar danos a tecidos e nervos próximos a área afetada mas sem a morte das células, ou seja, podem oca-

sionar danos indesejáveis às células saudáveis sem eliminar o câncer;

- temperaturas acima de $8°C$ não provocam danos às células e tecidos.

4.3 Curvas de DOB na ausência e presença de *Helicobacter pylori*

O teste respiratório da úreia ^{13}C é um método não invasivo para detecção de *Helicobacter pylori*. Embora já seja um exame consolidado, ainda são poucos os estudos que buscam avaliar o tempo ótimo, com menor número de falsos positivos e negativos, após a administração oral da uréia ^{13}C para realização do teste, um dos objetivos principais do estudo de Garcia (2017), do qual temos acesso a parte dos dados.

O DOB (*Delta Over Baseline*) é uma medida calculada através do enriquecimento isotópico relativo $\delta^{13}C(‰)$, o qual é dado por (Garcia, 2017):

$$\delta^{13}C(‰) = \left(\frac{R_{Amostra}}{R_{Padrão(PDB)}} - 1\right) \times 1000 \tag{72}$$

em que $R_{Amostra}$ é a razão isotópica $(^{13}C/^{12}C)$ expirado durante o exame e $R_{Padrão(PDB)} = 0,0112372$ é a razão isotópica $(^{13}C/^{12}C)$ do padrão internacional para o carbono (PDB). O DOB é dado pela expressão (Garcia, 2017):

$$DOB(‰) = \delta^{13}C_{Amostra} - \delta^{13}C_{Basal}. \tag{73}$$

É válido notar que o valor de $\delta^{13}C_{Basal}$ é aquele medido antes a ingestão da uréia. Portanto, o DOB no tempo de 0 minutos será sempre igual a 0. Por tal razão, na fase de processamento, descrita a seguir, as observações no tempo 0 foram removidas.

O banco de dados utilizado conta com informações de 120 pacientes discriminados pelo resultado dos testes histológicos, teste UBT (o paciente foi classificado como positivo quando a medida de DOB foi superior a 4‰), peso, altura, IMC, sexo, data de nascimento, data em que os pacientes participaram do estudo,

idade, raça e as medidas de DOB em 15 instantes de tempo pré-determinados: 0 (momento de ingestão da uréia); 5; 7,5; 10; 12,5; 15; 17,5; 20; 22,5; 25; 27,5; 30; 35; 40 e 45 minutos após ingestão da uréia, sendo as observações no tempo 0 removidas na etapa de processamento. Anteriormente a importação dos dados no R, 8 pacientes, que não apresentavam medidas de DOB, foram excluídos. Os pacientes foram divididos conforme o resultado, positivo (54 pacientes) ou negativo (58 pacientes), do exame histológico.

O ajuste Bayesiano das curvas foi feito sem considerar a dependência temporal dos dados com a função `bass` do pacote **BASS** (Francom & Sansó, 2019) do R que utiliza o método RJMCMC para simulação das cadeias. Considerando os diagnósticos de convergência e autocorrelação nas cadeias foi utilizado *thin* de 100 ("thin"), *burn-in* de 1000 ("nburn") e um total de 1001000 iterações antes do *burn-in* e do *thin* ("nmcmc") para a simulação de cada uma das cadeias. Foram simuladas um total de 3 cadeias para cada curva, com sementes iguais a 138 e 367 para a curva dos pacientes com exame histológico positivo e negativo, respectivamente, para reprodutibilidade do estudo.

Dessa forma, tem-se um total de 30000 densidades preditivas para cada subdivisão do banco de dados (pacientes com exame histológico positivo e negativo), ou seja, para cada uma das curvas ajustadas. O ajuste foi feito utilizando *splines* cúbicas, sendo indicado o grau 3 da *spline* no parâmetro "degree" da função `bass`.

Para o novo conjunto de dados a função `pred_line` original foi modificada, para se comportar adequadamente ao intervalo das variáveis independentes.

Os intervalos e curvas geradas pela função `bass` foram comparadas às geradas pela função `fit.search.numknots` do pacote **freeknotsplines**. Sendo encontrada a função *spline* cúbica ótima dentre as com número mínimo de nós 0 e máximo 10. As sementes utilizadas foram 993 para a curva dos pacientes com teste histológico positivo e 561 para a curva dos pacientes com teste histológico negativo.

A função `fpred_line` foi modificada da mesma forma que a `pred_line`.

Os intervalos de credibilidade 95% plotados para o modelo BASS seguem das equações:

$$LI = \hat{y} - 1{,}96 \times \sqrt{\sigma^2}; \quad (74)$$

$$LS = \hat{y} + 1{,}96 \times \sqrt{\sigma^2}, \quad (75)$$

em que $\hat{y}$ e σ^2 representam a densidade preditiva e a variância, respectivamente, para cada interação da cadeia de Markov. O intervalo final foi obtido tomando o quantil 0,025 dos limites inferiores calculados em todas as interações como limite inferior e o quantil 0,975 dos limites superiores calculados como limite superior.

Os intervalos de confiança 95% plotados para o modelo obtido pelo algoritmo `freeknotsplines` seguem as equações:

$$LI = \hat{y} - 1{,}96 \times \sqrt{Var(\hat{y} - y)}; \quad (76)$$

$$LS = \hat{y} + 1{,}96 \times \sqrt{Var(\hat{y} - y)}, \quad (77)$$

em que Var é a função utilizada para estimar variância de uma distribuição normal, dada, para determinado conjunto de valores observados $\{X_1, \ldots, X_n\}$ de uma variável X, por:

$$Var(X) = \frac{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}{n-1}, \quad (78)$$

em que $\overline{X}$ é a média aritmética dos valores observados de X.

Para análise de resíduos e avaliação do modelo foram plotados uma sequência de gráficos que contemplam: resíduos vs tempo (variável independente), resíduos vs DOB estimado, DOB estimado vs observado e gráfico quantil-quantil da distribuição normal. Os mesmos foram obtidos por duas funções `BASSresiduals` e `plot.BASSresiduals` que são modificações das originais para se adequar ao novo conjunto de dados. Analogamente as funções `BASSresiduals` e `plot.BASSresiduals`, foram feitas as modificações das funções `fnsresiduals` e `plot.fnsresiduals`.

As funções criadas e a implementação para gerar os resultados estão no arquivo codigo_DOB.R da pasta com os arquivos suplementares.

5 RESULTADOS E DISCUSSÃO

5.1 Crioablação Vertebral

O gráfico de dispersão na Figura 10 mostra o comportamento da temperatura, nas diferentes vértebras, registradas no PROBE 5 durante o segundo ciclo de congelamento. As curvas ajustadas pelo método BASS e `freeknotsplines`, que não consideram a dependência temporal dos valores, pouco diferem entre si. Os intervalos de credibilidade (95%) do método BASS, com maior amplitude que os intervalos do método `freeknotsplines` mostram uma tendência dele ser mais conservador em suas inferências. Considerando os limiares de temperatura, abaixo de $-20°C$ indicando regiões de morte celular e acima de $8°C$ indicando regiões em que não ocorre dano celular, o PROBE 5 é uma em que ocorre a morte celular. Além do gráfico de dispersão convencional, é apresentado para o método BASS o gráfico com as 30000 densidades preditivas (10000 interações por cadeia em 3 cadeias) na Figura 11.

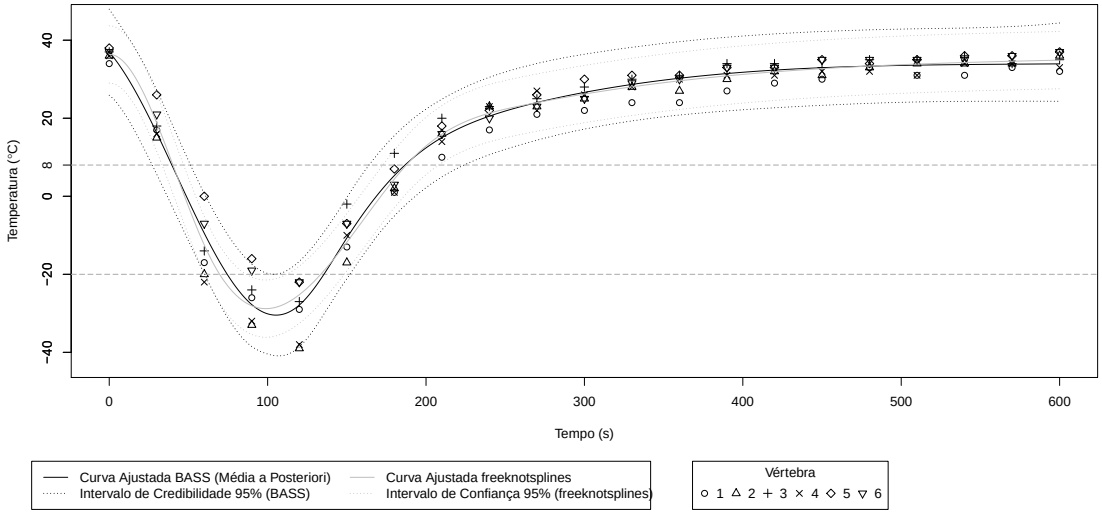

Figura 10 - Gráfico de dispersão com curvas dos modelos e intervalos de confiança ou de credibilidade utilizando dados do termopar PROBE 5 durante ciclo 2.

Os resultados gerados pelo método BASS e pacote **freeknotsplines** são bem diferentes, embora as curvas do modelo BASS e do pacote **freeknotsplines** sejam semelhantes na Figura 10. O BASS é composto de uma coleção de modelos, curvas, correspondentes a cada interação do algoritmo RJMCMC nas 3 diferentes cadeias. Assim, cada densidade preditiva observada na Figura 11 é a representação gráfica de cada um desses modelos e os pontos da curva final apresentada no diagrama de dispersão do BASS (Figura 10) são dados pela média aritmética dos pontos de todas as densidades preditivas para cada valor de tempo. Assim, embora seja possível obter a expressão de cada um dos modelos, a expressão da curva final e os nós do modelo final não podem ser apresentados, ao menos de forma conveniente. O pacote **freeknotsplines**, em contrapartida, tem como resultado um único modelo ótimo escrito em bases *B-Spline* para modelos com pelo menos um nó ou como curvas de bézier para modelos sem nenhum nó. O modelo obtido com o pacote **freeknotsplines** tem sequência aumentada de nós dada por

$\gamma = \{0; 0; 0; 0; 64{,}688; 192{,}213; 259{,}359; 600; 600; 600; 600\}$ e o modelo dado por:

$$\begin{aligned} f(x) &= 36{,}338 \times \mathfrak{B}_{1,4}(X) + 36{,}718 \times \mathfrak{B}_{2,4}(X) - 71{,}131 \times \mathfrak{B}_{3,4}(X) + \\ & 14{,}963 \times \mathfrak{B}_{4,4}(X) + 32{,}352 \times \mathfrak{B}_{5,4}(X) + 33{,}856 \times \mathfrak{B}_{6,4}(X) + \\ & 34{,}892 \times \mathfrak{B}_{7,4}(X) \end{aligned} \tag{79}$$

que é equivalente a:

$$f(X) = \begin{cases} p_1 = 36{,}33818 + 0{,}01762507 \times X - 0{,}02629412 \times X^2 \\ +0{,}0002073165 \times X^3,\ 0 \leq X < 64{,}688; \\ p_2 = 105{,}5829093 - 3{,}193714765 \times x + 0{,}023349627 \times X^2 \\ -0{,}0000484959 \times X^3,\ 64{,}688 \leq X < 192{,}213; \\ p_3 = -394{,}034145 + 4{,}60416312 \times X - 0.017219337 \times X^2 \\ +0{,}00002185846 \times X^3,\ 192{,}213 \leq X < 259{,}359; \\ p_4 = -16{,}2036114 + 0{,}233805428 \times X - 0{,}000368726 \times X^2 \\ +0{,}00000020106 \times X^3,\ 259{,}359 \leq X \leq 600 \end{cases} \tag{80}$$

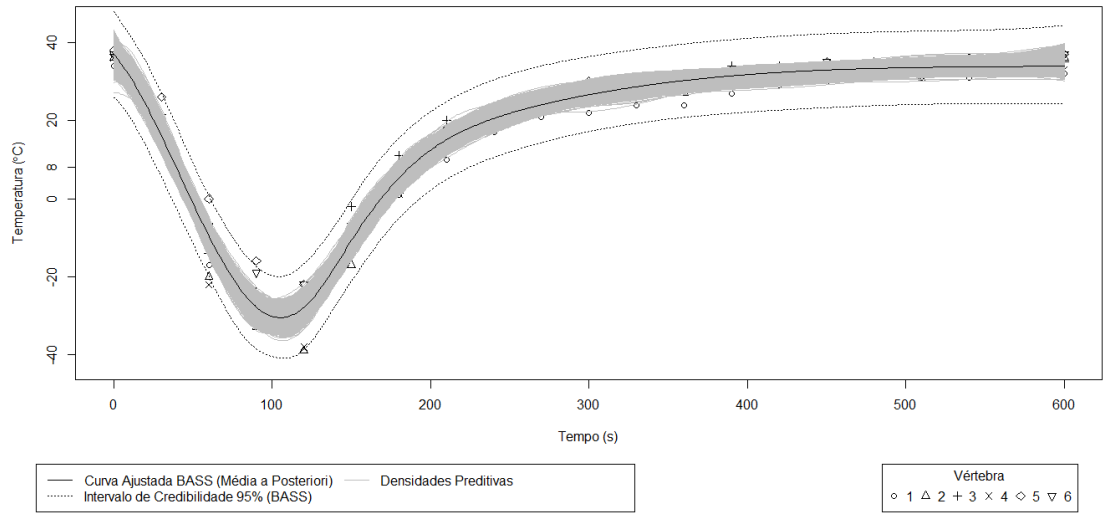

Figura 11 - Gráfico de dispersão com densidades preditivas do modelo BASS ajustado aos dados do termopar PROBE 5 durante o ciclo 2.

O gráfico de traços, apresentado na Figura 12 com o resultado das 3 cadeias, ilustra a convergência do algoritmo, considerando o comportamento do parâmetro σ^2, ilustrado pelo fato das cadeias estarem oscilando em torno de um mesmo valor. Resultado também evidente nos gráficos diagnósticos utilizando o critério de Castelloe & Zimmerman (2002), com os potencias de redução de escala, e seus limites superiores, próximos a 1 (Figura 13), os valores de V próximos aos de Wc (Figura 14) e os de Wm próximos aos de $WmWc$ (Figura 15).

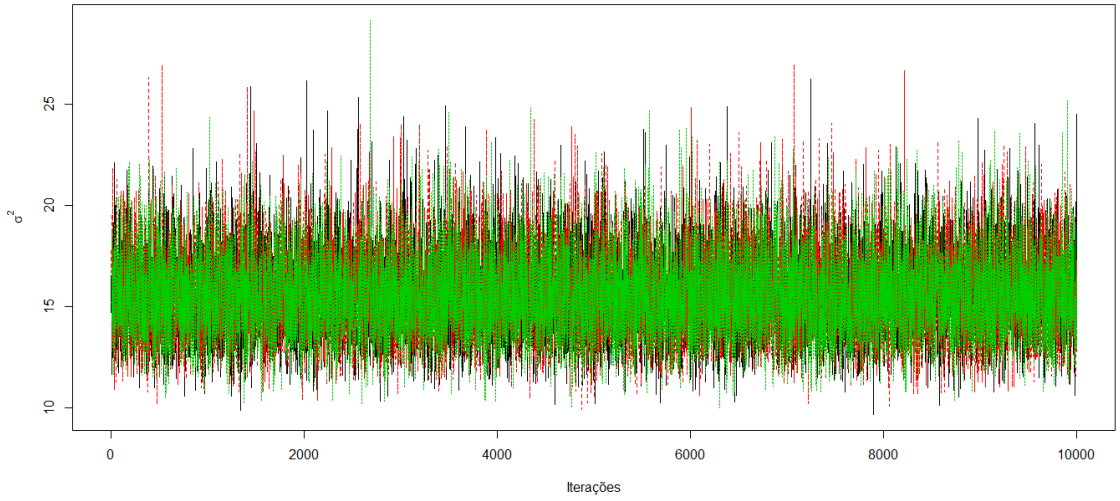

Figura 12 - Gráfico de traços do parâmetro σ^2 referente ao modelo ajustado para o temopar PROBE 5 durante o ciclo 2.

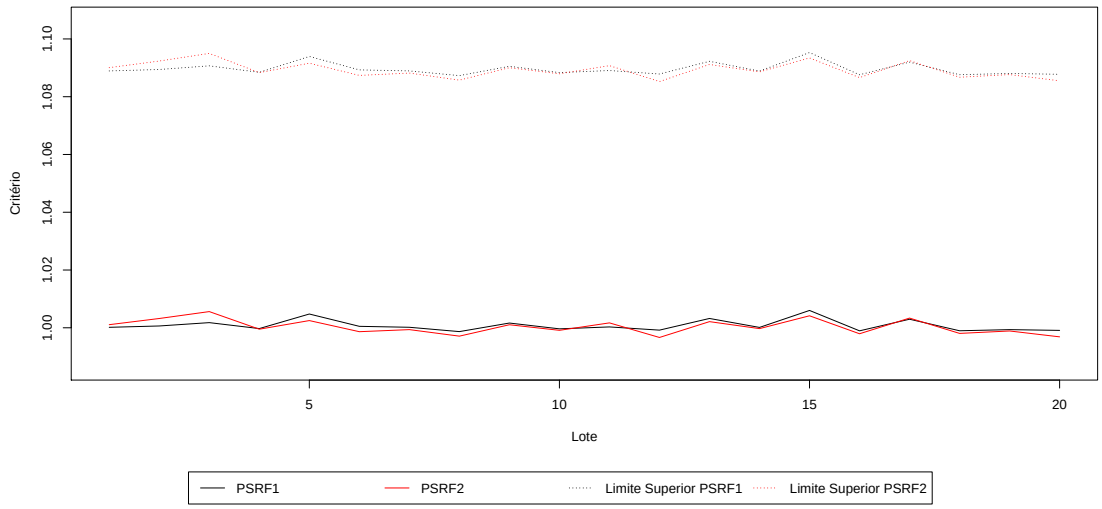

Figura 13 - Potenciais de redução do parâmetro σ^2 referente ao modelo ajustado para o temopar PROBE 5 durante o ciclo 2.

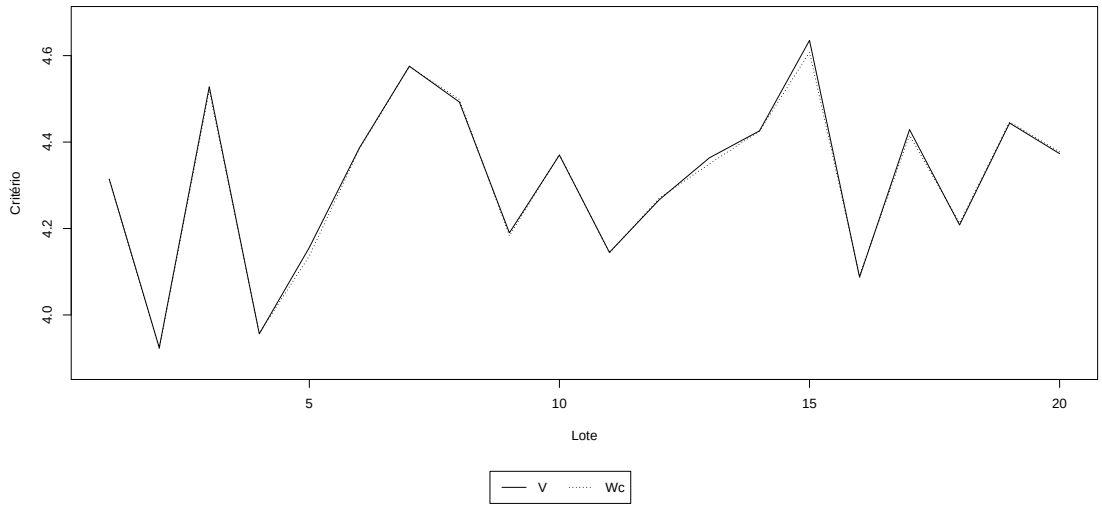

Figura 14 - Valores de V e Wc do parâmetro σ^2 referente ao modelo ajustado para o temopar PROBE 5 durante o ciclo 2.

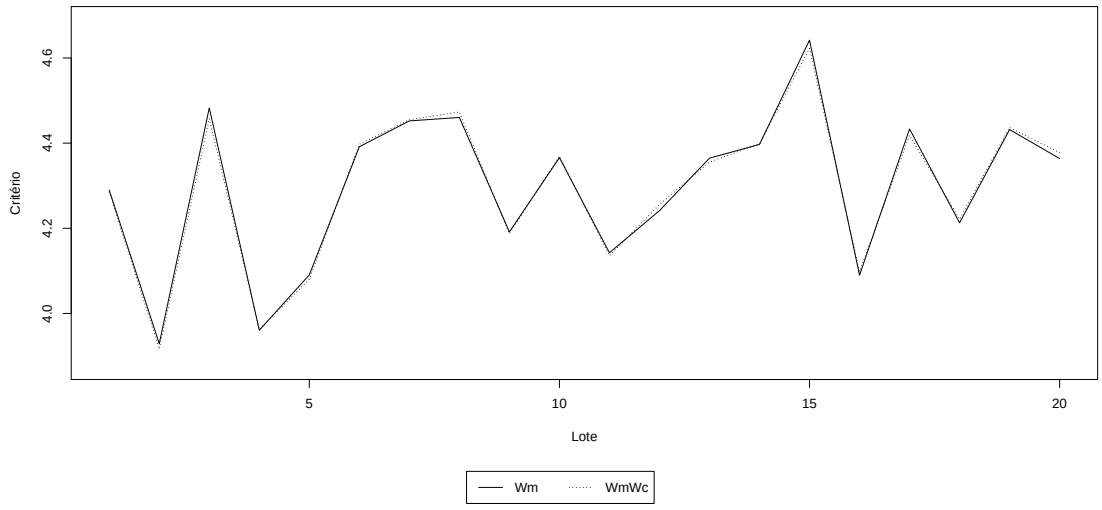

Figura 15 - Valores de Wm e $WmWc$ do parâmetro σ^2 referente ao modelo ajustado para o temopar PROBE 5 durante o ciclo 2.

A dependência foi avaliada usando gráficos de autocorrelação para o parâmetro σ^2. Na Figura 16 tem-se o gráfico de autocorrelação para a cadeia 3 do PROBE 5 para o ciclo 2, em que os valores de autocorrelação para atrasos diferentes de 0 oscilam próximo de 0, ilustrando que são independentes. Os gráficos de autocorrelação das outras cadeias indicam resultados similares.

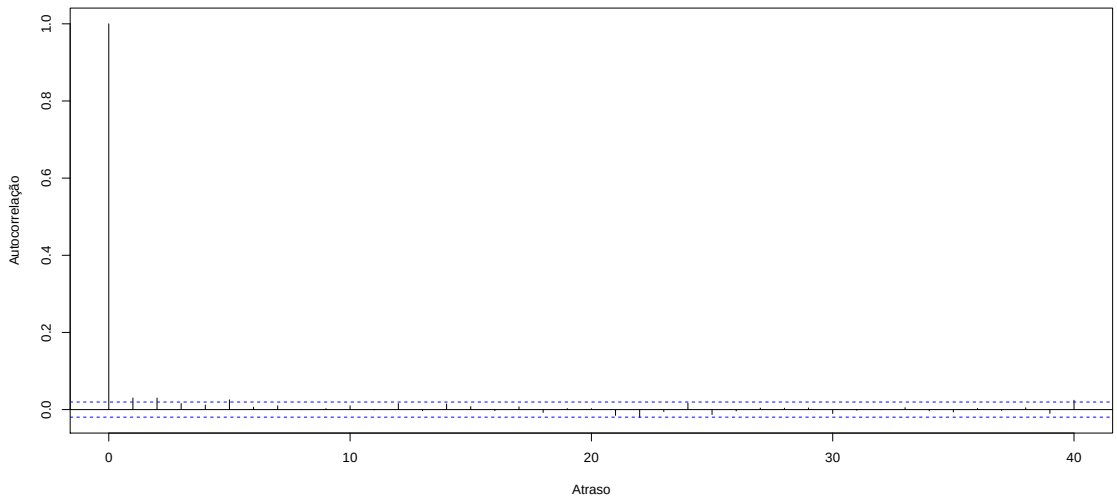

Figura 16 - Gráfico de autocorrelação do parâmetro σ^2 referente a cadeia 3 do modelo ajustado para o temopar PROBE 5 durante o ciclo 2.

Os gráficos de resíduos, como ilustrado na Figura 17, referente ao modelo ajustado pelo método BASS, mostram heterocedasticidade da variância dos resíduos, tal que a variância parece aumentar quanto mais baixas são as estimativas de temperatura pelo modelo. Os gráficos quantil-quantil apresentados ilustram fuga da normalidade, observados pela grande quantidade de pontos nas caudas que estão fora das bandas de confiança. Apesar de violar às hipóteses assumidas para os resíduos, a curva média ajustada ilustra bem o comportamento médio dos dados. Os resultados encontrados para o método `freeknotsplines` são semelhantes aos observados para o método BASS.

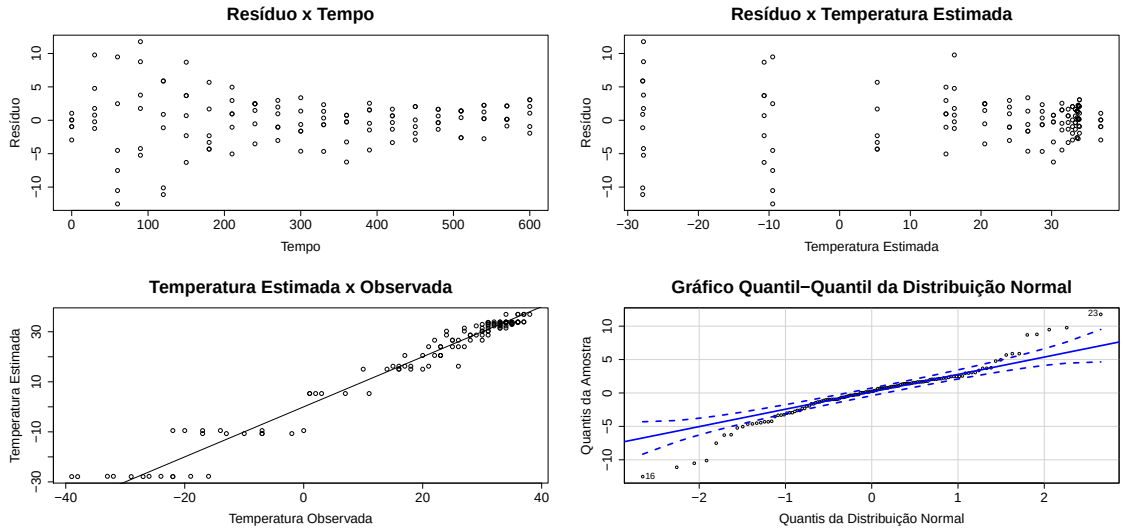

Figura 17 - Gráfico de resíduos do modelo BASS ajustado para o termopar PROBE 5 durante ciclo 2.

Analisando a complexidade do modelo, tem-se que a função ajustada pelo algoritmo `freeknotsplines` tem 3 nós com valores iguais a 64,7, 192,2 e 259,4 (segundos) e a Figura 18 mostra que o número mais frequente de funções de base no modelo BASS é 4. Dessa forma, o uso de funções *splines* são necessárias no contexto das expansões polinomiais cúbicas, visto que um polinômio cúbico simples seria incapaz de uma aproximação adequada.

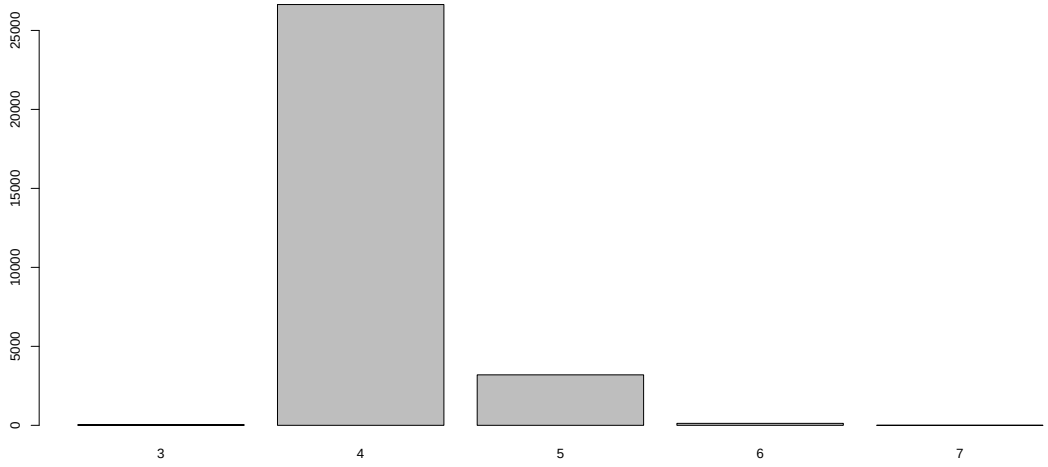

Figura 18 - Número de funções de base nas diferentes interações do modelo BASS ajustado aos dados do PROBE 5 durante o ciclo 2.

Além dos gráficos apresentados, outros, como os gráficos de autocorrelação para outras cadeias e os resíduos para o modelo do pacote **freeknotsplines** estão nos arquivos suplementares.

5.2 Curvas de DOB na ausência e presença de *Helicobacter pylori*

O gráfico de dispersão dos modelos ajustados para as medidas de DOB repetidas no tempo para os pacientes com exame histológico positivo na Figura 19 mostra claramente que existe uma tendência, pequena, de aumento do DOB até atingir um platô para finalmente voltar a cair com uma grande variabilidade. Enquanto que no gráfico com os modelos ajustados para os pacientes com exame histológico negativo na Figura 20 a maioria dos valores parecem permanecer constantes um pouco acima do 0, o que também é um indício de que uma estratégia para modelagem mais simples poderia ser utilizada. A Figura 21 traz os modelos ajustados pelo algoritmo

`freeknotsplines` para o caso positivo e negativo em conjunto, modelos que não consideram a dependência temporal dos dados. É notável que o comprimento do intervalo de credibilidade é bem extenso para a curva ajustada ao DOB dos pacientes com exame histológico positivo. Uma alternativa seria utilizar um intervalo de confiança assimétrico. Figura semelhante, com as curvas para os pacientes com exame histológico positivo e negativo em conjunto geradas pelo modelo BASS, também foi elaborada, sendo o resultado observado para os dois métodos bastante semelhante. Considerando o algoritmo `freeknotsplines`, é possível obter as expressões para as curvas ajustadas, em que $T \in [5, 45]$ é o tempo em minutos após ingestão da uréia ^{13}C:

- Curva ajustada aos pacientes com exame histológico negativo:

$$f(T) = 0{,}832 - 0{,}058 \times T + 0{,}002 \times T^2 - 1{,}849 \times 10^{-5} \times T^3 \qquad (81)$$

- Curva ajustada aos pacientes com exame histológico positivo:

$$f(T) = 12{,}541 + 3{,}448 \times T - 0{,}131 \times T^2 + 0{,}001 \times T^3 \qquad (82)$$

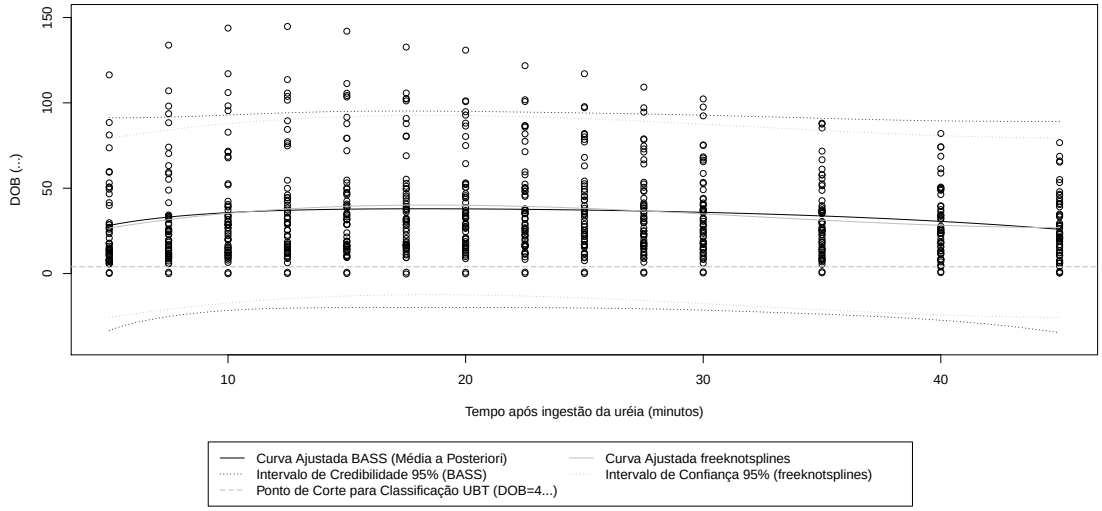

Figura 19 - Gráfico de dispersão dos modelos ajustados para os pacientes com exame histológico positivo.

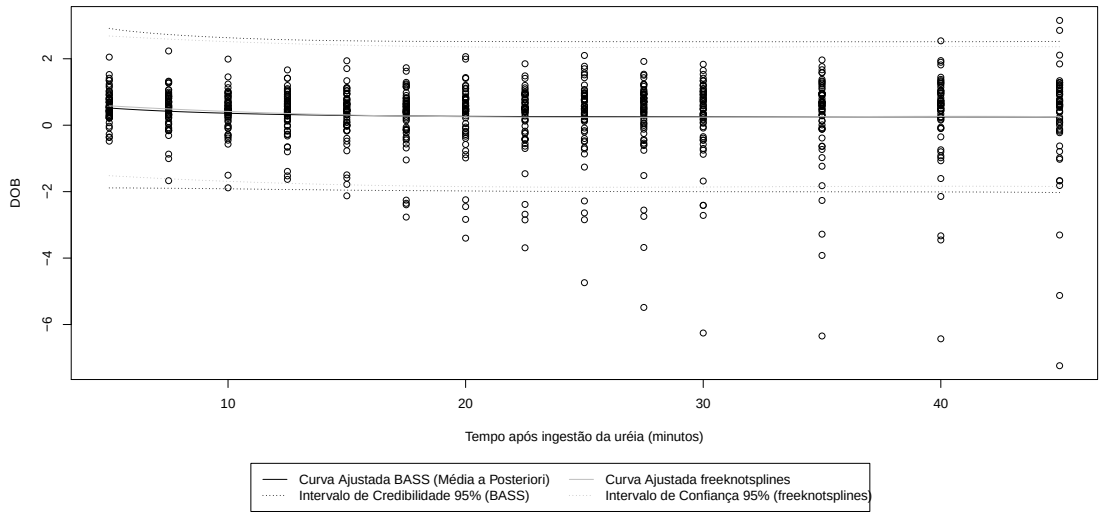

Figura 20 - Gráfico de dispersão dos modelos ajustados para os pacientes com exame histológico negativo.

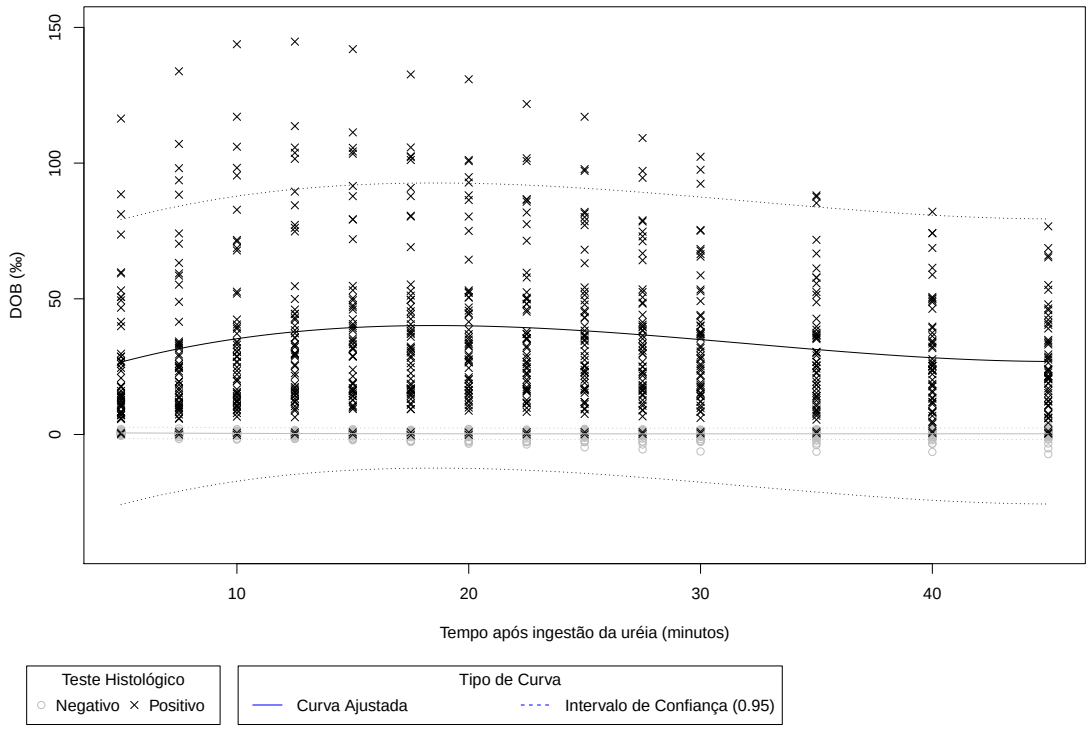

Figura 21 - Gráfico de dispersão do modelo ajustado pelo algoritmo `freeknotsplines` com curvas ajustadas para os pacientes com exame histológico positivo e negativo.

O gráfico com as densidades preditivas para os pacientes com exame histológico positivo na Figura 22 mostra que todas se aproximam muito da média, apesar da grande variabilidade nos dados. Resultado similar foi observado para os gráficos dos pacientes com exame histológico negativo.

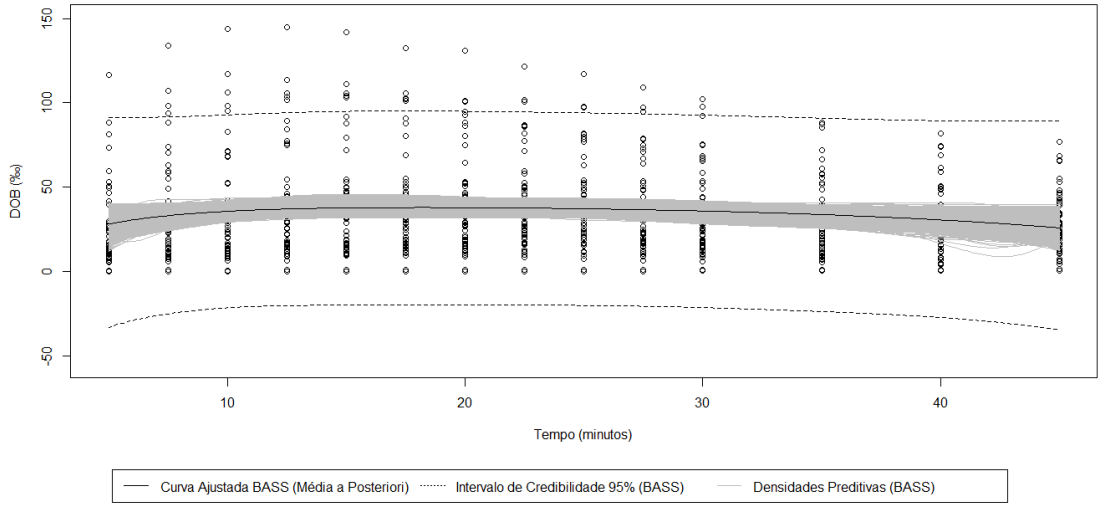

Figura 22 - Gráfico de dispersão dos modelos ajustados com densidades preditivas para os pacientes com exame histológico positivo.

Os gráficos de resíduos para o modelo gerado pelo algoritmo `freeknotsplines` ajustado aos dados dos pacientes com exame histológico positivo na Figura 23 mostra graves problemas relacionados à heterocedasticidade e não adequação ao modelo normal, semelhante ao que também ocorre no gráfico de resíduos para o modelo ajustado para os dados dos pacientes com exame histológico negativo observado na Figura 24. Dessa forma, constatou-se que, para este conjunto de dados, ambas as metodologias não são flexíveis o suficiente para assumir outras estruturas de variância para o resíduo, apesar das curvas ajustadas representarem bem o comportamento médio dos dados.

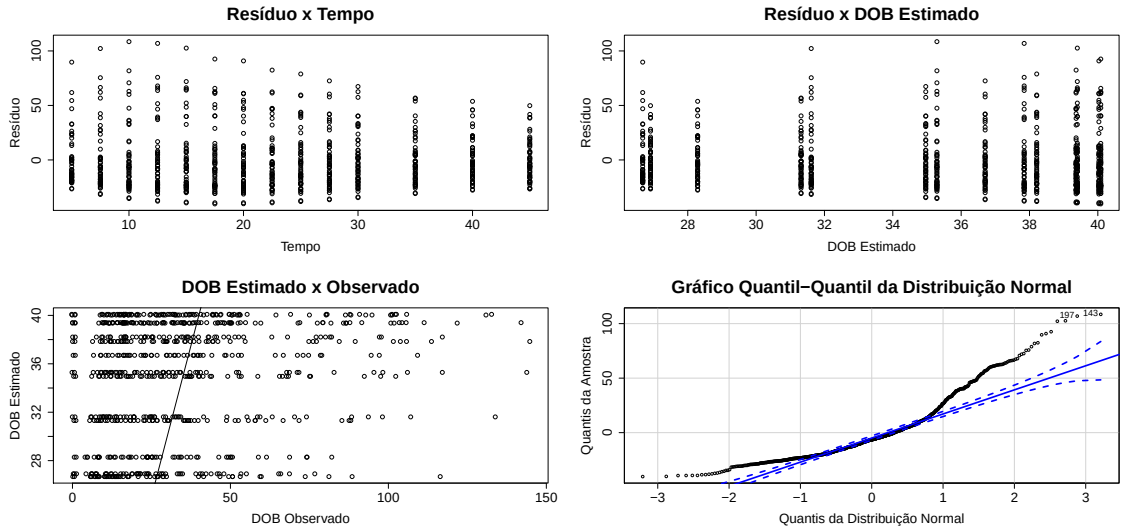

Figura 23 - Gráficos de resíduos do modelo gerado pelo algoritmo `freeknotsplines` ajustado para as medidas de DOB dos pacientes com exame histológico positivo.

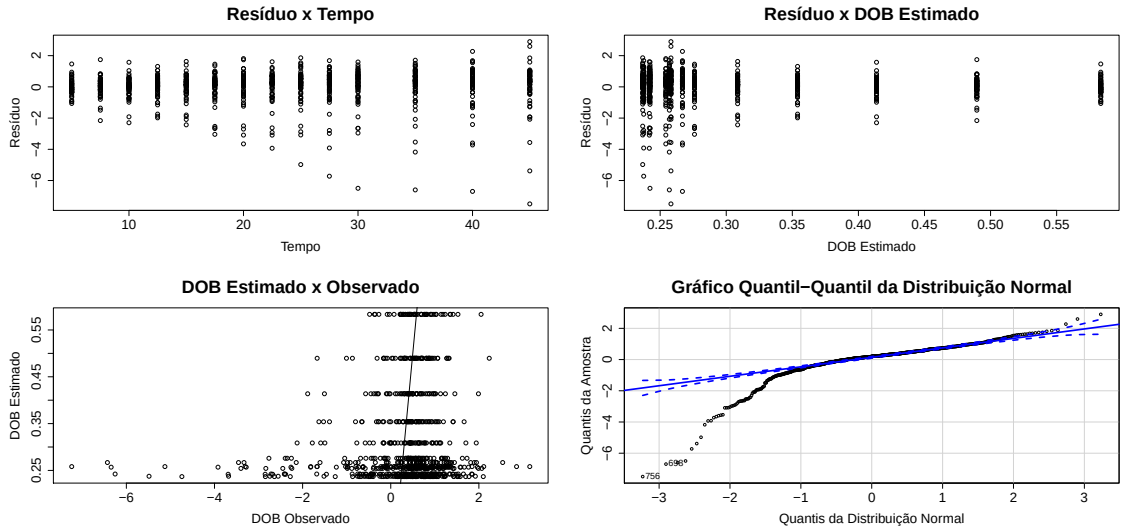

Figura 24 - Gráficos de resíduos do modelo gerado pelo algoritmo `freeknotsplines` ajustado para as medidas de DOB dos pacientes com exame histológico negativo.

Os problemas relacionados a heterocedasticidade dos dados são, provavelmente, relacionado aos diferentes comportamentos entre pacientes. Os gráficos de interação nas Figuras 25 e 26 mostram, para pacientes com exame histológico negativo e positivo, respectivamente, que o comportamento de alguns indivíduos difere muito dos demais. Ilustrando ser preferível utilizar metodologias que considerem as medidas repetidas por pacientes para o presente banco de dados.

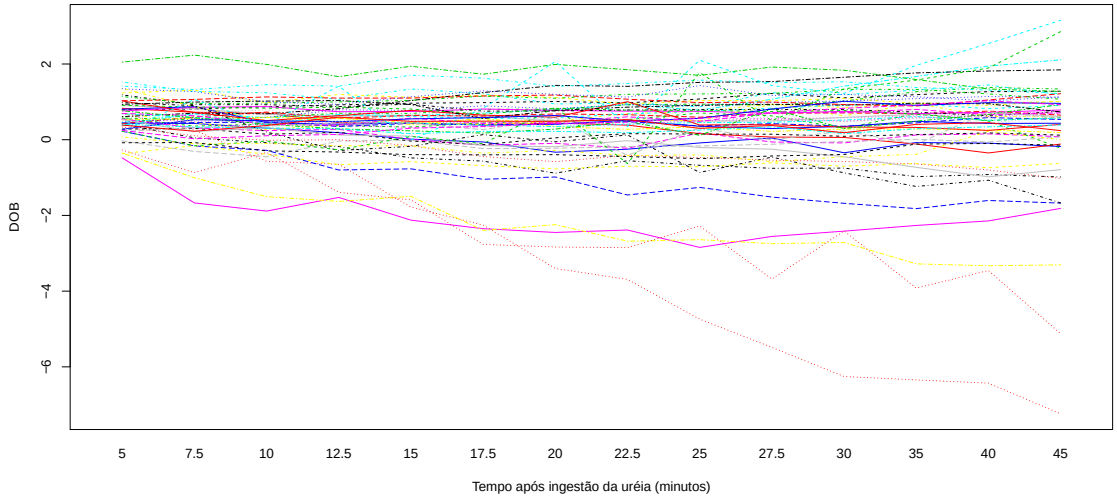

Figura 25 - Gráfico de interação entre DOB e tempo discriminado por paciente com exame histológico negativo. Pacientes diferentes estão em linhas diferentes.

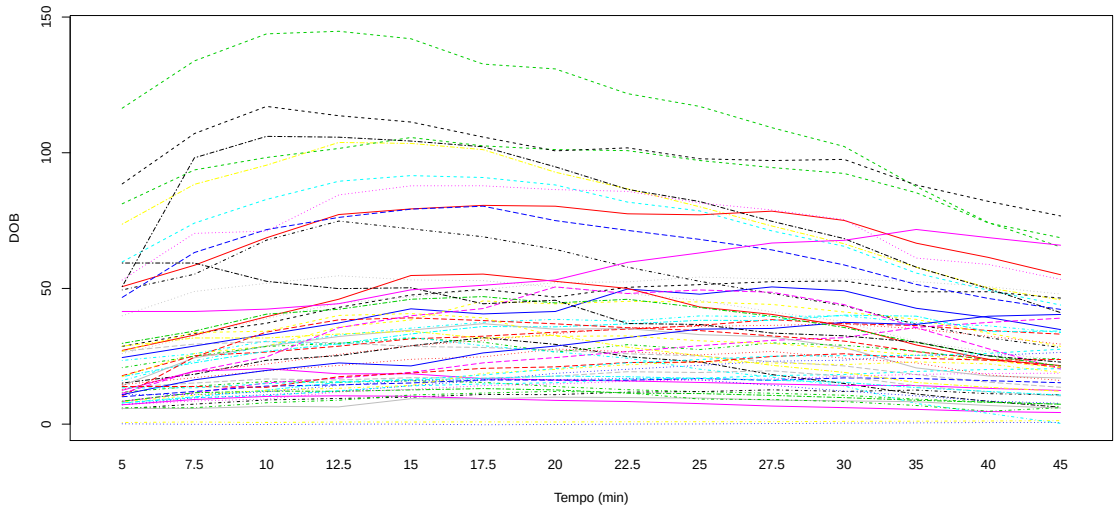

Figura 26 - Gráfico de interação entre DOB e tempo discriminado por paciente com exame histológico positivo. Pacientes diferentes estão em linhas diferentes.

A Figura 27 apresenta a menor credibilidade necessária para que os intervalos não tenham intersecção calculada para cada momento. Dessa forma, o intervalo ótimo para ser realizado o teste respiratório da uréia está entre 16 e 21 minutos. Considerando os resultados observados em 17,5 e 20 minutos e utilizando o ponto de corte de $DOB > 4‰$, temos 52 classificados corretamente como positivos e 58 como negativos, nenhum classificado incorretamente como positivo e 2 incorretamente como negativo, resultado idêntico ao teste UBT aplicado a todo o perfil, o que mostra que esses tempos são adequados para realizar o teste.

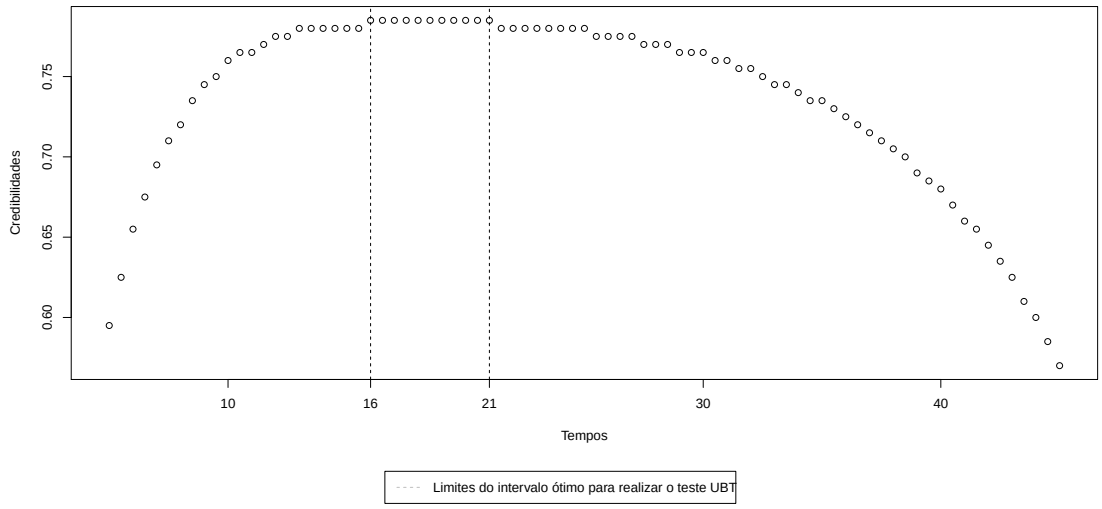

Figura 27 - Gráficos das credibilidades máximas dos intervalos para cada tempo, de modo que os intervalos não se cruzem e seja constatada diferença entre as curvas médias.

Analisando a complexidade das funções para o caso negativo e positivo, o modelo ótimo obtido pelo método `freeknotsplines` seria uma função polinomial simples, ou seja, um polinômico cúbico seria suficiente em ambos os casos, com a possibilidade, inclusive, do uso de polinômios de graus inferiores. Ao analisarmos o número de funções de base escolhidos pelo modelo BASS, observa-se uma pequena diferença, pois nota-se que o mais frequente para o modelo com dados de DOB para pacientes com exame histológico positivo é 2 e negativo é 1, conforme as Figuras 28 e 29, respectivamente.

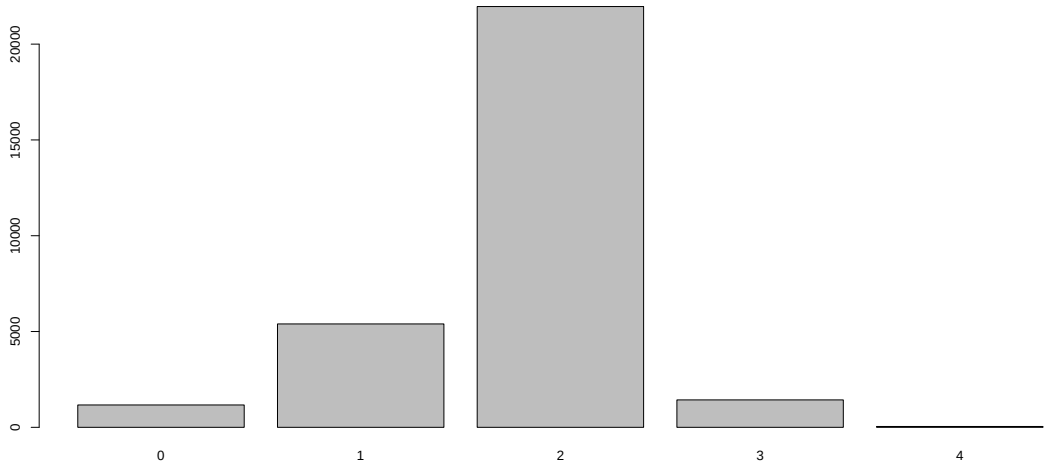

Figura 28 - Número de funções de base nas diferentes interações do modelo BASS ajustado aos dados dos pacientes com exame histológico positivo.

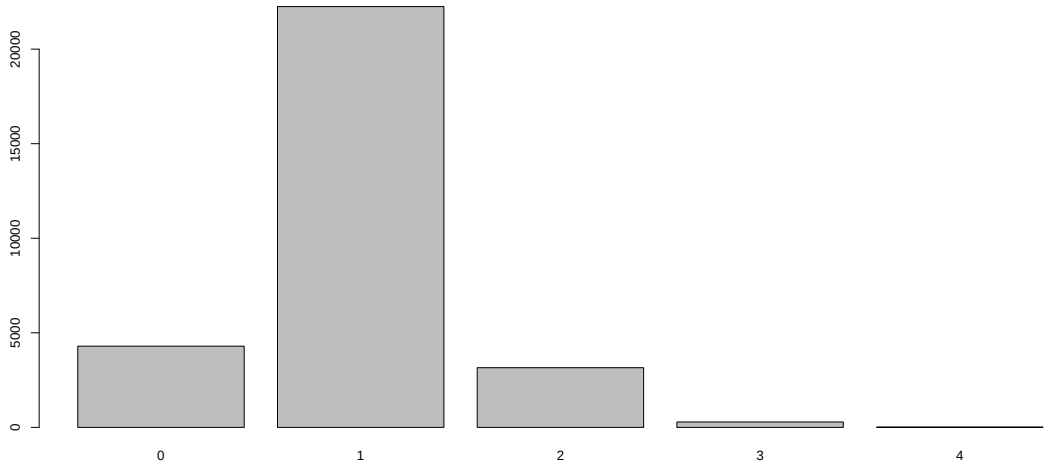

Figura 29 - Número de funções de base nas diferentes interações do modelo BASS ajustado aos dados dos pacientes com exame histológico negativo.

Demais gráficos, além desses discutidos no texto, como os gráficos de análise de convergência e dependência e demais diagramas de dispersão, estão na subpasta DOB de imagens nos arquivos suplementares.

6 CONCLUSÕES

A metodologia BASS e o método de splines de nós-livres baseado no algoritmo genético implementado pelo pacote **freeknotsplines** mostraram ser ferramentas eficientes para o ajuste de curvas mas, ambos, apresentam a mesma limitação, de assumir que a variância dos resíduos é constante, homocedástica. Constatou-se que as metodologias não são adequadas para os problemas avaliados nesse trabalho, podendo levar a erros de inferência. Dentre as alternativas possíveis, tem-se o uso de modelos de regressão linear com as bases ou com os nós encontrados pelo algoritmo do pacote **freeknotsplines**, em conjunto com outras metodologias como mínimos quadrados ponderados ou estimadores robustos para a variância. Alternativa, todavia, que não pode ser diretamente implementada para o contexto da metodologia BASS, visto que as previsões são feitas através de um conjunto de modelos e não um único modelo ótimo. Neste caso, a metodologia BASS teria que ser adaptada para assumir heterocedasticidade o que, provavelmente, geraria modelos muito específicos para cada problema. Essas questões ilustram que há uma demanda por métodos de *splines* de nós livres que sejam mais flexíveis com a estrutura de variância e permitam trabalhar com dados de medidas repetidas, que é o caso dos conjuntos de dados utilizados nesse trabalho.

Quanto às questões dos problemas práticos da área da saúde do trabalho, o mesmo conseguiu oferecer pistas que podem auxilar a respondê-las, todavia, para que a resposta oferecida seja confiável seriam necessários modelos que assumisem uma estrutura mais flexível de variância. No caso do tempo ótimo para realização do teste UBT, é notável que uma metodologia de ajuste de curvas conseguirá oferecer pistas mas não responder adequadamente qual seria esse período ótimo, visto ser

um problema, notadamente, de classificação. Assim, a busca para a solução mais adequada para esse problema está na procura de métodos de classificação que considerem essa natureza longitudinal dos dados. O período entre 16 e 21 minutos após a ingestão da uréia ^{13}C, com credibilidade de $78,5\%$, apontado pelos resultados como ótimo para a realização do teste parece ser um resultado coerente com a realidade.

Agora tratando das curvas de temperatura, por mais que conclusões a respeito da mortalidade e danos a tecidos tenham sido feitas considerando o termopar PROBE 5 ciclo 2, procedimento que pode ser repetido facilmente para o outros termopares, é notável que a interpretação correta desses resultados e qualquer tentativa de utilizá-los é extremamente dependente de conhecimentos especializados. Além disso, a quantidade disponível de dados, repetições, é extremamente pequena, o que provoca grandes dificuldades para uma inferência adequada, passível de ser extrapolada e utilizada dentro da prática médica com segurança, além dos aspectos relacionados à natureza longitudinal dos dados.

Além disso, quanto à complexidade dos modelos, no caso das curvas de DOB a relação tempo e DOB não é tão complexa e modelos mais simples podiam ser usados para estudá-la, o que implica que soluções considerando sua natureza longitudinal são mais simples de serem encontradas. No caso das curvas de termopares a relação é, de fato, complexa, o que implicaria em métodos mais específicos, como modelos não-lineares ou mesmo de *splines*, para o estudo do problema que considerem o contexto de medidas repetidas.

REFERÊNCIAS

BELIAKOV, G. Least squares splines with free knots: global optimization approach. **Applied mathematics and computation**, v.149, n.3, p.783–798, 2004.

BOJANOV, B. D.; HAKOPIAN, H.; SAHAKIAN, B. **Spline functions and multivariate interpolations**. Dordrecht: Springer Science & Business Media, 1993. 248v.

BRAIBANT, V.; FLEURY, C. Shape optimal design using B-splines. **Computer Methods in Applied Mechanics and Engineering**, v.44, n.3, p.247–267, 1984.

BROOKS, S. P.; GIUDICI, P. Markov chain Monte Carlo convergence assessment via two-way analysis of variance. **Journal of Computational and Graphical Statistics**, v.9, n.2, p.266–285, 2000.

BROOKS, S. P.; ROBERTS, G. O. Convergence assessment techniques for Markov chain Monte Carlo. **Statistics and Computing**, v.8, n.4, p.319–335, 1998.

CASTELLOE, J. M.; ZIMMERMAN, D. L. Convergence assessment for reversible jump MCMC samplers. Rel. téc., Department of Statistics and Actuarial Science, University of Iowa, 2002.

CRAVEN, P.; WAHBA, G. Smoothing noisy data with spline functions. **Numerische mathematik**, v.31, n.4, p.377–403, 1979.

DENISON, D. G. T.; MALLICK, B. K.; SMITH, A. F. M. Automatic Bayesian curve fitting. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, v.60, n.2, p.333–350, 1998.

DIMATTEO, I.; GENOVESE, C. R.; KASS, R. E. Bayesian curve-fitting with free-knot splines. **Biometrika**, v.88, n.4, p.1055–1071, 2001.

ERINJERI, J. P.; CLARK, T. W. Cryoablation: Mechanism of Action and Devices. **Journal of Vascular and Interventional Radiology**, v.21, n.8, p.S187–S191, 2010.

FAN, Y.; SISSON, S. A. Reversible Jump MCMC. In: BROOKS, S.; GELMAN, A.; JONES, G. L.; MENG, X.-L., (Ed.), **Handbook of Markov Chain Monte Carlo.** Boca Raton: Chapman & Hall/CRC, p.67–87, 2011.

FARAWAY, J. J. **Linear models with R**. Boca Raton: CRC press, 2014.

FRANCOM, D.; SANSÓ, B. Bass: An R package for fitting and performing sensitivity analysis of Bayesian adaptive spline surfaces. **Journal of Statistical Software**, v.2, 2019.

FRANCOM, D.; SANSÓ, B.; BULAEVSKAYA, V.; LUCAS, D.; SIMPSON, M. Inferring Atmospheric Release Characteristics in a Large Computer Experiment Using Bayesian Adaptive Splines. **Journal of the American Statistical Association**, v.114, n.528, p.1450–1465, 2019.

FRANCOM, D.; SANSO, B.; KUPRESANIN, A.; JOHANNESSON, G. Sensitivity Analysis and Emulation for Functional Data using Bayesian Adaptive Splines. **Statistica Sinica**, 2018.

FREITAS, R. M. C. D. Desenvolvimento de um modelo experimental de crioablação vertebral em suínos guiada por tomografia computadorizada de feixe cônico, 2015. Tese (Doutorado) - Universidade de São Paulo.

FREITAS, R. M. C. D.; ANDRADE, C. S.; CALDAS, J. G. M. P.; TSUNEMI, M. H.; FERREIRA, L. B.; ARANA-CHAVEZ, V. E.; CURY, P. M. Image-Guided Cryoablation of the Spine in a Swine Model: Clinical, Radiological, and Pathological Findings with Light and Electron Microscopy. **CardioVascular and Interventional Radiology**, v.38, n.5, p.1261–1270, 2015.

FRIEDMAN, J. H. Multivariate Adaptive Regression Splines. **The Annals of Statistics**, v.19, n.1, p.1–67, 1991.

FRIEDMAN, J. H.; SILVERMAN, B. W. Flexible Parsimonious Smoothing and Additive Modeling. **Technometrics**, v.31, n.1, p.3–21, 1989.

GÁLVEZ, A.; IGLESIAS, A. Efficient particle swarm optimization approach for data fitting with free knot B-splines. **Computer-Aided Design**, v.43, n.12, p.1683–1692, 2011.

GÁLVEZ, A.; IGLESIAS, A.; AVILA, A.; OTERO, C.; ARIAS, R.; MANCHADO, C. Elitist clonal selection algorithm for optimal choice of free knots in B-spline data fitting. **Applied Soft Computing**, v.26, p.90–106, 2015.

GAMERMAN, D.; LOPES, H. F. **Markov Chain Monte Carlo**. Taylor & Francis Inc, 2006.

GARCIA, B. D. O. Influência do jejum e do ácido cítrico no teste respiratório com isótopos estáveis do carbono para detecção da infecção por Helicobacter pylori. Botucatu, 2017. Dissertação (Mestrado) - Universidade Estadual Paulista (UNESP).

GELMAN, A.; RUBIN, D. B. Inference from Iterative Simulation Using Multiple Sequences. **Statistical Science**, v.7, n.4, p.457–472, 1992.

GRAYBILL, F. A. **An introduction to linear statistical models**. New York: McGraw-Hill, 1961.

GREEN, P. J. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. **Biometrika**, v.82, n.4, p.711–732, 1995.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning: data mining, inference, and prediction**. 2. ed. 12ª reimpressão corrigida. New York: Springer-Verlag, 2017. Springer Series in Statistics.

JOHS, B.; HALE, J. S. Dielectric function representation by B-splines. **Physica Status Solidi (a)**, v.205, n.4, p.715–719, 2008.

JUPP, D. L. Approximation to data by splines with free knots. **SIAM Journal on Numerical Analysis**, v.15, n.2, p.328–343, 1978.

LIMA, E. L. **Análise real volume 1**. Rio de Janeiro: Impa, 2009. Coleção Matemática Universitária.

LINDSTROM, M. J. Penalized estimation of free-knot splines. **Journal of Computational and Graphical Statistics**, v.8, n.2, p.333–352, 1999.

LINDSTROM, M. J. Bayesian estimation of free-knot splines using reversible jumps. **Computational statistics & data analysis**, v.41, n.2, p.255–269, 2002.

MARQUES, T. P.; TSUNEMI, M. H. **convRJMCMC: Convergence assessment of RJMCMC algorithms**, 2021. R package version 0.1.0.

MEINARDUS, G.; NÜRNBERGER, G.; SOMMER, M.; STRAUSS, H. Algorithms for piecewise polynomials and splines with free knots. **Mathematics of computation**, v.53, n.187, p.235–247, 1989.

MICULA, G.; MICULA, S. **Handbook of splines**. Dordrecht: Springer Science & Business Media, 2012.

MOLINARI, N.; DURAND, J.-F.; SABATIER, R. Bounded optimal knots for regression splines. **Computational statistics & data analysis**, v.45, n.2, p.159–178, 2004.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. **Introduction to linear regression analysis**. Hoboken: John Wiley & Sons, 2012.

NOTT, D. J.; KUK, A. Y. C.; DUC, H. Efficient sampling schemes for Bayesian MARS models with many predictors. **Statistics and Computing**, v.15, n.2, p.93–101, 2005.

PAULINO, C. D.; AMARAL TURKMAN, M. A.; MURTEIRA, B.; SILVA, G. L. **Estatística Bayesiana**. 2. ed. Lisboa: Fundação Calouste Gulbenkian, 2018.

PIEGL, L.; TILLER, W. **The NURBS book**. 2. ed. New York: Springer-Verlag, 1996.

PLUMMER, M.; BEST, N.; COWLES, K.; VINES, K. CODA: Convergence Diagnosis and Output Analysis for MCMC. **R News**, v.6, n.1, p.7–11, 2006.

R CORE TEAM. R: A Language and Environment for Statistical Computing, 2019.

SCHUMAKER, L. **Spline functions: basic theory**. 3. ed. New York: Cambridge University Press, 2007.

SHIKIN, E. V.; PLIS, A. I. **Handbook on Splines for the User**. Boca Raton: CRC Press, 1995.

SONDEREGGER, D. L.; HANNIG, J. Fiducial theory for free-knot splines. In: **Contemporary Developments in Statistical Theory** Cham: Springer International Publishing, Springer Proceedings in Mathematics & Statistics, p.155–189, 2014.

SPIRITI, S.; EUBANK, R.; SMITH, P. W.; YOUNG, D. Knot selection for least-squares and penalized splines. **Journal of Statistical Computation and Simulation**, v.83, n.6, p.1020–1036, 2013.

SPIRITI, S.; SMITH, P.; LECUYER, P. freeknotsplines: Algorithms for Implementing Free-Knot Splines, 2018. R package version 1.0.1.

VENABLES, W. N. **Modern applied statistics with S**. New York: Springer, 2002.

APÊNDICES

Repositório do GitHub com arquivos suplementares:
https://github.com/TPMarques/suppFilesdissertation
Repositório do GitHub do pacote convRJMCMC:
https://github.com/TPMarques/convRJMCMC